On the Approximation Power of Two-Layer Networks of Random ReLUs

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Clayton Sanford CLAYTON@CS.COLUMBIA.EDU
Rocco A. Servedio ROCCO@CS.COLUMBIA.EDU
Emmanouil V. Vlatakis Gkaragkounis EMVLATAKIS@CS.COLUMBIA.EDU
Columbia University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

This paper considers the following question: how well can depth-two ReLU networks with randomly initialized bottom-level weights represent smooth functions? We give near-matching upper- and lower-bounds for L_2-approximation in terms of the Lipschitz constant, the desired accuracy, and the dimension of the problem, as well as similar results in terms of Sobolev norms. Our positive results employ tools from harmonic analysis and ridgelet representation theory, while our lower-bounds are based on (robust versions of) dimensionality arguments.

Keywords: Function representation, random initialization, deep learning, ReLU networks

1. Introduction

1.1. Background and motivation

Celebrated results of Cybenko (1989), Funahashi (1989), and Hornik et al. (1989) establish the universality of depth-2 neural networks by showing that any continuous function on $\mathbb{R}^d$ can be approximated by a neural network with a single hidden layer. However, these results offer no upper-bound (e.g., in terms of d) on the width (number of bottom-level gates) required, leaving unanswered many natural questions about the approximation power of neural networks, including:

- Which functions can be approximated by two-layer neural networks of subexponential width?
- Can tradeoffs be achieved between depth and width for neural network function approximation?
- Given the practical importance of random weight initialization, what are the representational capabilities of neural networks with some randomly drawn weights (say, at the bottom level)?

The first two questions above have been studied intensely in the approximation-theoretic and depth-separation literature; this paper focuses on the third question. Random weight initializations play an important role in training neural networks in practice, and are also of theoretical interest; as we discuss later in this introduction, they have been well studied as a way of understanding different aspects of approximation and generalization.

In this work, we study the representational ability of depth-2 random bottom-layer (RBL) ReLU networks. Such a network is equivalent to a linear combination of rectified linear units (ReLUs), where the weight vector and bias of each ReLU are randomly and independently chosen from a fixed distribution, but the top-level combining weights of the ReLUs are allowed to be arbitrary (we give precise definitions in Section 2.2). This particular setting is of interest because, as discussed later, a number of papers have given approximation-theoretic results in this regime. We choose the ReLU activation due to its popularity in both theory and practice; we expect that the results of our paper could be generalized to a range of other activation functions.

Our main goal is to understand the abilities and limitations of depth-2 RBL ReLU networks for approximating smooth functions of various types. We focus on smooth functions both because they are a natural class of functions to consider, and because non-smooth functions have been shown to be difficult to approximate by various types of neural networks. Indeed, several authors (e.g., Telgarsky (2016) and Daniely (2017)) have established lower-bounds on the width of neural networks that approximate certain non-smooth functions by taking advantage of the fact that such functions can be highly oscillatory (have many "bumps") and can require many gates to approximate each "bump."

Our chief focus is on functions over the d-dimensional solid cube $[-1,1]^d$ (though we also consider functions over d-dimensional Gaussian space in Appendix E) whose smoothness is measured in two different ways. Our main results are about approximating functions on $[-1,1]^d$ with bounded *Lipschitz constants*; in Appendix D, we also consider functions on $[-1,1]^d$ (satisfying certain periodicity conditions) with bounded *Sobolev norms*.

1.2. Our results

The main contributions of this work are to pose and answer the following question:

> *What is the minimum number of random ReLU features required so that (with high probability) there exists some linear combination of those features that closely approximates any sufficiently smooth function?*

This minimum number of random ReLU features is equivalent to the minimum width required for a depth-2 RBL ReLU network to approximate the smooth function in question. We give full details about our setting in Section 2.2, and here only touch on some of the main aspects:

- "Random ReLU features" are functions from $\mathbb{R}^d$ to $\mathbb{R}$ that are drawn independently from some fixed distribution. These take the form $x \mapsto \sigma_{\mathrm{ReLU}}(\langle \mathbf{w}, x\rangle + \mathbf{b})$ where $\sigma_{\mathrm{ReLU}}(z) := \max(z, 0)$ and $\mathbf{w}$ and $\mathbf{b}$ are random variables taking values in $\mathbb{S}^{d-1}$ and $\mathbb{R}$ respectively.
- Our notion of "close approximation" refers to the L_2 distance between functions with respect to the uniform distribution on the solid cube; we say that f is an ϵ-approximator for g if $\|f-g\|_{[-1,1]^d} \leq \epsilon$. In Appendix E, we sketch how analyses similar to our analysis over $[-1,1]^d$ can be used to study approximation with respect to the Gaussian measure over $\mathbb{R}^d$.
- As mentioned above, we chiefly measure the smoothness of a function by its Lipschitz constant. In Appendix D, we extend our results to measure smoothness in terms of Sobolev norms.

Our main results give tight upper- and lower-bounds on the minimum width required for both Lipschitz and Sobolev smooth functions. The upper- and lower-bounds match up to polynomial factors (equivalently, up to constant factors in the exponent). The sharpest forms of our bounds involve the number of integer points in certain Euclidean balls; below, we present informal statements of our upper- and lower-bounds for Lipschitz functions with explicit asymptotics given for clarity:

Theorem 1 (Informal upper-bound for L-Lipschitz functions) *Fix any $\epsilon, L > 0$ that satisfy $L/\epsilon \geq 2$, and let $f : [-1,1]^d \to \mathbb{R}$ be any L-Lipschitz function. For*

$$r = \exp\left(O\left(\min\left(\frac{L^2}{\epsilon^2}\log\left(\frac{d\epsilon^2}{L^2}+2\right), d\log\left(\frac{L^2}{\epsilon^2 d}+2\right)\right)\right)\right),$$

with probability 0.9 *(over a draw of* r *i.i.d. random ReLU features* $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(r)}$ *from a suitable distribution) there exists a depth-2 RBL ReLU network* h *with* $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(r)}$ *as the bottom-level features satisfying* $\|f - h\|_{[-1,1]^d} \leq \epsilon$*.*

Theorem 2 (Informal lower-bound for L-Lipschitz functions) *Fix any* $\epsilon, L > 0$*. There exists an* L*-Lipschitz function* $f : [-1,1]^d \to \mathbb{R}$ *such that with probability at least* $\frac{1}{2}$ *over a draw of*

$$r = \exp\left(\Omega\left(\min\left(\frac{L^2}{\epsilon^2}\log\left(\frac{d\epsilon^2}{L^2}+2\right), d\log\left(\frac{L^2}{\epsilon^2 d}+2\right)\right)\right)\right)$$

many i.i.d. random ReLU gates $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(r)}$*, every depth-2 ReLU network* h *of width* r *with* $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(r)}$ *as its bottom-layer gates has* $\|f - h\|_{[-1,1]^d} > \epsilon$*.*

Table 1 summarizes these results, as well as our analogues for functions in Sobolev balls.

Bound	Smoothness	Minimum Width	Theorem
Upper	Lipschitz $\leq L$	$\exp\left(O\left(\min\left(\frac{L^2}{\epsilon^2}\log\left(\frac{d\epsilon^2}{L^2}+2\right), d\log\left(\frac{L^2}{\epsilon^2 d}+2\right)\right)\right)\right)$	Thm. 1 / 6
Lower	Lipschitz $\leq L$	$\exp\left(\Omega\left(\min\left(\frac{L^2}{\epsilon^2}\log\left(\frac{d\epsilon^2}{L^2}+2\right), d\log\left(\frac{L^2}{\epsilon^2 d}+2\right)\right)\right)\right)$	Thm. 2 / 10
Upper	H^s norm $\leq \gamma$	$\exp\left(O\left(\min\left(d\log\left(\frac{s\gamma^{2/s}}{d\epsilon^{2/s}}+2\right), \frac{s\gamma^{2/s}}{\epsilon^{2/s}}\log\left(\frac{d\epsilon^{2/s}}{s\gamma^{2/s}}+2\right)\right)\right)\right)$	Thm. 35
Lower	H^s norm $\leq \gamma$	$\exp\left(\Omega\left(\min\left(d\log\left(\frac{\gamma^{2/s}}{d\epsilon^{2/s}}+2\right), \frac{\gamma^{2/s}}{\epsilon^{2/s}}\log\left(\frac{d\epsilon^{2/s}}{\gamma^{2/s}}+2\right)\right)\right)\right)$	Thm. 39

Table 1: Our upper- and lower-bounds on the minimum width needed for an RBL ReLU network to ϵ-approximate a function over $L_2([-1,1]^d)$ with either bounded Lipschitz constant L, or bounded order-s Sobolev norm γ (and periodic boundary conditions).

Discussion. Our results shed light on a question posed by Safran et al. (2019) about the approximation power of unconstrained depth-2 networks. They ask whether there exists a d-dimensional 1-Lipschitz function f that can be represented by a depth-3 neural network with $\mathrm{poly}(d)$ neurons but requires width $\exp(\Omega(d))$ to be approximated by a depth-2 network. As one of their main results, they answer this question in the negative for pointwise approximation when f is a radial function (depending only on $\|x\|_2$) over the unit ball, by showing that any such function can be efficiently approximated by a $\mathrm{poly}(d)$ width depth-2 network. Our results imply that the answer is also negative for L_2-approximation of *arbitrary* 1-Lipschitz functions (which need not be radial) over $[-1,1]^d$; this follows from our upper-bounds for the case that $L = 1$ and ϵ is any constant, which establish the existence of approximators that are $\mathrm{poly}(d)$-width, depth-2 RBL networks. Our results do not answer their question outright, because showing that every 1-Lipschitz function can be approximated with respect to the L_2 norm over $[-1,1]^d$ by a depth-2 network of $\mathrm{poly}(d)$ width does *not* imply that every 1-Lipschitz function is uniformly approximable by such a network.

Our upper-bounds on the width that suffices to approximate Lipschitz functions are also useful for proving learnability hardness results for neural networks with more than two layers. Malach et al. (2021) establish this connection between hardness of approximation and hardness of learning by showing that any function that cannot be weakly approximated by a network with three layers cannot be learned by gradient descent applied to a neural network of *any* depth, given certain assumptions about the random weight initialization and bounds on the number of units in the network and number of steps of gradient descent. Their result hinges on a technical lemma (their Lemma B.2), which shows that L-Lipschitz functions can be approximated by three layer neural networks with bounded width. By replacing that lemma with our Theorem 6, their result can be strengthened to say that any function not weakly approximable by *two*-layer neural networks is not learnable by gradient descent for networks of any depth that obey their assumptions.

1.3. Our techniques

In this section we give a high-level overview of the ideas that underlie our upper and lower bounds.

1.3.1. UPPER-BOUNDS

Our width upper-bounds state that for any fixed function of the relevant sort, given a large enough number of independent random ReLU features, with high probability some linear combination of those features approximates the function. We argue this in three steps. (Below, we only discuss the Lipschitzness smoothness measure, but the Sobolev case follows the same basic steps.)

1. The first step shows that for any L-Lipschitz function f, there exists a low-degree trigonometric polynomial P that closely approximates f. We establish the existence of this trigonometric polynomial using the fact that any function in $L_2([-1,1]^d)$ can be expressed as a (potentially infinite) linear combination of sinusoidal functions, due to the existence of a Fourier representation for f. We use the Lipschitzness of f to show that high-frequency terms have negligibly small coefficients in the representation, which we drop to obtain a low-degree approximation P.
2. The second step expresses P as an infinite mixture of random ReLU features (à la Barron, 1993; Murata, 1996; Rubin, 1998; Candès, 1999). That is, for some distribution over biases $\mathbf{b}$ and weights $\mathbf{w}$ (which depends on L, ϵ, and d, but not f, and takes values in $\mathbb{R} \times \mathbb{S}^{d-1}$), P can be written as
$$P(x) = \mathop{\mathbb{E}}_{\mathbf{b},\mathbf{w}} \left[h(\mathbf{b},\mathbf{w})\sigma_{\mathrm{ReLU}}\left(\langle \mathbf{w}, x\rangle - \mathbf{b}\right)\right]$$
for some function $h(\mathbf{b},\mathbf{w})$. Intuitively, this is possible because each sinusoidal component of P is a ridge function (a function that depends only on a one-dimensional projection of its input).
3. Finally, using a standard concentration argument, we show that the empirical average of sufficiently many random ReLUs gives a close approximation to P with high probability. It follows that the overall weighted combination of random features closely approximates f.

1.3.2. LOWER-BOUNDS

Our lower-bounds are proved using a dimensionality argument, stemming from the simple observation that linear combinations of r features (functions) can span at most r dimensions in the function space $L_2([-1,1]^d)$. The key is to give $N \gg r$ candidate functions $\varphi_1, \ldots, \varphi_N$ that are orthonormal in $L_2([-1,1]^d)$. With such a set of functions in hand, any fixed outcome of a draw of r random features will be such that linear combinations of those r features cannot closely approximate more than a small fraction of the N functions, because no r-dimensional subspace can be close to a large fraction of N orthonormal functions. (This kind of dimensionality argument has been used in a number of prior works, including Barron (1993); Yehudai and Shamir (2019); Kamath et al. (2020) and elsewhere.)

Specializing to our context, to give a lower-bound on the minimum width of RBL ReLU networks needed to approximate L-Lipschitz functions, it suffices to construct a large family of orthonormal L-Lipschitz functions. We do this with L-Lipschitz sinusoidal functions of the form $\sqrt{2}\sin\left(\pi\left\langle K,x\right\rangle\right)$ where $K \in \mathbb{Z}^d$. The quantity $\|K\|_2$ controls the Lipschitz constant of these functions, and as our analysis shows, the tradeoff between the number of functions in the family (which increases with the allowed range of $\|K\|_2$ and controls our width bound r) and the Lipschitz constant L yields a lower-bound that is quite close to our upper-bound for L-Lipschitz functions.

The simple dimensionality argument sketched above establishes that some function among the N orthonormal functions is hard to approximate (in fact, that most of them are hard), but it does not yield an *explicit* hard function. By requiring the N orthonormal functions $\varphi_1, \ldots, \varphi_N$ to satisfy a natural symmetry property with respect to the random ReLU features, it is possible to get a lower bound for a single explicit function φ_1. Following this approach, we also give a quantitatively slightly weaker lower-bound on the minimum width that random ReLU networks need in order to approximate an explicit function φ_1.

1.4. Related work

Since the pioneering universal approximation results for (non-RBL) depth-2 networks (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989) mentioned in the introduction, many subsequent works have established quantitative bounds on the width that such networks require to approximate certain functions.[1] RBL networks have also been the subject of considerable study owing to their connection to kernel methods (Neal, 1996; Rahimi and Recht, 2008; Cho and Saul, 2009) and, in particular, the Neural Tangent Kernel (NTK). Jacot et al. (2018) argue that training neural networks with gradient descent with small step-sizes results in a learning rule similar to that obtained by a kernel method with the NTK. When the network weights are randomly initialized, then a finite-width NTK corresponds to a linear combination of random ReLUs. Both RBL ReLU networks and the finite-width NTK enjoy the same universal approximation property of non-RBL networks (Sun et al., 2018; Ji et al., 2019), and hence quantitative bounds on the network width required to approximate families of functions are of significant interest.

Upper-bounds. A line of inquiry starting with Barron (1993) (see also Klusowski and Barron, 2018) investigates upper-bounds on the width of (non-RBL) depth-2 networks needed to approximate functions whose smoothness is measured in terms of their Fourier transforms. Although these results do not deal with RBL networks and hence are incomparable to ours, they do use randomization in the proof. Specifically, a target function is represented as a mixture of activation functions drawn from a target-specific distribution, and a finite-width depth-2 network approximating the function is obtained by sampling. Our results use a similar overall approach, but with the crucial difference that in our RBL setting, our distribution of ReLUs does not depend on the target function.

Perhaps the works on RBL networks that are most closely related to our own upper-bounds are those of Andoni et al. (2014), Yehudai and Shamir (2019), Bach (2017), and Ji et al. (2019), all of which prove approximation-theoretic results by representing a target function as the expected value of weighted activation functions drawn from some distribution.

- Theorem 3.1 of Andoni et al. (2014) shows how neural networks with complex-valued weights and exponential activation functions can approximate polynomials of bounded degree. Their

[1] Our discussion here focuses on works that give non-asymptotic bounds. Pinkus (1999, Section 6) gives a review of asymptotic rates of approximation by neural networks of width r as $r \to \infty$ (regarding the dimension d as fixed).

bounds have an exponential dependence on that degree, which translates to an exponential dependence on the Lipschitz constant L even for constant dimension d; in contrast, our bounds are exponential in $\min\{d, L^2/\epsilon^2\}$, which can be much better if d is small.

- Yehudai and Shamir (2019) study depth-2 RBL ReLU networks (as we do), but like Andoni et al. (2014) focus on approximating polynomials of bounded degree. Since they consider a more stringent notion of L_∞-approximation (over the unit ball), their upper-bounds on network width (see their Theorems 3.3 and 3.4) are more pessimistic than ours and depend exponentially on the square of the polynomial degree.
- Proposition 3 of Bach (2017) and Theorem E.1 of Ji et al. (2019) imply (or directly give) upper-bounds on the width of depth-2 RBL ReLU networks (or finite-width NTK) to approximate Lipschitz functions. Similar to Yehudai and Shamir (2019), they consider an L_∞ notion of approximation, so they obtain upper-bounds that always are exponential in the dimension d.

Lower-bounds. A number of recent and classical papers give width lower-bounds for arbitrary (non-RBL) depth-2 networks that approximate certain types of multivariate functions. Maiorov (1999) gives asymptotically tight upper- and lower-bounds on the error in approximating functions from a Sobolev class achieveable by any two-layer network of a given width. The asymptotic nature of Maiorov's results (and proof techniques) means that the results do not imply lower-bounds on the network width required to achieve a given error rate ϵ unless ϵ is sufficiently small, possibly as a function of dimension. Our results differs from Maiorov's and other related results from the approximation theory literature by elucidating the interplay between the dimension and the error in both upper- and lower-bounds.

More recently, Eldan and Shamir (2016) and Safran and Shamir (2017) give $\exp(d)$-type lower-bounds on the width that depth-2 networks require to L_2-approximate certain simple functions under certain probability measures on $\mathbb{R}^d$. In Eldan and Shamir (2016) the function being approximated is not explicit, and in Safran and Shamir (2017) the lower-bound is only for very high-accuracy approximation (to error at most $1/d^4$). In both works the relevant probability measures are rather involved. In contrast, our lower bounds hold only for depth-2 RBL networks, but they are for simple explicit functions, for large (constant) values of the approximation parameter, and for L_2-approximation with respect to the uniform distribution over $[-1,1]^d$. In other relevant work on depth-2 lower-bounds, Martens et al. (2013) and Daniely (2017) give $\exp(d)$-type (or better) width lower bounds for depth-2 networks approximating certain functions with large Lipschitz constants, but these lower-bounds require a weight bound on the top-level combining gate. In contrast, our lower bonds for RBL networks have no restrictions on the weights of the top-level gate.

The work of Sonoda et al. (2020), which analyzes limitations on the approximation abilities of two-layer networks of random ReLU activation functions, is relevant to our lower-bounds. Their lower-bounds are independent of the width of the network; they give functions that cannot be approximated by RBL networks of *any* (potentially infinite) width. However, their lower-bounds are for an extremely strong notion of approximation, namely L_2 approximation over all of $\mathbb{R}^d$ (without any weighting by a probability distribution).

Our lower-bound idea of exploiting symmetry to obtain an *explicit* function that is difficult to approximate was inspired by Yehudai and Shamir (2019). Our approach for non-explicit lower bounds is quite similar to Theorem 19 of Kamath et al. (2020), which bounds the dimension of the space of all linear combinations of feature functions; similar to the lower-bound of Kamath et al.

(2020) (but unlike Yehudai and Shamir (2019)), our lower-bounds hold regardless of the size of the weights used in the linear combination of the bottom-level random features.

Finally, we remark that while we do not consider networks of depth larger than two, our paper was in large part inspired by results from the literature on depth separation. Telgarsky (2016), Eldan and Shamir (2016), and Daniely (2017) all prove lower-bounds by constructing highly oscillatory functions and showing that shallow networks must be wide in order to approximate these functions. Safran et al. (2019) prove lower-bounds on 1-Lipschitz functions that are non-oscillatory, such as $x \mapsto \max\{0, -\|x\| + 1\}$; however, these bounds only hold in the high-accuracy regime with small ϵ. These works motivated us to directly study the relationship between the Lipschitz constant of a target function and the width needed to approximate it.

2. Preliminaries

2.1. Notations

For a positive integer $d \in \mathbb{Z}^+$, let $[d] := \{1, 2, \ldots, d\}$. The vectors $\vec{0} := (0, \ldots, 0) \in \mathbb{R}^d$ and $\vec{1} := (1, \ldots, 1) \in \mathbb{R}^d$ are, respectively, the all-zeros and all-ones vectors. Let $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^d$. Let $\|f\|_{\text{Lip}}$ denote the Lipschitz constant of $f \colon \mathbb{R}^d \to \mathbb{R}$ with respect to the Euclidean metric (i.e., the least L s.t. f is L-Lipschitz w.r.t. $\|\cdot\|_2$).

We use the following notations for a multi-index $K \in \mathbb{N}^d$ (where $\mathbb{N} := \{z \in \mathbb{Z} : z \geq 0\}$). Let $|K| := \sum_{i=1}^d K_i$, $\|K\|_2 := (\sum_{i=1}^d K_i^2)^{1/2}$, and $K! := \prod_{i=1}^d (K_i!)$. Let $x^K := \prod_{i=1}^d x_i^{K_i}$ for $x \in \mathbb{R}^d$. Lastly, let $D^{(K)} f$ be the order-$|K|$ partial derivative of a function $f(x)$ with respect to x^K.

We use bold font to denote random variables and write "$\mathbf{x} \sim \mathcal{D}$" to indicate that random variable $\mathbf{x}$ is distributed according to distribution $\mathcal{D}$.

We use $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean inner product in $\mathbb{R}^d$ (and occasionally regard multi-indices $K \in \mathbb{N}^d$ as elements of $\mathbb{R}^d$). For a probability measure μ on $\mathbb{R}^d$, $L_2(\mu)$ denotes the space of square-integrable functions with inner product denoted by $\langle f, g \rangle_\mu := \mathbb{E}_{\mathbf{x} \sim \mu}[f(\mathbf{x}) g(\mathbf{x})] = \int_{\mathbb{R}^d} f(x) g(x) \mu(\mathrm{d}x)$. Many of our results concern the uniform probability measure on $[-1, 1]^d$. In these cases, we use the notations $L_2([-1, 1]^d)$ and $\langle \cdot, \cdot \rangle_{[-1,1]^d}$, and fix a particular orthonormal basis $\mathcal{T} = \{T_K : K \in \mathbb{Z}^d\}$ for $L_2([-1, 1]^d)$ based on trigonometric polynomials. See Appendix A for details. We also consider certain finite-dimensional subspaces of $L_2([-1, 1]^d)$ which are spanned by a set of functions indexed by $\mathcal{K}_{k,d} := \{K \in \mathbb{Z}^d : \|K\|_2 \leq k\}$. The dimensions $Q_{k,d} := |\mathcal{K}_{k,d}|$ of these subspaces are upper- and lower-bounded as follows (proof also given in Appendix A).

Fact 3 *For all $d \in \mathbb{Z}^+$ and $k \geq 1$, $Q_{k,d} = \exp\left(\Theta\left(\min\left(d \log\left(\frac{k^2}{d} + 2\right), k^2 \log\left(\frac{d}{k^2} + 2\right)\right)\right)\right)$.*

2.2. Random bottom layer neural network approximation

Throughout the paper, we treat a depth-2 random bottom layer (RBL) ReLU network as a random features model. The upper-bounds in this paper demonstrate the representational powers of linear combinations of these random features, while the lower-bounds demonstrate their limitations.

We define a family of distributions over the parameters of random ReLU activations. Note that our lower-bounds in Theorems 10, 13, 39, and 41 hold for *all* such distributions $\mathcal{D}$, while our upper-bounds in Theorems 6 and 35 hold for some fixed $\mathcal{D}$, which depends on an upper bound on the Lipschitz norm of the target function but not on the target function itself.

Definition 4 (Symmetric ReLU parameter distributions) *A product distribution $\mathcal{D} := \mathcal{D}_{\text{bias}} \times \mathcal{D}_{\text{weights}}$ over $\mathbb{R} \times \mathbb{S}^{d-1}$ is* a symmetric ReLU parameter distribution *if the coordinates of $\mathcal{D}_{\text{weights}}$ are invariant to permutation. That is, $\mathcal{D}_{\text{weights}} = \pi \circ \mathcal{D}_{\text{weights}}$ for any permutation π of $[d]$.*

Given a distribution over random ReLU parameters, we now introduce the full random ReLU features model. We define a notion of approximation and formalize the *minimum width* of the network (or the minimum number of random features to combine) needed to obtain a sufficiently accurate approximation with high probability.

Definition 5 (Minimum-width RBL ReLU network approximation) *Consider a symmetric ReLU parameter distribution $\mathcal{D}$, a measure μ over $\mathbb{R}^d$, and a network width $r \in \mathbb{Z}^+$. For all $i \in [r]$, we draw each* random network feature *$\mathbf{g}^{(i)} \in L_2(\mu)$ independently by drawing $(\mathbf{b}^{(i)}, \mathbf{w}^{(i)})$ from $\mathcal{D}$ and letting $\mathbf{g}^{(i)}(x) := \sigma_{\text{ReLU}}(\langle \mathbf{w}^{(i)}, x\rangle - \mathbf{b}^{(i)})$.*

Given $\epsilon, \delta > 0$ and a function $f : \mathbb{R}^d \to \mathbb{R}$ with bounded $\|f\|_\mu$, we define $\mathrm{MinWidth}_{f,\epsilon,\delta,\mu,\mathcal{D}}$ to be the smallest $r \in \mathbb{Z}^+$ such that the following holds: With probability at least $1 - \delta$ over $\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}$,

$$\inf_{g \in \mathrm{Span}\left(\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}\right)} \|f - g\|_\mu \le \epsilon.$$

3. Upper-bounds for Lipschitz functions in $L_2([-1,1]^d)$

Our upper-bounds on the minimum width RBL ReLU network that approximates a Lipschitz function are dominated by the quantity $Q_{k,d}$, which represents the number of integer points contained in a d-dimensional ball of radius k (see Section 2.1).

Theorem 6 (Formal version of Theorem 1: Upper-bound for L-Lipschitz functions) *Fix some $\delta \in (0, \frac{1}{2}]$ and $\epsilon, L > 0$ with $\frac{L}{\epsilon} \ge 2$. Then, there exists some symmetric ReLU parameter distribution $\mathcal{D}$ such that for any $f \in L_2([-1,1]^d)$ with $\|f\|_{\text{Lip}} \le L$ and $|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \le L$,*

$$\mathrm{MinWidth}_{f,\epsilon,\delta,[-1,1]^d,\mathcal{D}} \le O\left(\frac{L^6 d^2}{\epsilon^6} \ln\left(\frac{1}{\delta}\right) Q^2_{2L/\epsilon,d}\right).$$

Applying the asymptotics of $Q_{k,d}$ from Fact 3 reveals that the minimum width can also be bounded by the term in Theorem 1. That expression shows that the minimum width is polynomial in $\frac{L}{\epsilon}$ when d is a fixed constant, and polynomial in d when $\frac{L}{\epsilon}$ is a fixed constant.

To prove Theorem 6, we break the process of approximating a Lipschitz function f with an RBL ReLU network into two steps. We first approximate f with a bounded-degree trigonometric polynomial P in Lemma 7 and then approximate P with an RBL ReLU network in Lemma 9. We state the lemmas and discuss their proofs in Sections 3.1 and 3.2 respectively. Section 3.3 gives a formal proof of Theorem 6.

In Appendix D.1, we present and prove Theorem 35, a parallel result to Theorem 6 that instead considers the approximation of some function f that has a bounded Sobolev norm and which (along with its derivatives) satisfies periodic boundary conditions. The proof of Theorem 35 only differs from that of Theorem 6 by obtaining a trigonometric polynomial approximation for f from Lemma 38 (stated and proved in Appendix D.1) rather than Lemma 7.

3.1. Approximating Lipschitz functions with bounded-degree trigonometric polynomials

Lemma 7 *Fix some $L, \epsilon > 0$ with $\frac{L}{\epsilon} \geq 1$ and consider any function $f \in L^2([-1,1]^d)$ with $\|f\|_{\text{Lip}} \leq L$ and $|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \leq L$. Then, taking $k = \frac{L}{\epsilon}$, there exists a bounded-degree trigonometric polynomial*

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K\left(\frac{x}{2}\right)$$

such that $\|f - P\|_{[-1,1]^d} \leq \epsilon$. Moreover, $|\beta_K| \leq L$ for all K.

We formally prove this lemma (which we restate as Lemma 22) in Appendix B.1. Here we highlight a central part of the argument (used in the full proof) by stating and proving a special case of the lemma which additionally requires that f satisfy periodic boundary conditions.

Lemma 8 (Approximating Lipschitz functions with periodic boundary conditions) *Fix some $L, \epsilon > 0$ with $\frac{L}{\epsilon} \geq 2$. Consider any function $f \in L^2([-1,1]^d)$ such that f satisfies periodic boundary conditions, $\|f\|_{\text{Lip}} \leq L$, and $|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \leq \frac{L}{2}$. Then, taking $k = \frac{L}{2\epsilon}$, there exists a bounded-degree trigonometric polynomial*

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K(x)$$

such that $\|f - P\|_{[-1,1]^d} \leq \epsilon$. Moreover, $|\beta_K| \leq \frac{L}{2}$ for all K.

To prove Lemma 8, we consider the representation of f as an infinite linear combination of trigonometric basis elements from $\mathcal{T}$. We show that f can only be L-Lipschitz if all high-degree terms of this representation have vanishingly small coefficients. This requires the term-by-term differentiation of the trigonometric representation of f, which is possible due to its periodic boundary conditions (see Lemma 20 in Appendix A).

Proof. By appealing to a standard approximation argument (e.g., Folland, 1999, Proposition 8.17), we may assume that f is differentiable. Because $\mathcal{T}$ is an orthonormal basis over $L_2([-1,1]^d)$, we can express f as

$$f(x) = \sum_{K \in \mathbb{Z}^d} \alpha_K T_K(x).$$

The condition $\|f\|_{\text{Lip}} \leq L$ implies that $\|\nabla f(x)\|_2 \leq L$ for all $x \in [-1,1]^d$. Because f has periodic boundary conditions, f is differentiable, and $\partial f(x)/\partial x_i \in L_2([-1,1]^d)$ for all i, Lemma 20 can be applied to relate L to the coefficients $(\alpha_K)_{K \in \mathbb{Z}^d}$:

$$L^2 \geq \mathop{\mathbb{E}}_{\mathbf{x} \sim [-1,1]^d}\left[\|\nabla f(\mathbf{x})\|_2^2\right] = \sum_{i=1}^d \mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}\right)^2\right] = \sum_{i=1}^d \mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\sum_{K \in \mathbb{Z}^d} \alpha_K \frac{\partial T_K(\mathbf{x})}{\partial \mathbf{x}_i}\right)^2\right] \tag{1}$$

$$= \sum_{i=1}^d \sum_{K \in \mathbb{Z}^d} \alpha_K^2 \left\|\frac{\partial T_K}{\partial x_i}\right\|_{[-1,1]^d}^2 + 2\sum_{i=1}^d \sum_{K \in \mathbb{Z}^d} \sum_{K' \neq K} \alpha_K \alpha_{K'} \left\langle \frac{\partial T_K}{\partial x_i}, \frac{\partial T_{K'}}{\partial x_i} \right\rangle_{[-1,1]^d}$$

$$= \sum_{i=1}^d \sum_{K \in \mathbb{Z}^d} \alpha_K^2 \pi^2 K_i^2 = \pi^2 \sum_K \alpha_K^2 \|K\|_2^2 . \tag{2}$$

Equations (1) and (2) follow from Lemma 20 and Fact 18 respectively. An immediate consequence of the above inequality is that $|\alpha_K| \leq L/\pi \leq L/2$ as long as $K \neq \vec{0}$. Because $|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \leq L/2$, $|\alpha_{\vec{0}}| \leq L/2$ as well. We define the trigonometric polynomial $P = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K$ by letting $\beta_K := \alpha_K$ for all K with $\|K\|_2 \leq k$. Parseval's identity (Fact 15) and the inequality ending on line (2) guarantee that

$$\begin{aligned}\|f - P\|^2_{[-1,1]^d} &= \sum_{K \in \mathbb{Z}^d \setminus \mathcal{K}_{k,d}} \alpha_K^2 \leq \sum_{K \in \mathbb{Z}^d \setminus \mathcal{K}_{k,d}} \alpha_K^2 \cdot \frac{\|K\|_2^2}{k^2} \leq \frac{1}{k^2} \sum_{K \in \mathbb{Z}^d} \alpha_K^2 \|K\|_2^2 \\ &\leq \frac{L^2}{\pi^2 k^2} \leq \frac{L^2}{2^2 k^2} = \epsilon^2. \qquad \blacksquare\end{aligned}$$

The proof of Lemma 7 is a reduction to Lemma 8. Instead of approximating f with a low-degree trigonometric polynomial, we approximate $\tilde{f}$, a scaled, shifted, and reflected version of f that has periodic boundary conditions and thus can be differentiated term-by-term. The bulk of the proof involves transforming f into $\tilde{f}$ and transforming $\tilde{P}$ (the trigonometric polynomial obtained by applying Lemma 8 to $\tilde{f}$) back into P. This scaling and reflection argument is why we approximate f with combinations of trigonometric polynomials of the form $T_K(x/2)$, rather than $T_K(x)$.

3.2. Approximating bounded-degree trigonometric polynomials with RBL ReLU nets

Lemma 9 *Fix some $\delta \in (0, 1/2]$, $\epsilon > 0$, $\rho \in (0, 1]$, $k \geq 1$, and $d \in \mathbb{Z}^+$. Then, there exists some symmetric ReLU parameter distribution $\mathcal{D}_k$ such that for any trigonometric polynomial*

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K(\rho x)$$

with $|\beta_K| \leq \beta_{\max}$ for all $K \in \mathcal{K}_{k,d}$,

$$\mathrm{MinWidth}_{P,\epsilon,\delta,[-1,1]^d,\mathcal{D}_k} \leq O\left(\frac{\beta_{\max}^2 d^2 k^4}{\epsilon^2} Q_{k,d}^2 \ln\left(\frac{1}{\delta}\right)\right).$$

We prove this lemma in Appendix B.2 as Lemma 23. We take advantage of the fact that every low-degree trigonometric polynomial can be expressed as a linear combination of ridge functions. As shown in Lemma 25, each of those ridge functions can in turn be represented as an infinite mixture of ReLUs. We then represent the entire trigonometric polynomial as an expectation over weighted random ReLU features with parameters drawn from a symmetric ReLU parameter distribution $\mathcal{D}_k$ (Definition 24). By bounding the maximum norm of every random ReLU drawn from $\mathcal{D}_k$, a concentration bound (Lemma 26) can show that this expectation can be closely approximated with a sufficiently large finite linear combination of randomly sampled ReLUs.

3.3. Proof of Theorem 6

Consider any $f \in L_2([-1,1]^d)$ with $\|f\|_{\mathrm{Lip}} \leq L$ and $|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \leq L$. By Lemma 7, there exists a bounded-degree trigonometric polynomial $P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K(x/2)$ with $k := 2L/\epsilon$ and

$|\beta_K| \le L$ for all $K \in \mathcal{K}_{k,d}$, such that $\|f - P\|_{[-1,1]^d} \le \epsilon/2$. By applying Lemma 9 to P with $\rho = 1/2$,

$$\mathrm{MinWidth}_{P,\epsilon/2,\delta,[-1,1]^d,\mathcal{D}_k} \le O\left(\frac{\beta_{\max}^2 d^2 k^4}{\epsilon^2} Q_{k,d}^2 \ln\left(\frac{1}{\delta}\right)\right) \le O\left(\frac{d^2 L^6}{\epsilon^6} Q_{2L/\epsilon,d}^2 \ln\left(\frac{1}{\delta}\right)\right).$$

Thus (see Definition 5) there exists an RBL ReLU network g of width $\mathrm{MinWidth}_{P,\epsilon/2,\delta,[-1,1]^d,\mathcal{D}_k}$ such that $\|P - g\|_{[-1,1]^d} \le \epsilon/2$. By the triangle inequality, $\|f - g\|_{[-1,1]^d} \le \epsilon$. We conclude that

$$\mathrm{MinWidth}_{f,\epsilon,\delta,[-1,1]^d,\mathcal{D}_k} = O\left(\frac{d^2 L^6}{\epsilon^6} Q_{2L/\epsilon,d}^2 \ln\left(\frac{1}{\delta}\right)\right). \qquad \blacksquare$$

4. Lower-bounds for Lipschitz functions in $L_2([-1,1]^d)$

We give lower-bounds on the minimum width needed to ϵ-approximate L-Lipschitz functions using depth-2 RBL ReLU networks. Below we present a formal statement of Theorem 2, which shows that a particular family of "simple" functions must contain some hard-to-approximate function. Like the upper-bounds in Section 3, the minimum width is polynomial (in fact linear) in the quantity $Q_{k,d}$, where $k = \Theta(L/\epsilon)$.

Theorem 10 (Formal version of Theorem 2: Lower-bound for L-Lipschitz functions) *Fix any $\epsilon, L > 0$ and fix any symmetric ReLU parameter distribution $\mathcal{D}$. Then, there exists some multi-index $K \in \mathbb{N}^d$ with $\|K\|_2 \le L/18\epsilon$ such that the function $f(x) := 4\epsilon T_K$ (recall that $T_K \in \mathcal{T}$) satisfies $\|f\|_{\mathrm{Lip}} \le L$ and*

$$\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},[-1,1]^d,\mathcal{D}} \ge \frac{1}{4} Q_{L/18\epsilon,d}.$$

The informal version, Theorem 2, follows by applying Fact 3 to lower-bound $Q_{k,d}$. We note that the function f used in the lower-bound aligns nicely with the approximation techniques from Section 3 because f is *(i)* a ridge function and *(ii)* a scalar multiple of a sinusoidal function from the trigonometric basis $\mathcal{T}$.

We prove Theorem 10 in stages by proving a sequence of claims which are successively more closely tailored to our RBL ReLU model.

1. In Appendix C.1 we state and prove Theorem 11, which gives a general result about the limitations of linear combinations of r random features. This theorem states that a large fraction of any set of N orthonormal functions must be inapproximable by linear combinations of r random features when $N \gg r$. We state a simplified version of the theorem below:

   Theorem 11 (Simplification of Theorem 29) *Let $\Phi = \{\varphi_1, \dots, \varphi_N\} \subset L_2(\mu)$ be a family of N functions such that $\langle \varphi_i, \varphi_{i'} \rangle_\mu = \mathbb{1}\{i = i'\}$. Let $\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}$ be i.i.d. copies of an $L_2(\mu)$-valued random variable. Then, there exists some $\varphi_i \in \Phi$ such that*

   $$\mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}}\left[\inf_{g \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2\right] \ge 1 - \frac{r}{N}.$$

   The proof hinges on an intuitive linear algebraic fact generalized to function spaces: N orthogonal vectors cannot all be close to the span of r vectors when $N \gg r$. It does so by applying the

Hilbert Projection Theorem (Fact 30). The full generality of Theorem 29 also includes function families Φ that are "nearly orthonormal" rather than strictly orthonormal (this generalization is useful for extending our results to Gaussian space, as discussed in Appendix E). It also proves the inapproximability of some explicit function φ_1 when the family Φ satisfies a suitable notion of symmetry relative to $\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}$.

2. Lemma 32 of Appendix C.2 adapts Theorem 29 to our random ReLU features by giving a lower-bound on the minimum width RBL network needed to ϵ-approximate some function for any $\epsilon > 0$. Below is a simplified version of the lemma that is restricted to orthonormal function families, considers only the uniform measure over $[-1,1]^d$, and omits the special "symmetric case" for Φ.

   Lemma 12 (Simplification of Lemma 32) *Let $\mathcal{D}$ be a symmetric ReLU parameter distribution. Fix any $\Phi = \{\varphi_1, \dots, \varphi_N\} \subset L_2([-1,1]^d)$ such that $\langle \varphi_i, \varphi_{i'} \rangle_{[-1,1]^d} = \mathbb{1}\{i = i'\}$. Then, for any $\epsilon > 0$, there exists some $\varphi_i \in \Phi$ such that $\mathrm{MinWidth}_{4\epsilon\varphi_i,\epsilon,1/2,[-1,1]^d,\mathcal{D}} \geq N/4$.*

   The proof combines a scaling argument with the definition of $\mathrm{MinWidth}$ to provide lower-bounds for any choice of the error parameter ϵ.

3. We conclude the proof of Theorem 10 in Appendix C.3. Lemma 33 shows the existence of a low-degree element of the sinusoidal basis $\mathcal{T}$ that cannot be approximated over $[-1,1]^d$ by an RBL ReLU network of small width. It does so by defining the orthonormal family of functions to be $\Phi := \{T_K \in \mathcal{T} : K \in \mathcal{K}_{k,d}\}$ and invoking Lemma 32. The proof of Theorem 10 only requires applying Lemma 33 for some $k = \Theta(L/\epsilon)$ and showing that all $T_K \in \Phi$ have $\|T_K\|_{\mathrm{Lip}} \leq L$.

   Lemma 33 also yields an immediate proof of Theorem 39, the Sobolev analogue of Theorem 10, in Appendix D.2. Theorem 39 uses the same function family Φ, but must bound the Sobolev norm of all functions in Φ rather than the Lipschitz constant.

The lower-bound established in Theorem 10 is non-explicit; it guarantees the existence of some inapproximable function in $\mathcal{T}$, but does not by itself let us deduce the specific identity of a hard function. Since it is desirable to have a lower-bound for a fully explicit function, we also give a variant that achieves this goal at only a small cost in the resulting quantitative lower-bound:

Theorem 13 (Explicit lower-bound for an L-Lipschitz function) *For some $\epsilon, L > 0$, let $\ell := \min(\lceil d/2 \rceil, \lfloor L^2/32\pi^2\epsilon^2 \rfloor)$. Fix any symmetric ReLU parameter distribution $\mathcal{D}$. Then the function $f(x) := 4\sqrt{2}\epsilon \sin(\pi \sum_{i=1}^{\ell} x_i)$ satisfies $\|f\|_{\mathrm{Lip}} \leq L$ and*

$$\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},[-1,1]^d,\mathcal{D}} \geq \frac{1}{4}\binom{d}{\ell} \geq \exp\left(\Omega\left(\min\left(\frac{L^2}{\epsilon^2}\log\left(\frac{d\epsilon^2}{L^2}+2\right), d\right)\right)\right).$$

Comparing the quantitative lower-bounds of Theorem 10 and Theorem 13, we see that the latter is weaker only by a logarithmic factor in the exponent.

We prove the explicit lower-bound Theorem 13 in Appendix C.4. The only difference between the proofs of Theorems 10 and 13 is in the last step. Theorem 13 relies on Lemma 34, an analogue of Lemma 33, which invokes Lemma 32 with a different family Φ of trigonometric polynomials that are symmetric up to a permutation of variables. That is, for every $T_K, T_{K'} \in \Phi$, there exists some permutation π over $[d]$ such that $T_K = T_{K'} \circ \pi$. (Roughly speaking, the larger family of orthonormal functions used in the proof of Theorem 10 consists of functions of the form $\sin(\pi \langle K, x\rangle)$ where $K \in \mathbb{N}^d$ is only constrained by having $\|K\|$ satisfy some bound, whereas the smaller family of orthonormal functions used in the proof of Theorem 34 consists of functions of the form $\sin(\pi \langle K, x\rangle)$ where K is restricted to be a 0/1 vector of some specific Hamming weight. The latter family is easily seen to satisfy symmetry with respect to any permutation π of the d coordinates, whereas the former family does not satisfy such a symmetry condition.) This symmetry condition makes it easy to argue that all functions in the symmetric family Φ are "equally hard," from which a lower bound follows straightforwardly.

Finally, we mention that Lemma 34 also supports a proof of the inapproximability of an explicit function with bounded Sobolev norm; this is established in Theorem 41 of Appendix D.2.

Acknowledgments

In this work, D. Hsu is supported by NSF grants CCF-1740833, IIS-1563785, and a Sloan Research Fellowship. C. Sanford gratefully acknowledges CCF-1563155 and is partially supported by Google Faculty Research Award. R.A. Servedio is supported by NSF grants CCF-1814873, IIS-1838154, CCF-1563155, and by the Simons Collaboration on Algorithms and Geometry. E.V. Vlatakis-Gkaragkounis is grateful to be supported by NSF grants CCF-1703925, CCF-1763970, CCF-1814873, CCF-1563155, and by the Simons Collaboration on Algorithms and Geometry and by the Onassis Foundation under Scholarship ID: F ZN 010-1/2017-2018. Finally, the authors would like to thank Shivam Nadimpalli for the helpful discussions about the necessary conditions of Lemma 20. This material is based upon work supported by the National Science Foundation under grant numbers listed above. 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 National Science Foundation (NSF).

References

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, 2014.

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, May 1993.

Richard Bellman. Almost orthogonal series. *Bulletin of the American Mathematical Society*, 50: 517–519, 1944.

Ralph P. Boas, Jr. A general moment problem. *American Journal of Mathematics*, 63:361, 1941.

Emmanuel J. Candès. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis*, 6(2):197–218, 1999.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems 22*, 2009.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, 2017.

Harry Dym and Henry P. McKean. *Fourier Series and Integrals*. Academic Press, 1972.

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

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 1999.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, 2018.

Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2019.

Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Conference on Learning Theory*, 2020.

Jason M. Klusowski and Andrew R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12), Dec 2018.

Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

V.E Maiorov. On best approximation by ridge functions. *Journal of Approximation Theory*, 99 (1):68 – 94, 1999. ISSN 0021-9045. doi: https://doi.org/10.1006/jath.1998.3304. URL http://www.sciencedirect.com/science/article/pii/S0021904598933044.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks, 2021.

James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems 26*, 2013.

Noboru Murata. An integral representation of functions using three-layered networks and their approximation bounds. *Neural Networks*, 9(6):947–956, 1996.

Radford M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 2008.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, 2009.

Boris Rubin. The Calderón reproducing formula, windowed X-ray transforms, and radon transforms in L^p-spaces. *Journal of Fourier Analysis and Applications*, 4(2):175–197, 1998.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International Conference on Machine Learning*, 2017.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: What is actually being separated? In *Conference on Learning Theory*, 2019.

Sho Sonoda, Ming Li, Feilong Cao, Changqin Huang, and Yu Guang Wang. On the approximation lower bound for neural nets with random weights. *arXiv preprint arXiv:2008.08427*, 2020.

Yitong Sun, Anna Gilbert, and Ambuj Tewari. On the approximation properties of random ReLU features. *arXiv preprint arXiv:1810.04374*, 2018.

Gabor Szegö. *Orthogonal Polynomials*, volume XXIII of *Americam Mathematical Society Colloquium Publications*. A.M.S, Providence, 1989.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory*, 2016.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Vadim Vladimirovich Yurinskiĭ. Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis*, 6(4):473–499, 1976.

Appendix A. Key facts about trigonometric polynomial basis

In this appendix, we supplement Section 2.1 by introducing the family of trigonometric polynomials that we use in our proofs and by proving properties related to their orthonormality. We recall the definition of an orthonormal basis for the space $L_2(\mu)$:

Definition 14 (Orthonormal basis) *A countable set $\mathcal{G} \subset L_2(\mu)$ is an* orthonormal basis *for $L_2(\mu)$ if $\langle g, \tilde{g}\rangle_\mu = \mathbb{1}\{g = \tilde{g}\}$ for all $g, \tilde{g} \in \mathcal{G}$ and $\mathrm{Span}(\mathcal{G}) = L_2(\mu)$.*

We frequently apply the following standard facts about orthonormal bases:

Fact 15 (Facts about orthonormal bases) *For some measure μ, let $\mathcal{G}$ be an orthonormal basis for $L_2(\mu)$. For any $f, \tilde{f} \in L_2(\mu)$ we have that $f = \sum_{g \in \mathcal{G}} \alpha_g g$ and $\tilde{f} = \sum_{g \in \mathcal{G}} \beta_g g$ for some real $(\alpha_g)_{g \in \mathcal{G}}$ and $(\beta_g)_{g \in \mathcal{G}}$, and moreover*

- $\alpha_g = \langle f, g \rangle_\mu$;
- $\|f\|_\mu^2 = \sum_{g \in \mathcal{G}} \alpha_g^2$ *(Parseval); and*
- $\langle f, \tilde{f} \rangle_\mu = \sum_{g \in \mathcal{G}} \alpha_g \beta_g$ *(Plancherel).*

We define the basis of trigonometric polynomials $\mathcal{T}$ as

$$\mathcal{T} := \left\{ T_K : K \in \mathbb{Z}^d \right\},$$

where

$$T_K(x) := \begin{cases} 1 & K = \vec{0} \\ \sqrt{2} \sin\left(\pi \left\langle K, x \right\rangle\right) & K \in \mathcal{K}_{\text{sin}} \\ \sqrt{2} \cos\left(\pi \left\langle K, x \right\rangle\right) & K \in \mathcal{K}_{\text{cos}}, \end{cases} \tag{3}$$

and $\mathcal{K}_{\text{sin}}$ and $\mathcal{K}_{\text{cos}}$ form a partition of $\mathbb{Z}^d \setminus \{\vec{0}\}$[2] and are defined as

$$\mathcal{K}_{\text{sin}} := \left\{ K \in \mathbb{Z}^d \setminus \{\vec{0}\} : K_i > 0, \text{ where } i = \min\left\{ j \in [d] : x_j \neq 0 \right\} \right\},$$
$$\mathcal{K}_{\text{cos}} := \left\{ K \in \mathbb{Z}^d \setminus \{\vec{0}\} : K_i < 0, \text{ where } i = \min\left\{ j \in [d] : x_j \neq 0 \right\} \right\}.$$

The set $\mathcal{T}$ is a useful family of functions for both our upper- and our lower-bounds on the minimum width RBL ReLU network needed to approximate Lipschitz functions. The fact that $\mathcal{T}$ is an orthonormal basis for $L_2([-1,1]^d)$ (Fact 17) permits us to express other functions in $L_2([-1,1]^d)$ as a linear combination of the elements of $\mathcal{T}$. As we show in Fact 18, those orthogonality properties of the elements of $\mathcal{T}$ are maintained even after taking partial derivatives. In addition, every function in $\mathcal{T}$ is a ridge function (that is, $T_K(x) = \phi_K(\langle K, x \rangle)$ for some $\phi_K : \mathbb{R} \to \mathbb{R}$), which, as we will see later, means (very usefully for us) that T_K is easily approximated by linear combinations of shifted ReLUs. Finally, the Lipschitz constant of all functions in $\mathcal{T}$ is bounded: $\|T_K\|_{\text{Lip}} \leq \sqrt{2}\pi \|K\|_2$.

To prove that $\mathcal{T}$ is orthogonal, we rely on the following fact from integral calculus.

Fact 16 (Integrals of multivariate sinusoids) *For each $K \in \mathbb{Z}^d$,*

$$\int_{[-1,1]^d} \cos\left(\pi \left\langle K, x \right\rangle\right) \mathrm{d}x = 2^d \cdot \mathbb{1}\{K = \vec{0}\} \quad \& \quad \int_{[-1,1]^d} \sin\left(\pi \left\langle K, x \right\rangle\right) \mathrm{d}x = 0.$$

Proof. We use a simple inductive argument on d to evaluate the first integral. The base case $d = 1$ is straightforward, so assume $d > 1$ and define $x_{-1} = (x_2, \ldots, x_d) \in \mathbb{R}^{d-1}$ for any $x \in \mathbb{R}^d$. Assume inductively that

$$\int_{[-1,1]^{d-1}} \cos\left(\pi \left\langle K_{-1}, x_{-1} \right\rangle\right) \mathrm{d}x_{-1} = 2^{d-1} \mathbb{1}\{K_{-1} = \vec{0}\}.$$

[2] Note that this partition of $\mathbb{Z}^d - \{\vec{0}\}$ is an arbitrary one. The only property this partition is designed to satisfy is that if K corresponds to $\sin(\pi\langle K, x\rangle)$, then $-K$ must correspond to $\cos(-\pi\langle K, x\rangle)$ (and vice versa).

By the cosine addition formula, we have that:

$$\begin{aligned}
&\int_{[-1,1]^d} \cos\left(\pi\langle K, x\rangle\right) \mathrm{d}x \\
&= \int_{[-1,1]^d} \left[\cos\left(\pi K_1 x_1\right) \cos\left(\pi\langle K_{-1}, x_{-1}\rangle\right) - \sin\left(\pi K_1 x_1\right) \sin\left(\pi\langle K_{-1}, x_{-1}\rangle\right)\right] \mathrm{d}x \\
&= \left[\int_{-1}^{1} \cos\left(\pi K_1 x_1\right) \mathrm{d}x_1\right] \left[\int_{[-1,1]^{d-1}} \cos\left(\pi \left\langle K_{-1}, x_{-1}\right\rangle\right) \mathrm{d}x_{-1}\right] \\
&\qquad - \left[\int_{-1}^{1} \sin\left(\pi K_1 x_1\right) \mathrm{d}x_1\right] \left[\int_{[-1,1]^{d-1}} \sin\left(\pi \left\langle K_{-1}, x_{-1}\right\rangle\right) \mathrm{d}x_{-1}\right] \\
&= 2 \cdot \mathbb{1}\{K_1 = 0\} \left[\int_{[-1,1]^{d-1}} \cos\left(\pi \left\langle K_{-1}, x_{-1}\right\rangle\right) \mathrm{d}x_{-1}\right] = 2^d \cdot \mathbb{1}\{K = \vec{0}\}.
\end{aligned}$$

The second claim follows by a nearly identical inductive argument, which we omit. ■

Fact 17 *$\mathcal{T}$ is an orthonormal basis for $L_2([-1,1]^d)$.*

Proof. First, we make use of the well-known fact that the constant 1 function, along with $z \mapsto \sqrt{2}\sin(\pi k z)$ and $z \mapsto \sqrt{2}\cos(\pi k z)$ for all $k \in \mathbb{Z}^+$, collectively form an orthonormal basis for $L_2([-1,1])$. (For details, see Dym and McKean, 1972.) Thus, the d-fold Cartesian product of this collection is an orthonormal basis for $L_2([-1,1]^d)$.[3] Each function in this basis is a product of d functions—one per variable, and each being either a constant, sine, or cosine as above—and can be rewritten as a linear combination of functions from $\mathcal{T}$ using basic product-to-sum trigonometric identities. Thus, $\mathrm{Span}\left(\mathcal{T}\right) = L_2([-1,1]^d)$.

To complete our proof, it remains to show that all elements of $\mathcal{T}$ are orthogonal and have unit norm. It suffices to show that $\langle T_K, T_{K'}\rangle_{[-1,1]^d} = \mathbb{1}\{K = K'\}$ for all $K, K' \in \mathbb{Z}^d$. There are six possible scenarios for this claim depending on which partitioning subsets of $\mathbb{Z}^d$ contain K and K': *(1)* $K, K' \in \mathcal{K}_{\text{cos}}$; *(2)* $K, K' \in \mathcal{K}_{\text{sin}}$; *(3)* $K = K' = \vec{0}$; *(4)* $K \in \mathcal{K}_{\text{cos}}, K' = \vec{0}$ or $K = \vec{0}, K' \in \mathcal{K}_{\text{cos}}$; *(5)* $K \in \mathcal{K}_{\text{sin}}, K' = \vec{0}$ or $K = \vec{0}, K' \in \mathcal{K}_{\text{sin}}$; and *(6)* $K \in \mathcal{K}_{\text{sin}}, K' \in \mathcal{K}_{\text{cos}}$ or $K \in \mathcal{K}_{\text{cos}}, K' \in \mathcal{K}_{\text{sin}}$. For the sake of simplicity, we only explicitly prove the claim for scenario *(1)*. The other cases can be proved with similar trigonometric arguments, all of which involve applying Fact 16. For scenario *(1)*, we observe that

$$\begin{aligned}
\langle T_K, T_{K'}\rangle_{[-1,1]^d} &= \frac{1}{2^d} \int_{[-1,1]^d} 2\cos\left(\pi \left\langle K, x\right\rangle\right) \cos\left(\pi \left\langle K', x\right\rangle\right) \mathrm{d}x \\
&= \frac{1}{2^d} \int_{[-1,1]^d} \left[\cos\left(\pi \left\langle K - K', x\right\rangle\right) - \cos\left(\pi \left\langle K + K', x\right\rangle\right)\right] \mathrm{d}x \\
&= \frac{1}{2^d} \left[2^d \mathbb{1}\left\{K - K' = 0\right\} - 2^d \mathbb{1}\left\{K + K' = 0\right\}\right] \\
&= \mathbb{1}\left\{K = K'\right\}.
\end{aligned}$$

[3] This is also an orthonormal basis, so we could similarly represent functions in $L_2([-1,1]^d)$ as linear combinations of the elements of this basis and apply the properties of Fact 15. However, this representation is unhelpful for our analysis because its elements have large Lipschitz constants and are not ridge functions.

The last equality holds because if $K + K' = 0$, then either K or K' must belong to $\mathcal{K}_{\sin}$ by the definitions of $\mathcal{K}_{\sin}$ and $\mathcal{K}_{\cos}$. ■

We additionally derive the following useful fact about the partial derivatives of elements of the trigonometric basis $\mathcal{T}$.

Fact 18 (Orthogonality of derivatives of $\mathcal{T}$) *For all $M \in \mathbb{N}^d$ and for all $K, K' \in \mathbb{Z}^d$,*

$$\left\langle D^{(M)} T_K, D^{(M)} T_{K'} \right\rangle_{[-1,1]^d} = \mathbb{1}\left\{K = K'\right\} \pi^{2|M|} K^{2M}.$$

Proof. The partial derivatives of T_K for every $\mathcal{K} \in \mathbb{Z}^d$ can be exactly characterized by inductively taking derivatives of sin and cos functions:

$$D^{(M)} T_K(x) = \begin{cases} \pi^{|M|} T_K(x) K^M & |M| \equiv 0 \pmod 4 \\ \pi^{|M|} T_{-K}(x) K^M & |M| \equiv 1 \pmod 4 \ \&\ K \in \mathcal{K}_{\sin} \\ -\pi^{|M|} T_{-K}(x) K^M & |M| \equiv 1 \pmod 4 \ \&\ K \in \mathcal{K}_{\cos} \cup \{\vec{0}\} \\ -\pi^{|M|} T_K(x) K^M & |M| \equiv 2 \pmod 4 \\ -\pi^{|M|} T_{-K}(x) K^M & |M| \equiv 3 \pmod 4 \ \&\ K \in \mathcal{K}_{\sin} \\ \pi^{|M|} T_{-K}(x) K^M & |M| \equiv 3 \pmod 4 \ \&\ K \in \mathcal{K}_{\cos} \cup \{\vec{0}\}. \end{cases} \tag{4}$$

The conclusion follows by applying the orthonormality of trigonometric basis elements from Fact 17 to Equation (4). ■

To prove that a function $f \in L_2([-1,1]^d)$ can be represented by a linear combination of sufficiently many random ReLUs, we first show that f can be approximated by a low-degree trigonometric polynomial. To do so, we upper-bound the higher-order coefficients of the trigonometric expansion of f. Obtaining these bounds requires taking partial derivatives of f by differentiating term-by-term the trigonometric expansion of f. However, this is not always possible; for instance, if $f(x) = x_1$, the terms of the trigonometric expansion of $\partial f / \partial x_1$ do not correspond to the term-by-term derivatives of the expansion of f.[4] We define a notion of *boundary periodicity* that lets us perform term-by-term differentiation:

Definition 19 (Periodic boundary conditions) $f \in L_2([-1,1]^d)$ *satisfies the* periodic boundary conditions *if for all $i \in [d]$ and for all $x \in [-1,1]^d$*

$$f(x_1, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_d) = f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d).$$

Note that all basis elements in $\mathcal{T}$ satisfy the periodic boundary conditions. The next lemma gives sufficient conditions for term-by-term differentiation of a function's trigonometric representation.

Lemma 20 (Term-by-term differentiation of trigonometric basis representations) *Consider some $f \in L_2([-1,1]^d)$ and $i \in [d]$ such that f satisfies the periodic boundary conditions, f*

[4] Because $\partial f / \partial x_1 = 1$, its trigonometric expansion $\partial f / \partial x_1 = \sum_{K \in \mathbb{Z}^d} \beta_K T_K$ will have $\beta_K = \mathbb{1}\{K = \vec{0}\}$. Because $f = \sum_{K \in \mathbb{Z}^d} \alpha_K T_K$ will have $\alpha_K \neq 0$ for some $K \neq \vec{0}$, $\beta_K \neq 0$ if term-by-term differentiation were possible. Since this contradicts the expansion of $\partial f / \partial x_1$, term-by-term differentiation is impossible in this case.

is differentiable with respect to x_i, and $\partial f/\partial x_i \in L_2([-1,1]^d)$. Then, f and $\partial f/\partial x_i$ have trigonometric expansions of the form

$$f = \sum_{K \in \mathbb{Z}^d} \alpha_K T_K \qquad \& \qquad \frac{\partial f}{\partial x_i} = \sum_{K \in \mathbb{Z}^d} \beta_K T_K,$$

where their coefficients $(\alpha_K)_{K\in\mathbb{Z}^d}, (\beta_K)_{K\in\mathbb{Z}^d}$ are related as follows:

$$\beta_K = \begin{cases} \pi K_i \alpha_{-K} & K \in \mathcal{K}_{\cos} \\ -\pi K_i \alpha_{-K} & K \in \mathcal{K}_{\sin} \\ 0 & K = \vec{0}. \end{cases} \tag{5}$$

Therefore,

$$\frac{\partial f}{\partial x_i} = \sum_{K \in \mathbb{Z}^d} \alpha_K \frac{\partial T_K}{\partial x_i}.$$

Proof. Without loss of generality, let $i = 1$. Because each of f and $\partial f/\partial x_1$ is in $L_2([-1,1]^d)$, there exist α and β by Fact 17 such that f and $\partial f/\partial x_1$ are exactly represented by the expansions given in the lemma statement. It remains to show that (5) holds. We fix any $K \in \mathcal{K}_{\cos}$, where $T_K(x) = \sqrt{2}\cos(\pi\langle K, x\rangle)$ and $\partial T_K(x)/\partial x_1 = -\sqrt{2}\pi K_1 \sin(\pi\langle K, x\rangle)$. By Fact 15, each coefficient of the representation is an inner-product: $\alpha_K = \langle f, T_K\rangle_{[-1,1]^d}$ and $\beta_K = \langle \partial f/\partial x_1, T_K\rangle_{[-1,1]^d}$. Moreover, β_K is related to α_{-K}, as shown in the following:

$$\begin{aligned}
\beta_K = \left\langle \frac{\partial f}{\partial x_1}, T_K \right\rangle_{[-1,1]^d} &= \frac{\sqrt{2}}{2^d} \int_{[-1,1]^d} \frac{\partial f(x)}{\partial x_1} \cos(\pi\langle K, x\rangle)\,\mathrm{d}x \\
&= \frac{\sqrt{2}}{2^d} \int_{[-1,1]^{d-1}} \int_{-1}^{1} \frac{\partial f(x)}{\partial x_1} \cos(\pi\langle K, x\rangle)\,\mathrm{d}x_1\,\mathrm{d}x_{-1} \\
&= \frac{\sqrt{2}}{2^d} \int_{[-1,1]^{d-1}} \left[f(x)\cos(\pi\langle K, x\rangle)\Big|_{-1}^{1} + \int_{-1}^{1} f(x)\pi K_1 \sin(\pi\langle K, x\rangle)\,\mathrm{d}x_1 \right] \mathrm{d}x_{-1} && (6) \\
&= \frac{\sqrt{2}}{2^d} \int_{[-1,1]^d} f(x)\pi K_1 \sin(\pi\langle K, x\rangle)\,\mathrm{d}x = \pi K_1 \langle f, T_{-K}\rangle_{[-1,1]^d} = \pi K_1 \alpha_{-K}. && (7)
\end{aligned}$$

We integrate by parts for Equation (6) and take advantage of the periodic boundary conditions of f and T_K for Equation (7). A symmetric argument proves the claim for $K \in \mathcal{K}_{\sin}$. When $K = \vec{0}$, we repeat the above argument, and the periodic boundary conditions of f imply that $\beta_{\vec{0}} = 0$. ■

The subspaces of $L_2([-1,1]^d)$ of primary interest in our analysis are spanned by a set of orthonormal functions that are indexed by the integer lattice points contained in given Euclidean balls. The next fact upper- and lower-bounds the number of such points (and hence the dimension of such a subspace).

Fact 21 (Restatement of Fact 3) *For all $d \in \mathbb{Z}^+$ and $k \geq 1$,*

$$Q_{k,d} = \exp\left(\Theta\left(\min\left(d\log\left(\frac{k^2}{d}+2\right), k^2\log\left(\frac{d}{k^2}+2\right)\right)\right)\right).$$

Proof. For the upper bound, we use the fact that $\|K\|_1 \le \|K\|_2^2$ for all $K \in \mathbb{Z}^d$:

$$\begin{aligned} Q_{k,d} &= \left|\left\{K \in \mathbb{Z}^d : \|K\|_2 \le k\right\}\right| \le \left|\left\{K \in \mathbb{Z}^d : \|K\|_1 \le k^2\right\}\right| \\ &\le \left|\left\{K \in \mathbb{N}^{2d} : \|K\|_1 \le k^2\right\}\right| \qquad (8) \\ &\le \binom{\lceil k^2 \rceil + 2d - 1}{\lceil k^2 \rceil}. \qquad (9) \end{aligned}$$

Inequality (8) holds because we replace each integer in K from the previous line with two natural numbers (there would be equality if we forced one of each pair of natural numbers to equal zero). Line (9) follows from a standard stars-and-bars counting argument. Note that

$$\binom{\lceil k^2 \rceil + 2d - 1}{\lceil k^2 \rceil} = \binom{\lceil k^2 \rceil + 2d - 1}{2d - 1}.$$

We show two separate upper-bounds on that quantity, which together prove the claim:

$$\begin{aligned} Q_{k,d} &\le \binom{\lceil k^2 \rceil + 2d - 1}{2d - 1} \le \left(\frac{e\lceil k^2 \rceil}{2d-1} + e\right)^{2d-1} \le \exp\left(\Theta\left(d \log\left(\frac{k^2}{d} + 2\right)\right)\right); \\ Q_{k,d} &\le \binom{\lceil k^2 \rceil + 2d - 1}{\lceil k^2 \rceil} \le \left(\frac{2ed}{\lceil k^2 \rceil} + e\right)^{\lceil k^2 \rceil} \le \exp\left(\Theta\left(k^2 \log\left(\frac{d}{k^2} + 2\right)\right)\right). \end{aligned}$$

For the lower bound, we observe that

$$\min\left(d \log\left(\frac{k^2}{d} + 2\right), k^2 \log\left(\frac{d}{k^2} + 2\right)\right) = \begin{cases} d \log\left(\frac{k^2}{d} + 2\right) & \text{if } k^2 \ge d, \\ k^2 \log\left(\frac{d}{k^2} + 2\right) & \text{if } k^2 < d. \end{cases}$$

We will lower-bound $Q_{k,d}$ by the appropriate term in each of the two cases, $k^2 \ge d$ and $k^2 < d$.

For the case $k^2 < d$, we lower-bound $Q_{k,d}$ by a sum of binomial coefficients:

$$\begin{aligned} Q_{k,d} &= \left|\left\{K \in \mathbb{Z}^d : \sum_{i=1}^d K_i^2 \le k^2\right\}\right| \\ &\ge \left|\left\{K \in \{0,1\}^d : \sum_{i=1}^d K_i \le k^2\right\}\right| \\ &= \binom{d}{0} + \binom{d}{1} + \cdots + \binom{d}{\lfloor k^2 \rfloor}. \end{aligned}$$

If $\lfloor k^2 \rfloor \le d/2$, then the sum of binomial coefficients is at least the last one, which we bound using

$$\binom{d}{\lfloor k^2 \rfloor} \ge \exp\left(\lfloor k^2 \rfloor \ln \frac{d}{\lfloor k^2 \rfloor}\right) \ge \exp\left(\frac{\lfloor k^2 \rfloor}{2} \ln\left(\frac{d}{\lfloor k^2 \rfloor} + 2\right)\right) = \exp\left(\Theta\left(k^2 \ln\left(\frac{d}{k^2} + 2\right)\right)\right).$$

Otherwise, if $d/2 < \lfloor k^2 \rfloor < d$, the sum of binomial coefficients is at least $2^{\lfloor k^2 \rfloor}$, and

$$2^{\lfloor k^2 \rfloor} = \exp\left((\ln 2)\lfloor k^2 \rfloor\right) \ge \exp\left(\frac{\ln 2}{\ln 4} \lfloor k^2 \rfloor \ln\left(\frac{d}{\lfloor k^2 \rfloor} + 2\right)\right) = \exp\left(\Theta\left(k^2 \ln\left(\frac{d}{k^2} + 2\right)\right)\right).$$

When $k^2 \geq d$, we show that $Q_{k,d}$ grows at a rate similar to that of the volume of a d-dimensional ball of sufficiently large radius $\Theta(k)$. To do so, we regard each $K \in \mathcal{K}_{k,d}$ as an element of $\mathbb{R}^d$, and define

$$A_{k,d} := \left\{ x \in \mathbb{R}^d : \min_{K \in \mathcal{K}_{k,d}} \|x - K\|_\infty \leq \frac{1}{2} \right\}.$$

This is the Minkowski sum of $\mathcal{K}_{k,d}$ and the ℓ_∞ ball of radius $1/2$ in $\mathbb{R}^d$. Note that $A_{k,d}$ has Lebesgue measure $\mathrm{vol}(A_{k,d}) = |\mathcal{K}_{k,d}| = Q_{k,d}$. Let $B_2^d(r) := \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ be the d-dimensional Euclidean ball of radius r. We claim that $B_2^d(k - \sqrt{d}/2) \subset A_{k,d}$, which in turn implies

$$Q_{k,d} \geq \mathrm{vol}\left(B_2^d \left(k - \sqrt{d}/2 \right) \right).$$

To see why this claim holds, consider any $x \in B_2^d(k - \sqrt{d}/2)$. We'll show that $x \in A_{k,d}$. Indeed, there exists some $y \in \mathbb{Z}^d$ such that $\|x - y\|_\infty \leq 1/2$, and hence this y also satisfies $\|x - y\|_2 \leq \sqrt{d}/2$. By the triangle inequality,

$$\begin{aligned} \|y\|_2 &\leq \|x\|_2 + \|x - y\|_2 \\ &\leq \left(k - \frac{\sqrt{d}}{2} \right) + \frac{\sqrt{d}}{2} = k. \end{aligned}$$

Thus, $y \in \mathcal{K}_{k,d}$, which implies $x \in A_{k,d}$.

To complete our lower-bound on $Q_{k,d}$, we observe that

$$\begin{aligned} Q_{k,d} &\geq \mathrm{vol}\left(B_d \left(k - \frac{1}{2}\sqrt{d} \right) \right) \geq \mathrm{vol}\left(B_d \left(\frac{k}{2} \right) \right) \\ &= \frac{\pi^{d/2} (k/2)^d}{\Gamma\left(\frac{d}{2} + 1\right)} \geq \left(\frac{\pi k^2}{2d + 4} \right)^{d/2} \geq \exp\left(\Theta\left(d \log\left(\frac{k^2}{d} + 2 \right) \right) \right), \end{aligned}$$

where Γ is the gamma function and we have used a standard bound on the volume of the d-dimensional Euclidean ball. ■

Appendix B. Supporting lemmas for upper-bounds for Lipschitz functions

This appendix supports Section 3, which presents and proves Theorem 6, the main upper-bound on the minimum width RBL network needed to approximate a Lipschitz function. It contains the proofs of the key Lemmas 7 and 9, which are given in Appendices B.1 and B.2 respectively.

B.1. Trigonometric polynomial approximation for Lipschitz functions

Lemma 22 (Restatement of Lemma 7) *Fix some $L, \epsilon > 0$ with $\frac{L}{\epsilon} \geq 1$ and consider any function $f \in L^2([-1,1]^d)$ with $\|f\|_{\mathrm{Lip}} \leq L$ and $|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \leq L$. Then, taking $k = \frac{L}{\epsilon}$, there exists a bounded-degree trigonometric polynomial*

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K \left(\frac{x}{2} \right)$$

such that $\|f - P\|_{[-1,1]^d} \leq \epsilon$. Moreover, $|\beta_K| \leq L$ for all K.

Proof of Lemma 22. To give a low-degree trigonometric polynomial approximation for f, we transform f into a function $\tilde{f}$ that satisfies periodic boundary conditions, apply Lemma 8 to approximate $\tilde{f}$ with trigonometric polynomial $\tilde{P}$, and obtain P from $\tilde{P}$. Roughly, the argument proceeds as follows:

1. We define $\bar{f} : [0,1]^d \to \mathbb{R}$ to be a rescaling and shift of f so that its domain is the cube $[0,1]^d$. That is, for $x \in [-1,1]^d$ and $y \in [0,1]^d$, $\bar{f}(y) = f(2y - \vec{1})$ and $f(x) = \bar{f}((x+\vec{1})/2)$. Then it holds that $\|\bar{f}\|_{\mathrm{Lip}} \leq 2L$ and $|\,\mathbb{E}_{\mathbf{y}\sim[0,1]^d}[\bar{f}(\mathbf{y})]| = |\,\mathbb{E}_{\mathbf{x}\sim[-1,1]^d}[f(\mathbf{x})]| \leq L$.

2. We define $\tilde{f} : [-1,1]^d \to \mathbb{R}$ by reflecting $\bar{f}$ across orthants as follows: $\tilde{f}(x) = \bar{f}(\mathrm{sign}(x) \odot x)$, where $\mathrm{sign}\,(x) := (\mathrm{sign}(x_1), \dots, \mathrm{sign}(x_d))$ and $\odot$ represents element-wise multiplication. The function $\tilde{f}$ is $2L$-Lipschitz, satisfies the periodic boundary conditions, and has
$$\left| \mathop{\mathbb{E}}_{\mathbf{x}\sim[-1,1]^d} \left[\tilde{f}(\mathbf{x})\right] \right| = \left| \mathop{\mathbb{E}}_{\mathbf{y}\sim[0,1]^d} \left[\bar{f}(\mathbf{y})\right] \right| \leq L.$$

3. We find a low-degree trigonometric polynomial $\tilde{P}$ that ϵ-approximates $\tilde{f}$ over $[-1,1]^d$.

4. Such a $\tilde{P}$ must ϵ-approximate $\tilde{f}$ in at least one of the 2^d unit cubes contained in the orthants of $[-1,1]^d$. Therefore, there exists some sign vector $\nu \in \{-1,1\}^d$ such that $\bar{f}(y)$ is approximated by $\tilde{P}(\nu \odot y)$ on $[0,1]^d$.

5. By shifting and rescaling $\tilde{P}(\nu \odot y)$, we obtain a trigonometric polynomial P that ϵ-approximates f on $[-1,1]^d$ as desired.

Steps (1) and (2) are immediate.

Step (3) follows from Lemma 8. Because $\tilde{f}$ is $2L$-Lipschitz, $\tilde{f}$ satisfies the periodic boundary conditions, $|\,\mathbb{E}_{\mathbf{x}\sim[-1,1]^d}[\tilde{f}(\mathbf{x})]| \leq L$, and $2L/\epsilon \geq 2$, Lemma 8 guarantees the existence of some trigonometric polynomial
$$\tilde{P}(x) = \sum_{K \in \mathcal{K}_{k,d}} \tilde{\beta}_K T_K(x)$$
such that $\|\tilde{f} - \tilde{P}\|_{[-1,1]^d} \leq \epsilon$ and $|\tilde{\beta}_K| \leq L$ for all K.

For step (4), if $\tilde{P}$ is an ϵ-approximator for $\tilde{f}$ over $L_2([-1,1]^d)$, then there must exist a unit cube in some orthant corresponding to some $\nu \in \{-1,1\}^d$ where $\tilde{P}$ also ϵ-approximates $\tilde{f}$. That is,
$$\mathop{\mathbb{E}}_{\mathbf{y}\sim[0,1]^d} \left[\left(\tilde{P}(\nu \odot \mathbf{y}) - \bar{f}(\mathbf{y}) \right)^2 \right] \leq \epsilon^2.$$

For step (5), by translating the distribution from $[-1,1]^d$ to $[0,1]^d$ and taking $P(x) := \tilde{P}(\nu \odot (x+\vec{1})/2)$, we obtain
$$\mathop{\mathbb{E}}_{\mathbf{x}\sim[-1,1]^d} \left[(P(\mathbf{x}) - f(\mathbf{x}))^2 \right] = \mathop{\mathbb{E}}_{\mathbf{y}\sim[0,1]^d} \left[\left(\tilde{P}(\nu \odot \mathbf{y}) - \bar{f}(\mathbf{y}) \right)^2 \right]$$

It remains to show that we can represent P as a proper trigonometric polynomial with halved frequencies and bounded coefficients. We do so by examining each term of the expansion of $\tilde{P}$. Fix any $K \in \mathbb{Z}^d$ with $\|K\|_2 \leq k$ and $K \in \mathcal{K}_{\sin}$. Then, $T_K(y) = \sqrt{2}\sin(\pi\langle K, y\rangle)$. Consider the term corresponding to K of $P(x)$ represented as an expansion of $\tilde{P}$, $\tilde{\beta}_K T_K(\nu \odot (x+\vec{1})/2)$.

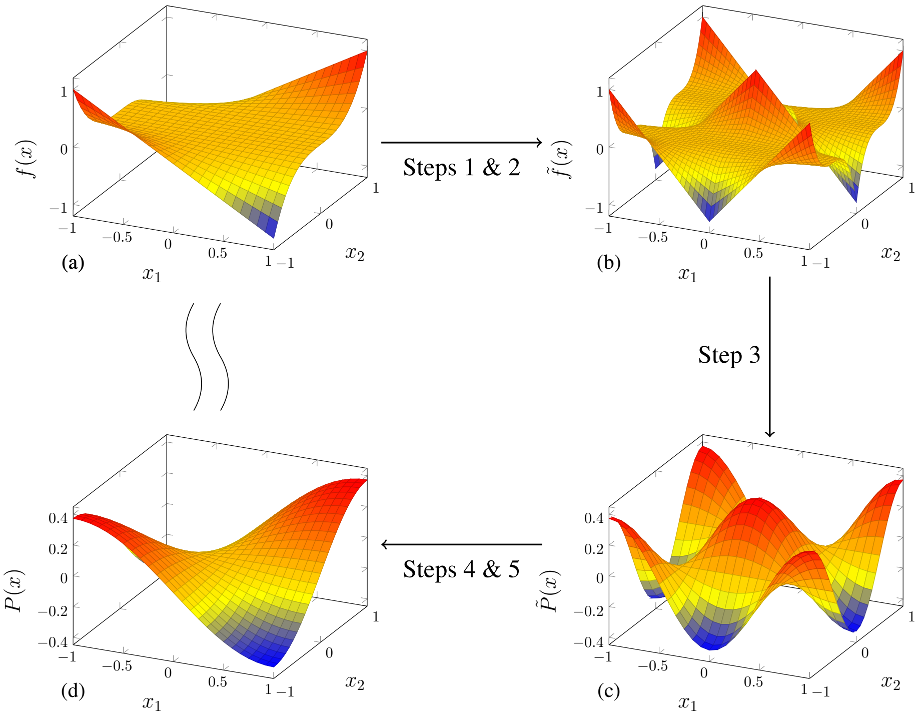

Figure 1: A depiction of the function transformations used to give an approximation of f in Lemma 7. The original function f is in (a), which is scaled and reflected to yield a function $\tilde{f}$ with periodic boundary conditions in (b), which is given a trigonometric polynomial approximation $\tilde{P}$ in (c), which is in turn scaled and shifted to obtain P approximating the original f in (d).

By rearranging its inner product and applying sum-of-angles trigonometric identities, we obtain the following identity:

$$
\begin{aligned}
T_K\left(\frac{1}{2}\nu \odot (x+\vec{1})\right) &= \sqrt{2}\sin\left(\frac{\pi}{2}\langle \nu \odot K, x\rangle + \frac{\pi}{2}\langle \nu \odot K, \vec{1}\rangle\right) \\
&= \begin{cases} \sqrt{2}\sin\left(\frac{\pi}{2}\langle \nu \odot K, x\rangle\right) & \langle \nu \odot K, \vec{1}\rangle \equiv 0 \pmod 4 \\ \sqrt{2}\cos\left(\frac{\pi}{2}\langle \nu \odot K, x\rangle\right) & \langle \nu \odot K, \vec{1}\rangle \equiv 1 \pmod 4 \\ -\sqrt{2}\sin\left(\frac{\pi}{2}\langle \nu \odot K, x\rangle\right) & \langle \nu \odot K, \vec{1}\rangle \equiv 2 \pmod 4 \\ -\sqrt{2}\cos\left(\frac{\pi}{2}\langle \nu \odot K, x\rangle\right) & \langle \nu \odot K, \vec{1}\rangle \equiv 3 \pmod 4. \end{cases}
\end{aligned}
$$

This yields the final representation for T_K functions:

$$
T_K\left(\frac{1}{2}\nu \odot (x+\vec{1})\right) = \begin{cases} T_{\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 0 \pmod 4 \\ T_{-\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 1 \pmod 4 \\ -T_{\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 2 \pmod 4 \\ -T_{-\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 3 \pmod 4. \end{cases}
$$

Similarly,

$$T_{-K}\left(\frac{1}{2}\nu \odot (x+\vec{1})\right) = \begin{cases} T_{-\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 0 \pmod 4 \\ -T_{\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 1 \pmod 4 \\ -T_{-\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 2 \pmod 4 \\ T_{\nu\odot K}(\frac{x}{2}) & \langle \nu \odot K, \vec{1}\rangle \equiv 3 \pmod 4. \end{cases}$$

Using these identities, we can rewrite P as its own trigonometric polynomial with coefficients β_K for all $K \in \mathbb{Z}^d$ such that $\beta_K \in \{\tilde{\beta}_{\nu\odot K}, -\tilde{\beta}_{\nu\odot K}\}$ if $\langle \nu \odot K, \vec{1}\rangle \equiv 0 \pmod 2$, and $\beta_K \in \{\tilde{\beta}_{-\nu\odot K}, -\tilde{\beta}_{-\nu\odot K}\}$ otherwise. Due to the existence of such β_K coefficients, the following trigonometric polynomial approximates f over $[-1,1]^d$:

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \tilde{\beta}_K T_K\left(\frac{1}{2}\nu \odot (x+\vec{1})\right) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K\left(\frac{x}{2}\right). \qquad \blacksquare$$

B.2. RBL ReLU network approximation for trigonometric polynomials

In this section, we give a general purpose lemma that bounds the width needed to approximate trigonometric polynomials of bounded degree.

Lemma 23 (Restatement of Lemma 9) *Fix some $\delta \in (0, 1/2]$, $\epsilon > 0$, $\rho \in (0,1]$, $k \geq 1$, and $d \in \mathbb{Z}^+$. Then, there exists some symmetric ReLU parameter distribution $\mathcal{D}_k$ such that for any trigonometric polynomial*

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K(\rho x)$$

with $|\beta_K| \leq \beta_{\max}$ for all $K \in \mathcal{K}_{k,d}$,

$$\mathrm{MinWidth}_{P,\epsilon,\delta,[-1,1]^d,\mathcal{D}_k} \leq O\left(\frac{\beta_{\max}^2 d^2 k^4}{\epsilon^2} Q_{k,d}^2 \ln\left(\frac{1}{\delta}\right)\right).$$

We first define the specific symmetric ReLU parameter distribution $\mathcal{D}_k$ used in the proof, which can be shown to meet the symmetry criteria spelled out in Definition 4. (As a result, the lower-bounds on the minimum width in Theorems 10 and 13 hold for $\mathcal{D}_k$.)

Definition 24 (Symmetric ReLU parameter distribution $\mathcal{D}_k$ for $[-1,1]^d$ upper-bounds) *Define $\mathcal{D}_k := \mathcal{D}_{\text{bias}} \times \mathcal{D}_{\text{weights},k}$ as a product distribution with the following components:*

- *$\mathcal{D}_{\text{bias}}$ is the uniform distribution over $[-2\sqrt{d}, 2\sqrt{d}]$; and*
- *$\mathcal{D}_{\text{weights},k}$ is a distribution over weights $\mathbf{w}$ taking value in $\mathbb{S}^{d-1}$. To draw $\mathbf{w}$ from $\mathcal{D}_{\text{weights},k}$, draw $\mathbf{K}$ uniformly at random from $\mathcal{K}_{k,d}$ and let $\mathbf{w} := \mathbf{K}/\left\|K\right\|_2$. (If $\mathbf{K} = \vec{0}$, let $\mathbf{w} := \vec{1}/\sqrt{d}$.)*

We also introduce notation to represent the set of vectors contained in $\mathcal{K}_{k,d}$ that generate each $w \in \mathrm{supp}(\mathcal{D}_{\text{weights},k}) \subset \mathbb{S}^{d-1}$:

$$\mathcal{K}_{k,d,w} := \begin{cases} \{K \in \mathcal{K}_{k,d} : K = \eta w, \eta \geq 0\} & w = \frac{1}{\sqrt{d}}\vec{1} \\ \{K \in \mathcal{K}_{k,d} : K = \eta w, \eta > 0\} & \text{otherwise.} \end{cases}$$

Note that every $w \in \operatorname{supp}(\mathcal{D}_{\text{weights},k})$ is drawn with probability $|\mathcal{K}_{k,d,w}|/Q_{k,d}$, which is at least $1/Q_{k,d}$ and at most $(k+1)/Q_{k,d}$.

To prove Lemma 9, we represent P as an expectation over random ReLU features with parameters drawn from $\mathcal{D}_k$. We first express each trigonometric basis element T_K as an expectation over random ReLUs. We leverage the fact that each individual T_K is a ridge function (that is, $T_K(x) = \phi(\langle K, x\rangle)$ for some ϕ). In the following lemma, we show that every ridge function on $[-1,1]^d$ can be represented as a mixture of ReLUs with random bias terms $\mathbf{b}$ drawn from $\mathcal{D}_{\text{bias}}$.

Lemma 25 (Representing ridge functions as a mixture of ReLUs) *Let $\phi : [-\sqrt{d}, \sqrt{d}] \to \mathbb{R}$ be twice differentiable and let $f : [-1,1]^d \to \mathbb{R}$ be $f(x) = \phi(\langle v, x\rangle)$ for some $v \in \mathbb{S}^{d-1}$. Then, for all $x \in [-1,1]^d$,*

$$f(x) = \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_{\text{bias}}}\left[\psi(\mathbf{b})\sigma_{\text{ReLU}}\left(\langle v, x\rangle - \mathbf{b}\right)\right],$$

where

$$\psi(b) := \begin{cases} 4\sqrt{d}a_0 := \frac{16}{\sqrt{d}}\phi(-\sqrt{d}) - 4\phi'(-\sqrt{d}) & b \in [-2\sqrt{d}, -\frac{3}{2}\sqrt{d}) \\ 4\sqrt{d}a_1 := -\frac{16}{\sqrt{d}}\phi(-\sqrt{d}) + 12\phi'(-\sqrt{d}) & b \in [-\frac{3}{2}\sqrt{d}, -\sqrt{d}) \\ 4\sqrt{d}\phi''(b) & b \in [-\sqrt{d}, \sqrt{d}] \\ 0 & b \in (\sqrt{d}, 2\sqrt{d}]. \end{cases}$$

Proof. We expand the expectation over $\mathbf{b}$. For $x \in [-1,1]^d$, let $z := \langle v, x\rangle \in [-\sqrt{d}, \sqrt{d}]$. We have the following:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_{\text{bias}}}\left[\psi(\mathbf{b})\sigma_{\text{ReLU}}\left(\langle v, x\rangle - \mathbf{b}\right)\right] \\
&= a_0 \int_{-2\sqrt{d}}^{-\frac{3}{2}\sqrt{d}} \sigma_{\text{ReLU}}(z-b)\,\mathrm{d}b + a_1 \int_{-\frac{3}{2}\sqrt{d}}^{-\sqrt{d}} \sigma_{\text{ReLU}}(z-b)\,\mathrm{d}b + \int_{-\sqrt{d}}^{\sqrt{d}} \phi''(b)\sigma_{\text{ReLU}}(z-b)\,\mathrm{d}b \\
&= a_0\left(zb - \frac{1}{2}b^2\right)\Big|_{-2\sqrt{d}}^{-\frac{3}{2}\sqrt{d}} + a_1\left(zb - \frac{1}{2}b^2\right)\Big|_{-\frac{3}{2}\sqrt{d}}^{-\sqrt{d}} + \int_{-\sqrt{d}}^{z} \phi''(b)(z-b)\,\mathrm{d}b \\
&= \frac{\sqrt{d}}{2}z\left(a_0 + a_1\right) + \frac{d}{8}\left(7a_0 + 5a_1\right) + \left(\phi'(b)(z-b)\right)\Big|_{-\sqrt{d}}^{z} - \int_{-\sqrt{d}}^{z} \phi'(b)\cdot(-1)\,\mathrm{d}b \\
&= z\phi'(-\sqrt{d}) + \phi(-\sqrt{d}) + \sqrt{d}\phi'(-\sqrt{d}) - \phi'(-\sqrt{d})(z+\sqrt{d}) + \phi(z) - \phi(-\sqrt{d}) \\
&= \phi(z) = f(x).
\end{aligned}$$

■

Once P is represented as an expectation over random ReLUs with parameters drawn from $\mathcal{D}_k$, we conclude the proof by arguing that this expectation can be closely approximated with high probability by a linear combination of sufficiently many randomly sampled ReLUs. We do so by applying a concentration bound due to Yurinskiĭ (1976) for sums of independent random variables taking values in a Hilbert space. We use a convenient version of the bound from Rahimi and Recht (2009, Lemma 4):

Lemma 26 (Concentration inequality for Hilbert spaces) *Let $\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(r)}$ be independent random variables that take values in a Hilbert space with norm $\|\cdot\|$ such that $\|\mathbf{h}^{(i)}\| \le m$ for all i.*

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\| \frac{1}{r}\sum_{i=1}^{r} \mathbf{h}^{(i)} - \mathbb{E}\left[\frac{1}{r}\sum_{i=1}^{r} \mathbf{h}^{(i)}\right] \right\| \leq \frac{m}{\sqrt{r}} \left(1 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right).$$

We are now prepared to formally prove Lemma 9.

Proof of Lemma 9. We first represent any trigonometric monomial T_K as an expected value over weighted ReLUs of the form $\sigma_{\mathrm{ReLU}}(\langle K/\left\|K\right\|_2, x\rangle + \mathbf{b})$ for $\mathbf{b} \sim \mathcal{D}_{\mathrm{bias}}$. For each K, we have $T_K(\rho x) = \phi_K(\langle K/\left\|K\right\|_2, x\rangle)$, where

$$\phi_K(z) = \begin{cases} \sqrt{2}\cos(\pi\rho\left\|K\right\|_2 z) & K \in \mathcal{K}_{\cos} \\ \sqrt{2}\sin(\pi\rho\left\|K\right\|_2 z) & K \in \mathcal{K}_{\sin} \\ 1 & K = \vec{0}. \end{cases}$$

By Lemma 25,

$$T_K(\rho x) = \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_{\mathrm{bias}}}\left[\psi_K(b)\sigma_{\mathrm{ReLU}}\left(\frac{1}{\left\|K\right\|_2}\langle K, x\rangle - \mathbf{b}\right)\right],$$

where ψ_K is the function defined in Lemma 25 for ϕ_K. Because $|\phi_K(z)| \leq \sqrt{2}$, $|\phi_K'(z)| \leq \sqrt{2}\pi\rho\left\|K\right\|_2$, and $|\phi_K''(z)| \leq \sqrt{2}\pi^2\rho^2\left\|K\right\|_2^2$ for all z, we can bound ψ_K:

$$|\psi_K(z)| \leq \max\left\{\frac{16}{\sqrt{d}} \cdot \sqrt{2} + 12 \cdot \sqrt{2}\pi\rho\left\|K\right\|_2, 4\sqrt{d}\sqrt{2}\pi^2\rho^2\left\|K\right\|_2^2\right\} \leq 60\sqrt{d}\left(\left\|K\right\|_2^2 + 1\right).$$

Because any sinusoidal basis element T_K can be expressed as an expectation of random ReLUs and because P is a linear combination of a finite number of those basis elements, we can also represent P as an expectation over ReLUs. We define $h : \mathbb{R} \times \mathbb{S}^{d-1} \to \mathbb{R}$ as

$$h(b,w) = \frac{Q_{k,d}}{|\mathcal{K}_{k,d,w}|}\sum_{K\in\mathcal{K}_{k,d,w}} \beta_K\psi_K(b) = \frac{1}{\Pr_{\mathbf{w}\sim\mathcal{D}_{\mathrm{weights},k}}[\mathbf{w} = w]}\sum_{K\in\mathcal{K}_{k,d,w}} \beta_K\psi_K(b),$$

and represent $P(x)$ as an infinite mixture of ReLU functions weighted by h over all $x \in [-1,1]^d$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{b},\mathbf{w}}\left[h(\mathbf{b},\mathbf{w})\sigma_{\mathrm{ReLU}}\left(\langle \mathbf{w}, x\rangle - \mathbf{b}\right)\right] \\
&= \sum_{w\in\mathrm{supp}(\mathcal{D}_{\mathrm{weights},k})} \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_{\mathrm{bias}}}\left[\sum_{K\in\mathcal{K}_{k,d,w}} \beta_K\psi_K(\mathbf{b})\sigma_{\mathrm{ReLU}}\left(\langle w, x\rangle - \mathbf{b}\right)\right] \\
&= \sum_{w\in\mathrm{supp}(\mathcal{D}_{\mathrm{weights},k})} \sum_{K\in\mathcal{K}_{k,d,w}} \beta_K \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_{\mathrm{bias}}}\left[\psi_K(\mathbf{b})\sigma_{\mathrm{ReLU}}\left(\frac{1}{\left\|K\right\|_2}\langle K, x\rangle - \mathbf{b}\right)\right] \\
&= \sum_{K\in\mathcal{K}_{k,d}} \beta_K \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_{\mathrm{bias}}}\left[\psi_K(\mathbf{b})\sigma_{\mathrm{ReLU}}\left(\frac{1}{\left\|K\right\|_2}\langle K, x\rangle - \mathbf{b}\right)\right] \\
&= \sum_{K\in\mathcal{K}_{k,d}} \beta_K T_K(\rho x) \\
&= P(x).
\end{aligned}$$

To conclude the proof, let $(\mathbf{w}^{(1)}, \mathbf{b}^{(1)}), \dots, (\mathbf{b}^{(r)}, \mathbf{w}^{(r)})$ be independent copies of $(\mathbf{w}, \mathbf{b})$, and define $\mathbf{h}^{(i)} \in L_2([-1,1]^d)$ for $i = 1, \dots, r$ by

$$\mathbf{h}^{(i)}(x) := h(\mathbf{w}^{(i)}, \mathbf{b}^{(i)})\sigma_{\mathrm{ReLU}}(\langle \mathbf{w}^{(i)}, x\rangle - \mathbf{b}^{(i)}).$$

Now we apply Lemma 26 to the random variables $\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(r)}$. Note that $\mathbb{E}_{\mathbf{b}^{(i)}, \mathbf{w}^{(i)}}[\mathbf{h}^{(i)}(x)] = P(x)$. To apply the lemma, we first bound $\|\mathbf{h}^{(i)}\|_{[-1,1]^d}$:

$$\begin{aligned}
\left\|\mathbf{h}^{(i)}\right\|_{[-1,1]^d} &\leq \max_{b \in [-2\sqrt{d}, 2\sqrt{d}], w \in \mathbb{S}^{d-1}, x \in [-1,1]^d} |h(b,w)\sigma_{\mathrm{ReLU}}\left(\langle w, x\rangle - b\right)| \\
&\leq \left(\max_{b,w,x} |\sigma_{\mathrm{ReLU}}\left(\langle w, x\rangle - b\right)|\right)\left(\max_{b,w} |h(b,w)|\right) \\
&= \left(\max_{w,x} \|w\|_2 \|x\|_2 + \max_{b} |b|\right)\left(\max_{b,w} \frac{Q_{k,d}}{|\mathcal{K}_{k,d,w}|}\left|\sum_{K \in \mathcal{K}_{k,d,w}} \beta_K \psi_K(b)\right|\right) \\
&\leq 3\sqrt{d} Q_{k,d} \max_{w} \frac{1}{|\mathcal{K}_{k,d,w}|} \sum_{K \in \mathcal{K}_{k,d,w}} |\beta_K| \cdot 60\sqrt{d}\left(\|K\|_2^2 + 1\right) \\
&\leq 360 d Q_{k,d} \beta_{\max} k^2.
\end{aligned}$$

Therefore, with probability $1 - \delta$,

$$\begin{aligned}
\inf_{g \in \mathrm{Span}\left(\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}\right)} \|P - g\|_{[-1,1]^d} &\leq \left\|\frac{1}{n}\sum_{i=1}^{r} \mathbf{h}^{(i)} - \mathbb{E}\left[\frac{1}{r}\sum_{i=1}^{n} \mathbf{h}^{(i)}\right]\right\|_{[-1,1]^d} \\
&\leq \frac{360 d \beta_{\max} k^2 Q_{k,d}}{\sqrt{r}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right) \leq \epsilon,
\end{aligned}$$

which holds as long as we choose r with

$$r \geq \frac{360^2 d^2 \beta_{\max}^2 k^4 Q_{k,d}^2}{\epsilon^2}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)^2.$$

Based on Definiton 5, this gives the desired upper-bound on MinWidth. ■

Appendix C. Supporting lemmas for lower-bounds for Lipschitz functions

This appendix supports Section 4 by proving Theorems 10 and 13.

C.1. General lower-bounds for random features

In Theorem 29, we give the most general form of our lower-bound. In this setting, we consider linear combinations of features drawn independently from some distribution over functions (which are not required to be ReLUs or even ridge functions). We argue that the span of any r such random functions in $L_2(\mu)$ cannot cover more than r dimensions of that function space and that we therefore cannot approximate most of the members of a family of N orthonormal functions if $N \gg r$.

If the family of N functions satisfies a suitable notion of symmetry with respect to the random features, then we can additionally argue that each function in that family is equally likely to be inapproximable. This makes it possible to construct a single explicit function that cannot be approximated with high probability by linear combinations of random features. We give the relevant notion of symmetry below:

Definition 27 (Symmetry of random functions) *Let* $\mathbf{g}$ *be an* $L_2(\mu)$*-valued random variable for some measure* μ*. We say* $\mathbf{g}$ *is* symmetric *with respect to the set of functions* $\Phi = \{\varphi_1, \dots, \varphi_N\} \subset L_2(\mu)$ *if the distribution of* $\langle \mathbf{g}, \varphi_i \rangle_\mu$ *is the same for all* $i = 1, \dots, N$*.*

In fact, strict orthonormality of the hard functions is not needed for our approach; we introduce a notion of "average coherence,"

which allows us to quantify how far the family is from being orthogonal and prove lower-bounds that depend on this quantity.

Definition 28 (Average coherence) *For any set of functions* $\Phi = \{\varphi_1, \dots, \varphi_N\} \subset L_2(\mu)$ *with* $\|\varphi_i\|_\mu = 1$ *for all* $i = 1, \dots, N$*, its* (average) coherence *is* $\kappa(\Phi) := \sqrt{\sum_{i \neq j} \langle \varphi_i, \varphi_j \rangle_\mu^2}$*.*

We are particularly interested in large collections of functions with low coherence. Note that a collection of orthogonal functions has zero coherence. Our main approximation lower bounds in Theorems 10 and 13 are achieved using an orthogonal collection. However, our general lower bound (Theorem 29) extends to the case where the family of functions has small (but nonzero) coherence, and indeed this version for families with small coherence is useful in extending our general approach to functions over Gaussian space, as we sketch in Appendix E.

The following general lower bound works for any distribution over random features that meets the above symmetry condition and for any set of "nearly-orthonormal" functions that have a bounded average coherence κ. It is akin to Theorem 19 of Kamath et al. (2020) although that result does not involve a symmetry notion (and hence does not yield an explicit hard function).

Theorem 29 (Lower-bound for linear combinations of random features) *Fix a family of functions* $\Phi = \{\varphi_1, \dots, \varphi_N\} \subset L_2(\mu)$ *with* $\|\varphi_i\|_\mu^2 = 1$ *for all* $i = 1, \dots, N$*. Let* $\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}$ *be i.i.d. copies of an* $L_2(\mu)$*-valued random variable. Then, there exists some* $\varphi_i \in \Phi$ *such that*

$$\mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2 \right] \geq 1 - \frac{r\,(1+\kappa(\Phi))}{N}. \tag{10}$$

In particular, for any $\alpha \in [0,1]$*,*

$$\Pr_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2 \geq \alpha \left(1 - \frac{r\,(1+\kappa(\Phi))}{N} \right) \right] \geq (1-\alpha) \left(1 - \frac{r\,(1+\kappa(\Phi))}{N} \right). \tag{11}$$

Moreover, if $\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}$ *are symmetric with respect to* Φ*, then* (10) *and* (11) *hold for* $i = 1$*.*

We recall two tools that will be used in the proof of Theorem 29, namely the Hilbert projection theorem and the Boas-Bellman inequality.

Fact 30 (Hilbert projection theorem (Rudin, 1987)) *For some measure μ and $g^{(1)}, \dots, g^{(r)} \in L_2(\mu)$, consider the subspace $W = \operatorname{Span}(g^{(1)}, \dots, g^{(r)})$ of $L_2(\mu)$. For any $f \in L_2(\mu)$, it holds that*

$$\inf_{g \in W} \|g - f\|_\mu^2 = \|\Pi_W f - f\|_\mu^2 = \|f\|_\mu^2 - \|\Pi_W f\|_\mu^2 \,, \tag{12}$$

where $\Pi_W \colon L_2(\mu) \to W$ is the orthogonal projection operator for W. Moreover, the orthogonal projection $\Pi_W f$ depends on f only through $(\langle g^{(1)}, f\rangle_\mu, \dots, \langle g^{(r)}, f\rangle_\mu)$.

The following is a generalization of Bessel's inequality due to Boas (1941) and Bellman (1944), specialized to our present context.

Fact 31 (Boas-Bellman inequality) *For any $g, \varphi_1, \dots, \varphi_N \in L_2(\mu)$,*

$$\sum_{i=1}^N \langle g, \varphi_i \rangle_\mu^2 \leq \|g\|_\mu^2 \left(\max_{1 \leq i \leq N} \|\varphi_i\|_\mu^2 + \kappa(\{\varphi_1, \dots, \varphi_N\}) \right) . \tag{13}$$

Proof of Theorem 29. By the Hilbert projection theorem (Fact 30), for all $i \in [N]$ we have that

$$\mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2 \right] = 1 - \mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\left\| \Pi_{\operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_i \right\|_\mu^2 \right] .$$

We now upper-bound the sum of the expected norms of the projections of each function in Φ onto $\operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r$. Let $\mathbf{u}_1, \dots, \mathbf{u}_\mathbf{d}$ be an orthonormal basis for $\operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r$, where $\mathbf{d} := \dim \operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r$. Then

$$\begin{aligned}
\sum_{i=1}^N \left\| \Pi_{\operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_i \right\|_\mu^2 &= \sum_{i=1}^N \sum_{k=1}^\mathbf{d} \langle \mathbf{u}_k, \varphi_i \rangle_\mu^2 = \sum_{k=1}^\mathbf{d} \sum_{i=1}^N \langle \mathbf{u}_k, \varphi_i \rangle_\mu^2 && \text{(Plancherel's identity, Fact 15)} \\
&\leq \sum_{k=1}^\mathbf{d} (1 + \kappa(\Phi)) = \mathbf{d} \cdot (1 + \kappa(\Phi)) && \text{(Fact 31)} \\
&\leq r \cdot (1 + \kappa(\Phi)) && (\dim \operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r \leq r).
\end{aligned}$$

Hence, we conclude by linearity of expectation that

$$\frac{1}{N} \sum_{i=1}^N \mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2 \right] \geq 1 - \frac{r \cdot (1 + \kappa(\Phi))}{N}. \tag{14}$$

Therefore, there exists some $i \in [N]$ such that

$$\mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2 \right] \geq 1 - \frac{r \cdot (1 + \kappa(\Phi))}{N},$$

which gives us inequality (10). Inequality (11) follows by an application of Markov's inequality to the random variable $1 - \inf_{g \in \operatorname{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|g - \varphi_i\|_\mu^2$ (which is easily seen to be non-negative), which by the first part of the theorem has expected value at most $r \cdot (1 + \kappa(\Phi))/N$.

We conclude by proving the stronger version of the theorem, where we additionally assume that the random features are symmetric. Suppose $\mathbf{g}^{(1)}, \dots, \mathbf{g}^{(r)}$ are symmetric with respect to Φ. As mentioned in Fact 30, the orthogonal projection $\Pi_{\mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_1$ depends on φ_1 only through the (random) vector $(\langle \mathbf{g}^{(1)}, \varphi_1 \rangle_\mu, \dots, \langle \mathbf{g}^{(r)}, \varphi_1 \rangle_\mu)$. Therefore, by the symmetry assumption on the distribution of each $\mathbf{g}^{(i)}$, the orthogonal projection $\Pi_{\mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_1$ has the same distribution as $\Pi_{\mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_i$ for all $i \in [N]$. Then

$$\mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\left\| \Pi_{\mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_1 \right\|_\mu^2 \right] = \frac{1}{N} \sum_{i=1}^N \mathop{\mathbb{E}}_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\left\| \Pi_{\mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \varphi_i \right\|_\mu^2 \right]. \tag{15}$$

Plugging Equation (15) into Inequality (14) proves that Inequalities (10) and (11) hold for $i = 1$. ■

C.2. MinWidth lower-bounds for RBL ReLU networks

Here, we specialize Theorem 29 to the case of ReLU networks, which prepares us to prove the specific lower-bounds that will be given in the subsequent sections.

Lemma 32 *Let $\mathcal{D}$ be a symmetric ReLU parameter distribution and μ be some measure over $\mathbb{R}^d$. Fix any $\Phi = \{\varphi_1, \dots, \varphi_N\} \subset L_2(\mu)$ such that $\|\varphi_i\|_\mu^2 = 1$ for all $i \in [N]$. Then, for any $\epsilon > 0$, there exists some $\varphi_i \in \Phi$ such that*

$$\mathrm{MinWidth}_{4\epsilon\varphi_i, \epsilon, \frac{1}{2}, \mu, \mathcal{D}} \geq \frac{N}{4 + 4\kappa(\Phi)}. \tag{16}$$

Additionally, suppose that the functions in Φ are symmetric up to some permutation of variables and μ is invariant to permutation of variables. That is, for all $i, i' \in [N]$ there exists a permutation $\pi_{i,i'}$ over $[d]$ such that $\varphi_i \circ \pi_{i,i'} = \varphi_{i'}$. Then, Inequality (16) always holds for $i = 1$.

Proof. By applying Theorem 29 for any $r \leq N/(4 + 4\kappa(\Phi))$ and for $\alpha = 1/3$, there exists some $i \in [N]$ such that

$$\Pr_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|\varphi_i - g\|_\mu < \frac{1}{4} \right] < \frac{1}{2}.$$

Note that for all f, there exists $g \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r$ with $\|f - g\|_\mu < \epsilon$ if and only if there exists $g' \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r$ with $\|f/4\epsilon - g'\|_\mu < 1/4$. Thus, we conclude the following:

$$\Pr_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \|4\epsilon\varphi_i - g\|_\mu < \epsilon \right] = \Pr_{\mathbf{g}^{(1)},\dots,\mathbf{g}^{(r)}} \left[\inf_{g' \in \mathrm{Span}(\mathbf{g}^{(j)})_{j=1}^r} \left\|\varphi_i - g'\right\|_\mu < \frac{1}{4} \right] < \frac{1}{2}.$$

To prove the stronger version of the theorem that assumes permutation symmetry for Φ, we apply the stronger version of Theorem 29. To do so, we must show that each $\mathbf{g}^{(i)}$ is symmetric with respect to Φ.

Because the ReLU feature parameters $\mathbf{b}^{(i)}$ are chosen independently $\mathbf{w}^{(i)}$ and the distribution of $\mathbf{w}^{(i)}$ is invariant to variable permutation, each $\mathbf{g}^{(i)}$ is drawn from a distribution that is also invariant to permutation. We prove the symmetry property by showing that the inner product distributions are identical for $\mathbf{g}^{(1)}$, without loss of generality. Because each function in $\varphi_1, \dots, \varphi_N$ is symmetric to a permutation of variables, there exists some permutation $\pi_{i,i'}$ such that for all $x \in \mu$, $\varphi_i(x) = \varphi_{i'}(\pi_{i,i'}(x))$. To show that the two inner products induce the same distribution, consider any $z \in \mathbb{R}$. Then:

$$\begin{aligned}
&\Pr_{\mathbf{g}^{(1)}}[\langle \mathbf{g}^{(1)}, \varphi_i \rangle_\mu \geq z] \\
&= \Pr_{\mathbf{g}^{(1)}}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\mathbf{g}^{(1)}(\mathbf{x})\varphi_i(\mathbf{x})] \geq z\right] \\
&= \Pr_{\mathbf{g}^{(1)}}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\mathbf{g}^{(1)}(\mathbf{x})\varphi_j(\pi_{i,i'}(\mathbf{x}))] \geq z\right] && \text{(Existence of } \pi_{i,i'}\text{)} \\
&= \Pr_{\mathbf{g}^{(1)}}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\mathbf{g}^{(1)}(\pi_{i,i'}(\mathbf{x}))\varphi_j(\pi_{i,i'}(\mathbf{x}))] \geq z\right] && \text{(Symmetry of } \mathbf{g}^{(1)}\text{'s distribution)} \\
&= \Pr_{\mathbf{g}^{(1)}}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\mathbf{g}^{(1)}(\mathbf{x})\varphi_{i'}(\mathbf{x})] \geq z\right] && \text{(Symmetry of } \mu\text{)} \\
&= \Pr_{\mathbf{g}^{(1)}}[\langle \mathbf{g}^{(1)}, \varphi_{i'} \rangle_\mu \geq z]
\end{aligned}$$

Hence, recalling Definition 27, $\mathbf{g}^{(1)}$ is symmetric with respect to $\varphi_1, \dots, \varphi_N$. By invoking Theorem 29 with the additional symmetry assumption, inequality (16) holds when $i = 1$. ■

C.3. Asymptotically tight lower-bounds for RBL ReLU networks over $[-1, 1]^d$

To finalize the proof of Theorem 10, we first show that some low-degree trigonometric polynomial cannot be approximated by a combination of random ReLU features.[5]

Lemma 33 *For any $k > 0$, any $\epsilon > 0$, and any symmetric ReLU parameter distribution $\mathcal{D}$, there exists some $K \in \mathbb{N}^d$ with $\|K\|_2 \leq k$ such that*

$$\mathrm{MinWidth}_{4\epsilon T_K, \epsilon, \frac{1}{2}, [-1,1]^d, \mathcal{D}} \geq \frac{1}{4} Q_{k,d}.$$

Proof. Let $\mathcal{T}_k := \{T_K \in \mathcal{T} : K \in \mathcal{K}_{k,d}\}$ be a subset of trigonometric basis elements with bounded degree. Because $\mathcal{T}$ is an orthonormal family of functions, $\mathcal{T}_k$ is as well, and $\kappa(\mathcal{T}_k) = 0$. Then, Lemma 32 implies the existence of some $T_K \in \mathcal{T}_k$ such that

$$\mathrm{MinWidth}_{4\epsilon T_K, \epsilon, \frac{1}{2}, [-1,1]^d, \mathcal{D}} \geq \frac{|\mathcal{T}_k|}{4} = \frac{1}{4} Q_{k,d}. \quad ■$$

We prove Theorem 10 by applying Lemma 33 and bounding the Lipschitz constant of the inapproximable function.

[5] We prove Lemma 33 separately from Theorem 10 since we also make use of Lemma 33 in Appendix D.2 when proving lower-bounds based on the Sobolev norm of a function, rather than its Lipschitz constant.

Proof of Theorem 10. Consider any $T_K \in \mathcal{T}$ with $\|K\|_2 \le k$. Then, for all $x, x' \in [-1,1]^d$,

$$\left|T_K(x) - T_K(x')\right| \le \sqrt{2}\pi \langle K, x - x' \rangle \le \sqrt{2}\pi \left\|K\right\|_2 \left\|x - x'\right\|_2 \le \sqrt{2}\pi k \|x - x'\|_2.$$

Thus, $\|T_K\|_{\mathrm{Lip}} \le \sqrt{2}\pi k$ and $\|f\|_{\mathrm{Lip}} \le 4\sqrt{2}\pi k\epsilon \le 18k\epsilon$. By applying Lemma 33 with $k := L/18\epsilon$, there exists a satisfactory f such that $\|f\|_{\mathrm{Lip}} \le L$. ■

C.4. Explicit lower-bounds for RBL ReLU networks over $[-1,1]^d$

As in the previous section, we prove Lemma 34 by applying Lemma 32 to a family of orthonormal functions. In order to obtain an explicit function f that is hard to approximate, we invoke the stronger version of Lemma 32, which requires showing that that the family of functions exhibits symmetry up to a permutation of variables.

Lemma 34 *For any $\ell \in \mathbb{Z}^+$ with $\ell \le d$, any $\epsilon > 0$, and any symmetric ReLU parameter distribution $\mathcal{D}$, define $f : \mathbb{R}^d \to \mathbb{R}$ to be the function $f(x) := 4\sqrt{2}\epsilon \sin(\pi \sum_{i=1}^{\ell} x_i)$. Then,*

$$\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},[-1,1]^d,\mathcal{D}} \ge \frac{1}{4}\binom{d}{\ell}.$$

Proof. We prove the claim by constructing a family of functions Φ_ℓ with $\frac{1}{4\epsilon} f \in \Phi_\ell$ and applying Lemma 32. We define a family of functions

$$\Phi_\ell := \left\{ \varphi_S : x \mapsto \sqrt{2} \sin\left(\pi \sum_{i \in S} x_i\right) \;\middle|\; S \subseteq [d], |S| = \ell \right\}.$$

Note that $|\Phi_\ell| = \binom{d}{\ell}$ and that $\varphi_1 := \frac{1}{4\epsilon} f = \varphi_{[\ell]} \in \Phi_\ell$. Because $\Phi_\ell \subseteq \mathcal{T}$ and $\mathcal{T}$ is an orthonormal basis for $L_2([-1,1]^d)$ (Fact 17), the functions in Φ_ℓ are orthonormal and $\kappa(\Phi_\ell) = 0$. Thus, because the Φ_ℓ satisfies the symmetry conditions for the special case of Lemma 32,

$$\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},[-1,1]^d,\mathcal{D}} \ge \frac{1}{4}\binom{d}{\ell}.$$ ■

Proof of Theorem 13. This is immediate from Lemma 34 and from the fact that $\|f\|_{\mathrm{Lip}} = 4\pi\epsilon\sqrt{2\ell} \le L$. The right-hand side of the bound follows by lower-bounding $\binom{d}{\ell}$ for our choice of ℓ.

If $\ell = \lceil d/2 \rceil$ and $d \ge 2$,[6] then

$$\binom{d}{\ell} \ge \left(\frac{d}{\lceil d/2 \rceil}\right)^{\lceil d/2 \rceil} \ge \left(\frac{3}{2}\right)^{d/2} \ge \exp\left(\Theta(d)\right).$$

Otherwise, $\ell < d/2$ and

$$\binom{d}{\ell} \ge \left(\frac{d}{\ell}\right)^{\ell} \ge \exp\left(\Theta\left(\ell \log\left(\frac{d}{\ell} + 2\right)\right)\right) = \exp\left(\Theta\left(\frac{L^2}{\epsilon^2} \log\left(\frac{d\epsilon^2}{L^2} + 2\right)\right)\right).$$ ■

This matches the exponent asymptotically up to logarithmic factors of the corresponding Lipschitz upper-bound, Theorem 6.

[6] There is no need to consider the $d = 1$ case, because then $\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},[-1,1]^d,\mathcal{D}} \ge \frac{1}{4} = \exp(\Theta(1))$, which satisfies the claim.

Appendix D. Upper- and lower-bounds for Sobolev functions

In this section, we present upper- and lower-bounds on the width required for depth-2 RBL ReLU approximation of functions in a larger family of smooth functions, namely the order-s Sobolev functions. Sobolev spaces are normed function spaces arising in the study of partial differential equations, and their norms quantify the effective "bumpiness" of their constituent functions in terms of their weak derivatives. Let μ denote the uniform probability measure on an open subset of $\mathbb{R}^d$. Following Leoni (2017), we denote the *order-s Sobolev space of functions in $L_2(\mu)$* for $s \in \mathbb{N}$ by[7]

$$H^s(\mu) := \left\{ f : \mathbb{R}^d \to \mathbb{R} : \; D^{(M)} f \in L_2(\mu), \; \forall M \in \mathbb{N}^d \text{ s.t. } |M| \leq s \right\}.$$

The norm on this space is

$$\|f\|_{H^s(\mu)} := \sqrt{\sum_{|M| \leq s} \left\| D^{(M)} f \right\|_\mu^2}.$$

(We do not consider Sobolev spaces in $L_p(\mu)$ for $p \neq 2$ since we rely on Hilbert space structure.)

We focus on the classical spaces $H^s(\mu)$ in $L_2(\mu)$, where μ is the uniform product probability measure on the torus $\mathbb{T}^d$ and $\mathbb{T} = \mathbb{R}/(2\mathbb{Z})$. As a short-hand, we refer to this space as $H^s(\mathbb{T}^d)$ in $L_2(\mathbb{T}^d)$. Recall that $\mathbb{T}$ is obtained by identifying points in $\mathbb{R}$ that differ by $2z$ for some $z \in \mathbb{Z}$. Functions on $\mathbb{T}^d$ can be regarded as functions on $[-1, 1]^d$, which, along with their derivatives, satisfy the periodic boundary conditions. Note that $\mathcal{T}$ is also an orthonormal basis for $\mathbb{T}^d$, because all of the trigonometric polynomials in $\mathcal{T}$ and all their derivatives have periodic boundary conditions and because the probability density of the uniform distribution on $\mathbb{T}^d$ is the same as the density over the uniform distribution on $[-1, 1]^d$.

D.1. Upper-bounds for functions in $H^s(\mathbb{T}^d)$

We prove an analogue to Theorem 6 that places an upper-bound on the minimum width RBL ReLU network that approximates a function with bounded order-s Sobolev norm.

Theorem 35 *Fix some $\delta \in (0, 1/2]$, $\epsilon, \gamma > 0$, and $s \in \mathbb{Z}^+$. Let $k := \sqrt{s}\gamma^{1/s}/\epsilon^{1/s}$. Then, there exists some ReLU parameter distribution $\mathcal{D}$ such that for any fixed $f \in H^s(\mathbb{T}^d)$ that satisfies $\|f\|_{H^s(\mathbb{T}^d)} \leq \gamma$, we have*

$$\mathrm{MinWidth}_{f,\epsilon,\delta,\mathbb{T}^d,\mathcal{D}} \leq O\left(\frac{s^2 \gamma^{2+4/s} d^2}{\epsilon^{2+4/s}} Q_{k,d}^2 \ln\left(\frac{1}{\delta}\right) \right).$$

Remark 36 *When $s = 1$,*

$$\mathrm{MinWidth}_{f,\epsilon,\delta,\mathbb{T}^d,\mathcal{D}} \leq O\left(\frac{\gamma^6 d^2}{\epsilon^6} Q_{\gamma/\epsilon,d}^2 \ln\left(\frac{1}{\delta}\right) \right),$$

which is a near-perfect match to the upper-bound for Lipschitz functions in Theorem 6. This is unsurprising, because all L-Lipschitz functions f with $|\mathbb{E}[f]| \leq L$ have a squared 1-order Sobolev norm with the following bound:

$$\|f\|_{H^s(\mathbb{T}^d)}^2 = \|f\|_{\mathbb{T}}^2 + \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{T}^d}\left[\|\nabla f(\mathbf{x})\|^2 \right] \leq O(L^2).$$

[7] Technically, $D^{(M)} f$ is interpreted as the M-th weak partial derivative of f. However, it satisfies the integration-by-parts formulas that appear in the proof of Lemma 20, which is all we require.

Thus, the two theorems give nearly identical upper-bounds for L-Lipschitz functions that satisfy periodic boundary conditions.

Remark 37 *Applying Fact 3 to Theorem 35 implies that*

$$\mathrm{MinWidth}_{f,\epsilon,\delta,\mathbb{T}^d,\mathcal{D}} \le \ln\left(\frac{1}{\delta}\right)\exp\left(O\left(\min\left(d\log\left(\frac{s\gamma^{2/s}}{d\epsilon^{2/s}}+2\right), \frac{s\gamma^{2/s}}{\epsilon^{2/s}}\log\left(\frac{d\epsilon^{2/s}}{s\gamma^{2/s}}+2\right)\right)\right)\right).$$

Like the proof of Theorem 6, we first show that every function in $H^s(\mathbb{T}^d)$ can be approximated by low-degree trigonometric polynomial in Lemma 38, which is a parallel result to Lemma 7. Unlike Theorem 6, however, we require that f and its first s derivatives satisfy the periodic boundary conditions, which is assured by the fact that $f \in H^s(\mathbb{T}^d)$. Thanks to this assumption, we eliminate the need for the “reflection” trick from Lemma 7, which simplifies the proof.

Lemma 38 (Approximating Sobolev functions with low-degree trigonometric polynomials) *Fix any values $\gamma, \epsilon > 0$ and $s \in \mathbb{Z}^+$. Consider any $f \in H^s(\mathbb{T}^d)$ with $\|f\|_{H^s(\mathbb{T}^d)} \le \gamma$. Let $k := \sqrt{s}\gamma^{1/s}/(2\epsilon)^{1/s}$. Then, there exists a trigonometric polynomial*

$$P(x) = \sum_{K \in \mathcal{K}_{k,d}} \beta_K T_K(x)$$

such that $\|f - P\|_{\mathbb{T}^d} \le \epsilon$. Moreover, $|\beta_K| \le \|f\|_{\mathbb{T}^d} \le \gamma$ for all $K \in \mathcal{K}_{k,d}$.

Proof. Because $\mathcal{T}$ is an orthonormal basis over $\mathbb{T}^d$, we express f as the expansion

$$f = \sum_{K \in \mathbb{Z}^d} \alpha_K T_K.$$

Since f can be regarded as a function on $[-1,1]^d$ whose first s partial derivatives satisfy boundary conditions, Lemma 20 implies that this expansion of f can be differentiated term-by-term. By taking term-by-term partial derivatives of f, applying Parseval’s identity (Fact 15), and using the known norms of partial derivatives of T_K (Fact 18), we obtain the following closed-form $L_2(\mathbb{T}^d)$ norm for $D^{(M)}f$ for all $M \in \mathbb{N}^d$ with $|M| \le s$:

$$\left\|D^{(M)}f\right\|_{\mathbb{T}^d}^2 = \sum_{K \in \mathbb{Z}^d} \alpha_K^2 (\pi K)^{2M}.$$

Therefore, the squared $H^s(\mathbb{T}^d)$-norm of f can be written as

$$\|f\|_{H^s(\mathbb{T}^d)}^2 = \sum_{|M| \le s} \left\|D^{(M)}f\right\|_{\mathbb{T}^d}^2 = \sum_{|M| \le s}\sum_{K \in \mathbb{Z}^d} \alpha_K^2 (\pi K)^{2M} = \sum_{K \in \mathbb{Z}^d} \alpha_K^2 c_{K,s}, \tag{17}$$

where

$$c_{K,s} := \sum_{|M| \le s} (\pi K)^{2M}.$$

We lower-bound $c_{K,s}$ in terms of s and $\|K\|_2$ with the multinomial theorem:

$$c_{K,s} \geq \sum_{|M|=s} (\pi K)^{2M} \geq \frac{\pi^{2s}}{s!} \sum_{|M|=s} \frac{s!}{M!} K^{2M} = \frac{\pi^{2s}}{s!} \|K\|_2^{2s} \geq \left(\frac{\pi^2 \|K\|_2^2}{s} \right)^s .$$

We define $\beta_K := \alpha_K$ for all $K \in \mathcal{K}_{k,d}$ and $\beta_K := 0$ for all other $K \in \mathbb{Z}^d$. Note that if $K \in \mathbb{Z}^d$ has $\|K\|_2 > k \geq \sqrt{s}\gamma^{1/s}/\pi\epsilon^{1/s}$, then $c_{K,s} > \gamma^2/\epsilon^2$. By Parseval's identity, we have $\beta_K^2 \leq \|f\|_{\mathbb{T}^d}^2$. Moreover,

$$\|f - P\|_{\mathbb{T}^d}^2 = \sum_{K \in \mathbb{Z}^d \setminus \mathcal{K}_{k,d}} \alpha_K^2 \leq \sum_{\substack{K \in \mathbb{Z}^d: \\ c_{K,s} > \gamma^2/\epsilon^2}} \alpha_K^2 \leq \sum_{\substack{K \in \mathbb{Z}^d: \\ c_{K,s} > \gamma^2/\epsilon^2}} \alpha_K^2 \cdot \frac{c_{K,S}}{\gamma^2/\epsilon^2} \leq \frac{\epsilon^2}{\gamma^2} \sum_{K \in \mathbb{Z}^d} \alpha_K^2 c_{K,s} \leq \epsilon^2.$$

Above, the first equality uses Parseval's identity, and the final equality uses Equation (17). ■

Proof of Theorem 35. This proof is identical to the proof of Theorem 6 in Section 3.3, except that we make use of Lemma 38 instead of Lemma 7, and instead set $k := \sqrt{s}\gamma^{1/s}/\epsilon^{1/s}$ and $\rho := 1$. ■

D.2. Lower-bounds for functions in $H^s([-1,1]^d)$

Similar to Section 4, we give lower-bounds on the width of RBL ReLU neural networks required to approximate certain functions (now ones with bounded s-order Sobolev norm). As before, we present two variants of the lower-bound, one non-explicit tight bound and one looser explicit bound.

- Theorem 39 is analogous to Theorem 10. It shows the existence of some sinusoidal function with bounded Sobolev norm which matches the upper-bound Theorem 35 by depending on the same combinatorial term.
- Theorem 41, like Theorem 13, offers an explicit sinusoidal function with bounded Sobolev norm whose minimum width can be bounded by a term that differs from the asymptotics of the exponent of the upper-bound by a logarithmic factor.

These results follow from proofs that directly apply Lemmas 33 and 34 respectively and bound the s-order Sobolev norms of the resulting functions.

D.2.1. A TIGHT LOWER-BOUND

We give a bound on the minimum width depth-2 RBL ReLU network needed to approximate some function with bounded Sobolev norm, which is a scaled version of some function in $\mathcal{T}$. The family of functions is identical to that of Theorem 39; the only difference is that we parameterize the bounds by the s-order Sobolev norm of the function, rather than its Lipschitz constant.

Theorem 39 *Fix some $\epsilon, \gamma > 0$ and $s \in \mathbb{Z}_+$ with $\gamma^2/\epsilon^2 \geq 16(s+1)$. Let*

$$k := \frac{\gamma^{1/s}}{\pi 4^{1/s} \epsilon^{1/s} (s+1)^{1/2s}}.$$

Then, there exists some $K \in \mathcal{K}_{k,d}$ such that for $f := 4\epsilon T_K$ and for any symmetric ReLU parameter distribution $\mathcal{D}$,

$$\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},\mathbb{T}^d,\mathcal{D}} \geq \frac{1}{4} Q_{k,d},$$

and $\|f\|_{H^s(\mathbb{T}^d)} \leq \gamma$.

Remark 40 *By invoking Fact 3, we have*

$$\mathrm{MinWidth}_{f,\epsilon,1/2,\mathbb{T}^d,\mathcal{D}} \geq \exp\left(\Omega\left(\min\left(d\log\left(\frac{\gamma^{2/s}}{d\epsilon^{2/s}}+2\right), \frac{\gamma^{2/s}}{\epsilon^{2/s}}\log\left(\frac{d\epsilon^{2/s}}{\gamma^{2/s}}+2\right)\right)\right)\right).$$

Note that we can drop $(s+1)^{1/s}$ terms from the asymptotics of the exponent, because $(s+1)^{1/s} \in (1,2]$ for all $s \in \mathbb{Z}^+$. The asymptotics of the exponents match the upper-bound on the minimum width presented in Remark 37, when $\delta = 1/2$ and s is regarded as a small constant.

Proof. To prove the existence of f, we need only invoke Lemma 33 for our choice of k. It remains to bound the s-order Sobolev norm of f. We do so by expanding the squared Sobolev norm of f and applying Fact 18 to obtain an exact representation of the norms of derivatives of the basis elements $T_K \in \mathcal{T}$.

$$\begin{aligned}\|f\|_{H^s(\mathbb{T}^d)}^2 &= \sum_{M:|M|\leq s} \left\|D^{(M)}f\right\|_{\mathbb{T}^d}^2 = 16\epsilon^2 \sum_{M:|M|\leq s} \left\|D^{(M)}T_K\right\|_{\mathbb{T}^d}^2 = 16\epsilon^2 \sum_{M:|M|\leq s} \pi^{2|M|}K^{2M} \\ &= 16\epsilon^2 \sum_{m=0}^{s} \pi^{2m} \sum_{|M|=m} K^{2M} \leq 16\epsilon^2 \sum_{m=0}^{s} \pi^{2m} \sum_{|M|=m} \frac{m!}{K!}K^{2M} = 16\epsilon^2 \sum_{m=0}^{s} \pi^{2m} \|K\|_2^{2m} \\ &\leq 16\epsilon^2 \sum_{m=0}^{s} \left(\pi^2 k^2\right)^m = 16\epsilon^2 \sum_{m=0}^{s} \left(\frac{\gamma^{2/s}}{16^{1/s}\epsilon^{2/s}(s+1)^{1/s}}\right)^m\end{aligned}$$

Because of our assumed lower-bound on γ^2/ϵ^2, the final term of the sum cannot be smaller than any preceding terms. Therefore, we conclude with the following trivial bound on the sum.

$$\|f\|_{H^s(\mathbb{T}^d)}^2 \leq 16\epsilon^2 \sum_{m=0}^{s} \left(\frac{\gamma^{2/s}}{16^{1/s}\epsilon^{2/s}(s+1)^{1/s}}\right)^m \leq 16\epsilon^2(s+1)\left(\frac{\gamma^{2/s}}{16^{1/s}\epsilon^{2/s}(s+1)^{1/s}}\right)^s = \gamma^2.$$

■

D.2.2. A LOWER-BOUND FOR AN EXPLICIT SINUSOIDAL FUNCTION

We give an explicit lower-bound that bounds the Sobolev norm of the function f used in Lemma 34. In that way, it is nearly identical to Theorem 13.

Theorem 41 *Fix some $\epsilon, \gamma > 0$ and $s \in \mathbb{Z}_+$ with $\gamma^2/\epsilon^2 \geq 16(s+1)$. Let*

$$\ell := \min\left(\left\lceil\frac{d}{2}\right\rceil, \left\lfloor\frac{\gamma^{2/s}}{\pi^2 16^{1/s}\epsilon^{2/s}(s+1)^{1/s}}\right\rfloor\right).$$

Fix any symmetric ReLU parameter distribution $\mathcal{D}$. Then, the function $f(x) := 4\sqrt{2}\epsilon \sin(\pi \sum_{i=1}^{\ell} x_i)$ satisfies $\|f\|_{H^s(\mathbb{T}^d)} \leq \gamma$ and

$$\mathrm{MinWidth}_{f,\epsilon,\frac{1}{2},\mathbb{T}^d,\mathcal{D}} \geq \frac{1}{4}\binom{d}{\ell} \geq \exp\left(\Omega\left(\min\left(\frac{\gamma^{2/s}}{\epsilon^{2/s}} \log\left(\frac{d\epsilon^{2/s}}{\gamma^{2/s}} + 2\right), d\right)\right)\right).$$

Proof. The width bound is immediate from Lemma 34 and from the lower-bounds on $\binom{d}{\ell}$ shown in the proof of Theorem 13. Note that f can be written as $f = 4\epsilon T_K$ for some K with

$$\|K\|_2 = \sqrt{\ell} \leq \frac{\gamma^{1/s}}{\pi 4^{1/s} \epsilon^{1/s} (s+1)^{1/2s}}.$$

Thus, we conclude that $\|f\|_{H^s(\mathbb{T}^d)} \leq \gamma$ by applying the same chain of inequalities from Theorem 39, making use of our lower-bound on γ^2/ϵ^2. ■

Appendix E. A similar approach for the Gaussian measure

The techniques underlying our upper- and lower-bounds on approximation by depth-2 RBL networks are rather general, and can be applied in a broader range of settings than are captured by Theorems 1 and 2. These settings include other activation functions beyond ReLU gates and other functions spaces beyond $L_2([-1,1]^d)$. In this Appendix, we briefly sketch how several of the key ingredients for Theorems 1 and 2 have analogues over *Gaussian space*, and how results similar to Theorems 1 and 2 can be proved over Gaussian space.[8]

E.1. The setting and key background results

We consider the domain $\mathbb{R}^d$ endowed with the standard d-dimensional Gaussian measure $\mathcal{N}(0, I_d)$ with mean zero and identity covariance matrix. It is well known (see e.g. Section 11.2 of O'Donnell (2014)) that the set $\{H_K\}_{K \in \mathbb{N}^d}$ of all *multivariate normalized Hermite polynomials* is an orthonormal basis for $L_2(\mathcal{N}(0, I_d))$, where for $K = (K_1, \ldots, K_d)$ the function H_K is

$$H_K = \prod_{j=1}^{d} h_{K_j}(x_j)$$

where h_i is the degree-i normalized univariate Hermite polynomial. These multivariate Hermite polynomials are analogous to the trigonometric basis polynomials T_K that are introduced in Appendix A for the function space $L_2([-1,1]^d)$.

Well known results (see, e.g., Section 5.5 of Szegö (1989)) show that partial derivatives of multivariate normalized Hermite polynomials can be conveniently expressed in terms of other multivariate normalized Hermite polynomials, very analogous to Equation 4. By combining this with a well-known recurrence relation for Hermite polynomials (again, see Szegö (1989)), it is possible to prove the following result, which is closely analogous to Lemma 20 but now for $L_2(\mathcal{N}(0, I_d))$ rather than $L_2([-1,1]^d)$:

[8] Coarse analogues of the results from Appendix D for Sobolev spaces may also be obtained with these techniques.

Lemma 42 (Term-by-term differentiation for Hermite representation) *Consider some $f \in L_2(\mathcal{N}(0, I_d))$ and $i \in [d]$ such that f is differentiable with respect to x_i and $\frac{\partial f}{\partial x_i} \in L_2(\mathcal{N}(0, I_d))$. Then, f and its partial derivative $\partial f / \partial x_i$ have Hermite expansions of the form*

$$f = \sum_{K \in \mathbb{N}^d} \alpha_K H_K \quad \& \quad \frac{\partial f}{\partial x_i} = \sum_{K \in \mathbb{N}^d} \alpha_K \frac{\partial H_K}{\partial x_i}.$$

E.2. The upper-bound approach

Recall that our positive results for depth-2 RBL ReLU approximation are proved in two stages. In the first stage (Lemma 7, restated as Lemma 22 in Appendix B.1), we argued that any Lipschitz function over $[-1, 1]^d$ can be approximated as a low-degree trigonometric polynomial with bounded coefficients. In the second stage (Lemma 9), we argued that low-degree trigonometric polynomials can be approximated with depth-2 RBL ReLU networks.

For the first stage, with Lemma 42 in hand as an analogue of Lemma 20, it is possible to obtain an analogue of Lemma 22; in the current Gaussian setting, this result shows that functions in $L_2(\mathcal{N}(0, I_d))$ with bounded Lipschitz constant can be approximated with low-degree Hermite polynomials whose coefficients (in terms of the orthonormal basis of normalized multivariate Hermite polynomials) are not too large. (The argument is in fact simpler than for Lemma 22 because there are no issues with periodic boundary conditions, which were responsible for steps 1, 2, 4 and 5 of the outline provided at the beginning of the proof of Lemma 22.)

For the second stage, some technical challenges arise because the Hermite basis functions (unlike the trigonometric polynomials defined in Appendix A) are not ridge functions. These challenges can be overcome: using techniques from Andoni et al. (2014), it is possible to show that the small-coefficient, low-degree Hermite polynomials we are dealing with can indeed be approximated by depth-2 RBL ReLU networks. It turns out that the resulting width of the RBL ReLU networks obtained using this approach is roughly $(dL/\epsilon)^{O(L^2/\epsilon^2)}$, i.e., polynomial in the dimension d but exponential in L/ϵ; this corresponds to a somewhat weaker analogue of Theorem 6 in which the "$\min(L^2/\epsilon^2, d)$" is replaced with just L^2/ϵ^2, and gives a good upper bound when d is large compared to L^2/ϵ^2. For the complementary regime where L^2/ϵ^2 is large compared to d, using different techniques[9] it can be shown that in fact depth-2 RBL ReLU networks of width roughly $(dL/\epsilon)^{O(d)}$ also suffice; combining these two regimes, this gives an overall approximation result for Lipschitz functions over Gaussian space that is quite closely analogous to Theorem 6. The arguments to establish these results are somewhat lengthy for each of the two regimes, though, so we omit both the arguments and detailed claims of the results in this paper.

[9] Roughly speaking, the approach (inspired by Ji et al. (2019)) is to (i) truncate the function by setting it to a constant outside of a ball of carefully chosen radius; (ii) approximate the truncated function with a superposition of "Gaussian bumps;" (iii) approximate this superposition of Gaussian bumps by a weighted average of random ReLU gates.

E.3. The lower-bound approach

Recall that our main lower bound tool, Theorem 29, only requires small average coherence (rather than strict orthogonality) for the set of "hard" functions . Exploiting this flexibility, it is not difficult to adapt Theorem 29 to the setting of Gaussian space.

In a bit more detail, it turns out that taking a family $\Phi = \{\varphi_1, \ldots, \varphi_N\}$ of "hard" functions that corresponds to points $v^{(1)}, \ldots, v^{(N)}$ in a suitable packing of the unit sphere, where the function $\varphi_i(x)$ is defined to be (a suitably normalized version of) $\sin(L\langle v^{(i)}, x\rangle)$, results in Φ having small average coherence, and from this it is not difficult to obtain lower-bounds on depth-2 RBL ReLU network width, following the approach of Section 4. The resulting lower bounds can be shown to be quite close to matching the upper-bounds for Gaussian space sketched in the previous subsection.