JMLR: Workshop and Conference Proceedings vol 49:1–34, 2016

The Power of Depth for Feedforward Neural Networks

Ronen Eldan RONEN.ELDAN@WEIZMANN.AC.IL
Weizmann Institute of Science, Rehovot, Israel

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science, Rehovot, Israel

Abstract

We show that there is a simple (approximately radial) function on $\mathbb{R}^d$, expressible by a small 3-layer feedforward neural networks, which cannot be approximated by any 2-layer network, to more than a certain constant accuracy, unless its width is exponential in the dimension. The result holds for virtually all known activation functions, including rectified linear units, sigmoids and thresholds, and formally demonstrates that depth – even if increased by 1 – can be exponentially more valuable than width for standard feedforward neural networks. Moreover, compared to related results in the context of Boolean functions, our result requires fewer assumptions, and the proof techniques and construction are very different.

Keywords: Neural networks, Depth vs. Width, Function approximation, Fourier transform

1. Introduction and Main Result

Learning via multi-layered artificial neural networks, a.k.a. deep learning, has seen a dramatic resurgence of popularity over the past few years, leading to impressive performance gains on difficult learning problems, in fields such as computer vision and speech recognition. Despite their practical success, our theoretical understanding of their properties is still partial at best.

In this paper, we consider the question of the *expressive power* of neural networks of *bounded* size. The boundedness assumption is important here: It is well-known that sufficiently large depth-2 neural networks, using reasonable activation functions, can approximate any continuous function on a bounded domain (Cybenko (1989); Hornik et al. (1989); Funahashi (1989); Barron (1994)). However, the required size of such networks can be exponential in the dimension, which renders them impractical as well as highly prone to overfitting. From a learning perspective, both theoretically and in practice, our main interest is in neural networks whose size is bounded.

For a network of bounded size, a basic architectural question is how to trade off between its width and depth: Should we use networks that are narrow and deep (many layers, with a small number of neurons per layer), or shallow and wide? Is the "deep" in "deep learning" really important? Or perhaps we can always content ourselves with shallow (e.g. depth-2) neural networks?

Overwhelming empirical evidence as well as intuition indicates that having depth in the neural network is indeed important: Such networks tend to result in complex predictors which seem hard to capture using shallow architectures, and often lead to better practical performance. However, for the types of networks used in practice, there are surprisingly few formal results (see related work below for more details).

In this work, we consider fully connected feedforward neural networks, using a linear output neuron and some non-linear activation function on the other neurons, such as the commonly-used

rectified linear unit (ReLU, $\sigma(z) = \max\{z, 0\}$), as well as the sigmoid ($\sigma(z) = (1 + \exp(-z))^{-1}$) and the threshold ($\sigma(z) = \mathbf{1}\{z \geq 0\}$). Informally speaking, we consider the following question: *What functions on $\mathbb{R}^d$ expressible by a network with ℓ-layers and w neurons per layer, that cannot be well-approximated by any network with $< \ell$ layers, even if the number of neurons is allowed to be much larger than w?*

More specifically, we consider the simplest possible case, namely the difficulty of approximating functions computable by 3-layer networks using 2-layer networks, when the networks are feedforward and fully connected. Following a standard convention, we define a 2-layer network of width w on inputs in $\mathbb{R}^d$ as

$$\mathbf{x} \mapsto \sum_{i=1}^{w} v_i \sigma\left(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i\right) \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, and $v_i, b_i \in \mathbb{R}$, $\mathbf{w}_i \in \mathbb{R}^d$, $i = 1, \ldots, w$ are parameters of the network. This corresponds to a set of w neurons computing $\mathbf{x} \mapsto \sigma(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i)$ in the first layer, whose output is fed to a linear output neuron $\mathbf{x} \mapsto \sum_{i=1}^{w} v_i x_i$ in the second layer[1]. Similarly, a 3-layer network of width w is defined as

$$\sum_{i=1}^{w} u_i \sigma\left(\sum_{j=1}^{w} v_{i,j} \sigma\left(\langle \mathbf{w}_{i,j}, \mathbf{x} \rangle + b_{i,j}\right) + c_i\right), \tag{2}$$

where $u_i, c_i, v_{i,j}, b_{i,j} \in \mathbb{R}, \mathbf{w}_{i,j} \in \mathbb{R}^d$, $i, j = 1, \ldots, w$ are parameters of the network. Namely, the outputs of the neurons in the first layer are fed to neurons in the second layer, and their outputs in turn are fed to a linear output neuron in the third layer.

Clearly, to prove something on the separation between 2-layer and 3-layer networks, we need to make some assumption on the activation function $\sigma(\cdot)$ (for example, if $\sigma(\cdot)$ is the identity, then both 2-layer and 3-layer networks compute linear functions, hence there is no difference in their expressive power). All we will essentially require is that $\sigma(\cdot)$ is *universal*, in the sense that a sufficiently large 2-layer network can approximate any univariate Lipschitz function which is non-constant on a bounded domain. More formally, we use the following assumption:

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

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

satisfies

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

1. Note that sometimes one also adds a constant bias parameter b to the output neuron, but this can be easily simulated by a "constant" neuron i in the first layer where $w_i = \mathbf{0}$ and v_i, b_i are chosen appropriately. Also, sometimes the output neuron is defined to have a non-linearity as well, but we stick to linear output neurons, which is a very common and reasonable assumption for networks computing real-valued predictions.

This assumption is satisfied by the standard activation functions we are familiar with. First of all, we provide in Appendix A a constructive proof for the ReLU function. For the threshold, sigmoid, and more general sigmoidal functions (e.g. monotonic functions which satisfy $\lim_{z\to\infty}\sigma(z) = a, \lim_{z\to-\infty}\sigma(z) = b$ for some $a \neq b$ in $\mathbb{R}$), the proof idea is similar, and implied by the proof of Theorem 1 of Debao (1993)[2]. Finally, one can weaken the assumed bound on w to any $\text{poly}(R, L, 1/\delta)$, at the cost of a worse polynomial dependence on the dimension d in Thm. 1 part 1 below (see Subsection 4.4 for details).

In addition, for technical reasons, we will require the following mild growth and measurability conditions, which are satisfied by virtually all activation functions in the literature, including the examples discussed earlier:

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

$$|\sigma(x)| \leq C(1 + |x|^{\alpha})$$

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

Our main result is the following theorem, which implies that there are 3-layer networks of width polynomial in the dimension d, which cannot be arbitrarily well approximated by 2-layer networks, unless their width is exponential in d:

Theorem 1 *Suppose the activation function $\sigma(\cdot)$ satisfies assumption 1 with constant c_σ, as well as assumption 2. Then there exist universal constants $c, C > 0$ such that the following holds: For every dimension $d > C$, there is a probability measure μ on $\mathbb{R}^d$ and a function $g : \mathbb{R}^d \to \mathbb{R}$ with the following properties:*

1. *g is bounded in $[-2, +2]$, supported on $\{\mathbf{x} : \|\mathbf{x}\| \leq C\sqrt{d}\}$, and expressible by a 3-layer network of width $Cc_\sigma d^{19/4}$.*
2. *Every function f, expressed by a 2-layer network of width at most ce^{cd}, satisfies*

$$\mathbb{E}_{\mathbf{x}\sim\mu}\left(f(\mathbf{x}) - g(\mathbf{x})\right)^2 \geq c.$$

The proof is sketched in Sec. 2, and is formally presented in Sec. 4. Roughly speaking, g approximates a certain radial function $\tilde{g}$, depending only on the norm of the input. With 3 layers, approximating radial functions (including $\tilde{g}$) to arbitrary accuracy is straightforward, by first approximating the squared norm function, and then approximating the univariate function acting on the norm. However, performing this approximation with only 2 layers is much more difficult, and the proof shows that exponentially many neurons are required to approximate $\tilde{g}$ to more than constant accuracy. We conjecture (but do not prove) that a much wider family of radial functions also satisfy this property.

We make the following additional remarks about the theorem:

2. Essentially, a single neuron with such a sigmoidal activation can express a (possibly approximate) single-step function, a combination of w such neurons can express a function with w such steps, and any L-Lipschitz function which is constant outside $[-R, R]$ can be approximated to accuracy δ with a function involving $\mathcal{O}(RL/\delta)$ steps.

Remark 2 (Activation function) *The theorem places no constraints on the activation function $\sigma(\cdot)$ beyond assumptions 1 and 2. In fact, the inapproximability result for the function $\tilde{g}$ holds even if the activation functions are different across the first layer neurons, and even if they are chosen adaptively (possibly depending on $\tilde{g}$), as long as they satisfy assumption 2.*

Remark 3 (Constraints on the parameters) *The theorem places no constraints whatsoever on the parameters of the 2-layer networks, and they can take any values in $\mathbb{R}$. This is in contrast to related depth separation results in the context of threshold circuits, which do require the size of the parameters to be constrained (see discussion of related work below).*

Remark 4 (Properties of g) *At least for specific activation functions such as the ReLU, sigmoid, and threshold, the proof construction implies that g is poly(d)-Lipschitz, and the 3-layer network expressing it has parameters bounded by poly(d).*

Related Work

On a qualitative level, the question we are considering is similar to the question of Boolean circuit lower bounds in computational complexity: In both cases, we consider functions which can be represented as a combination of simple computational units (Boolean gates in computational complexity; neurons in neural networks), and ask how large or how deep this representation needs to be, in order to compute or approximate some given function. For Boolean circuits, there is a relatively rich literature and some strong lower bounds. A recent example is the paper Rossman et al. (2015), which shows for any $d \geq 2$ an explicit depth d, linear-sized circuit on $\{0,1\}^n$, which cannot be non-trivially approximated by depth $d-1$ circuits of size polynomial in n. That being said, it is well-known that the type of computation performed by each unit in the circuit can crucially affect the hardness results, and lower bounds for Boolean circuits do *not* readily translate to neural networks of the type used in practice, which are real-valued and express continuous functions. For example, a classical result on Boolean circuits states that the parity function over $\{0,1\}^d$ cannot be computed by constant-depth Boolean circuits whose size is polynomial in d (see for instance Håstad (1986)). Nevertheless, the parity function can in fact be easily computed by a simple 2-layer, $\mathcal{O}(d)$-width *real-valued* neural network with most reasonable activation functions[3].

A model closer to ours is a *threshold circuit*, which is a neural network where all neurons (including the output neuron) has a threshold activation function, and the input is from the Boolean cube (see Parberry (1994) for a survey). For threshold circuits, the main known result in our context is that computing inner products mod 2 over d-dimensional Boolean vectors cannot be done with a 2-layer network with poly(d)-sized parameters and poly(d) width, but can be done with a small 3-layer network (Hajnal et al. (1993)). Note that unlike neural networks in practice, the result in Hajnal et al. (1993) is specific to the non-continuous threshold activation function, and considers hardness of exact representation of a function by 2-layer circuits, rather than merely approximating it. Following the initial publication of our paper, we were informed (Martens (2015)) that the proof technique, together with techniques in the papers (Maass et al. (1994); Martens et al. (2013))), can

3. See Rumelhart et al. (1986), Figure 6, where reportedly the structure was even found automatically by back-propagation. For a threshold activation function $\sigma(z) = \mathbf{1}\{z \geq 0\}$ and input $\mathbf{x} = (x_1, \ldots, x_d) \in \{0,1\}^d$, the network is given by $\mathbf{x} \mapsto \sum_{i=1}^{d+1} (-1)^{i+1} \sigma\left(\sum_{j=1}^{d} x_j - i + \frac{1}{2}\right)$. In fact, we only need σ to satisfy $\sigma(z) = 1$ for $z \geq \frac{1}{2}$ and $\sigma(z) = 0$ for $z \leq -\frac{1}{2}$, so the construction easily generalizes to other activation functions (such as a ReLU or a sigmoid), possibly by using a small linear combination of them to represent such a σ.

possibly be used to show that inner product mod 2 is also hard to approximate, using 2-layer neural networks with continuous activation functions, as long as the network parameters are constrained to be polynomial in d, and that the activation function satisfies certain regularity conditions[4]. Even so, our result does not pose any constraints on the parameters, nor regularity conditions beyond assumptions 1,2. Moreover, we introduce a new proof technique which is very different, and demonstrate hardness of approximating not the Boolean inner-product-mod-2 function, but rather functions in $\mathbb{R}^d$ with a simple geometric structure (namely, radial functions).

Moving to networks with real-valued outputs, one related field is arithmetic circuit complexity (see Shpilka and Yehudayoff (2010) for a survey), but the focus there is on computing polynomials, which can be thought of as neural networks where each neuron computes a linear combination or a product of its inputs. Again, this is different than most standard neural networks used in practice, and the results and techniques do not readily translate.

Recently, several works in the machine learning community attempted to address questions similar to the one we consider here. Pascanu et al. (2013); Montufar et al. (2014) consider the number of linear regions which can be expressed by ReLU networks of a given width and size, and Bianchini and Scarselli (2014) consider the topological complexity (via Betti numbers) of networks with certain activation functions, as a function of the depth. Although these can be seen as measures of the function's complexity, such results do not translate directly to a lower bound on the approximation error, as in Thm. 1. Delalleau and Bengio (2011); Martens and Medabalimi (2014) and Cohen et al. (2015) show strong approximation hardness results for certain neural network architectures (such as polynomials or representing a certain tensor structure), which are however fundamentally different than the standard neural networks considered here.

Quite recently, Telgarsky (2015) gave a simple and elegant construction showing that for any k, there are k-layer, $\mathcal{O}(1)$ wide ReLU networks on one-dimensional data, which can express a sawtooth function on $[0,1]$ which oscillates $\mathcal{O}(2^k)$ times, and moreover, such a rapidly oscillating function cannot be approximated by poly(k)-wide ReLU networks with $o(k/\log(k))$ depth. This also implies regimes with exponential separation, e.g. that there are k^2-depth networks, which any approximating k-depth network requires $\Omega(\exp(k))$ width. These results demonstrate the value of depth for arbitrarily deep, standard ReLU networks, for a single dimension and using functions which have an exponentially large Lipschitz parameter. In this work, we use different techniques, to show exponential separation results for general activation functions, even if the number of layers changes by just 1 (from two to three layers), and using functions in $\mathbb{R}^d$ whose Lipschitz parameter is polynomial in d.

2. Proof Sketch

In a nutshell, the 3-layer network we construct approximates a radial function with bounded support (i.e. one which depends on the input $\mathbf{x}$ only via its Euclidean norm $\|\mathbf{x}\|$, and is 0 for any $\mathbf{x}$ whose norm is larger than some threshold). With 3 layers, approximating radial functions is rather straightforward: First, using assumption 1, we can construct a linear combination of neurons expressing the univariate mapping $z \mapsto z^2$ arbitrarily well in any bounded domain. Therefore, by adding these combinations together, one for each coordinate, we can have our network first compute (approxi-

4. See remark 20 in Martens et al. (2013). These conditions are needed for constructions relying on distributions over a finite set (such as the Boolean hypercube). However, since we consider continuous distributions on $\mathbb{R}^d$, we do not require such conditions.

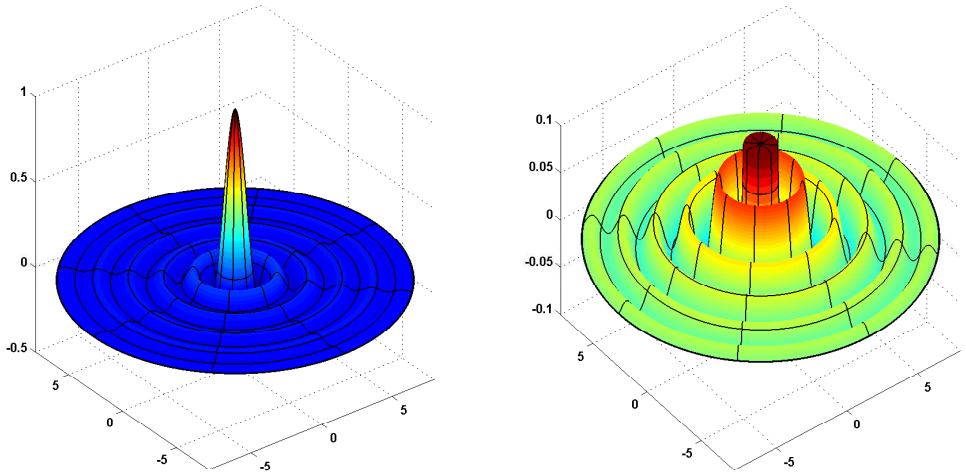

Figure 1: The left figure represents $\varphi(\mathbf{x})$ in $d = 2$ dimensions. The right figure represents a cropped and re-scaled version, to better show the oscillations of φ beyond the big origin-centered bump. The density of the probability measure μ is defined as $\varphi^2(\cdot)$

mately) the mapping $\mathbf{x} \mapsto \|\mathbf{x}\|^2 = \sum_i x_i^2$ inside any bounded domain, and then use the next layer to compute some univariate function of $\|\mathbf{x}\|^2$, resulting in an approximately radial function. With only 2 layers, it is less clear how to approximate such radial functions. Indeed, our proof essentially indicates that approximating radial functions with 2 layers can require exponentially large width.

To formalize this, note that if our probability measure μ has a well-behaved density function which can be written as $\varphi^2(\mathbf{x})$ for some function φ, then the approximation guarantee in the theorem, $\mathbb{E}_\mu(f(\mathbf{x}) - g(\mathbf{x}))^2$, can be equivalently written as

$$\int (f(\mathbf{x}) - g(\mathbf{x}))^2 \varphi^2(\mathbf{x}) d\mathbf{x} \;=\; \int (f(\mathbf{x})\varphi(\mathbf{x}) - g(\mathbf{x})\varphi(\mathbf{x}))^2 d\mathbf{x} \;=\; \|f\varphi - g\varphi\|_{L_2}^2. \tag{3}$$

In particular, we will consider a density function which equals $\varphi^2(\mathbf{x})$, where φ is the inverse Fourier transform of the indicator $\mathbf{1}\{\mathbf{x} \in B\}$, B being the origin-centered unit-volume Euclidean ball (the reason for this choice will become evident later). Before continuing, we note that a formula for φ can be given explicitly (see Lemma 6), and an illustration of it in $d = 2$ dimensions is provided in Figure 2. Also, it is easily verified that $\varphi^2(\mathbf{x})$ is indeed a density function: It is clearly non-negative, and by isometry of the Fourier transform, $\int \varphi^2(\mathbf{x}) d\mathbf{x} = \int \hat{\varphi}^2(\mathbf{x}) d\mathbf{x} = \int \mathbf{1}\{\mathbf{x} \in B\}^2 \, d\mathbf{x}$, which equals 1 since B is a unit-volume ball.

Our goal now is to lower bound the right hand side of Eq. (3). To continue, we find it convenient to consider the Fourier transforms $\widehat{f\varphi}, \widehat{g\varphi}$ of the functions $f\varphi, g\varphi$, rather than the functions themselves. Since the Fourier transform is isometric, the above equals

$$\|\widehat{f\varphi} - \widehat{g\varphi}\|_{L_2}^2.$$

Luckily, the Fourier transform of functions expressible by a 2-layer network has a very particular form. Specifically, consider any function of the form

$$f(\mathbf{x}) = \sum_{i=1}^{k} f_i(\langle \mathbf{v}_i, \mathbf{x} \rangle),$$

where $f_i : \mathbb{R} \to \mathbb{R}$ (such as 2-layer networks as defined earlier). Note that f may not be square-integrable, so formally speaking it does not have a Fourier transform in the standard sense of a function on $\mathbb{R}^d$. However, assuming $|f_i(x)|$ grows at most polynomially as $x \to \infty$ or $x \to -\infty$, it does have a Fourier transform in the more general sense of a tempered distribution (we refer the reader to the proof for a more formal discussion). This distribution can be shown to be supported on $\bigcup_i \mathrm{span}\{\mathbf{v}_i\}$: namely, a finite collection of lines[5]. The convolution-multiplication principle implies that $\widehat{f\varphi}$ equals $\hat{f} \star \hat{\varphi}$, or the convolution of $\hat{f}$ with the indicator of a unit-volume ball B. Since $\hat{f}$ is supported on $\bigcup_i \mathrm{span}\{\mathbf{v}_i\}$, it follows that

$$\mathrm{Supp}(\widehat{f\varphi}) \quad \subseteq \quad T \quad := \bigcup_{i=1}^{k} \left(\mathrm{span}\{\mathbf{v}_i\} + B\right).$$

In words, the support of $\widehat{f\varphi}$ is contained in a union of tubes of bounded radius passing through the origin. This is the key property of 2-layer networks we will use to derive our main theorem. Note that it holds regardless of the exact shape of the f_i functions, and hence our proof will also hold if the activations in the network are different across the first layer neurons, or even if they are chosen in some adaptive manner.

To establish our theorem, we will find a function g expressible by a 3-layer network, such that $\widehat{g\varphi}$ has a constant distance (in L_2 space) from any function supported on T (a union of k tubes as above). Here is where high dimensionality plays a crucial role: Unless k is exponentially large in the dimension, the domain T is very sparse when one considers large distances from the origin, in the sense that

$$\frac{Vol_{d-1}(T \cap r\mathbb{S}^{d-1})}{Vol_{d-1}(r\mathbb{S}^{d-1})} \lesssim ke^{-d}$$

(where $\mathbb{S}^{d-1}$ is the d-dimensional unit Euclidean sphere, and Vol_{d-1} is the $d-1$-dimensional Hausdorff measure) whenever r is large enough with respect to the radius of B. Therefore, we need to find a function g so that $\widehat{g\varphi}$ has a lot of mass far away from the origin, which will ensure that $\|\widehat{f\varphi} - \widehat{g\varphi}\|_{L_2}^2$ will be large. Specifically, we wish to find a function g so that $g\varphi$ is radial (hence $\widehat{g\varphi}$ is also radial, so having large mass in any direction implies large mass in all directions), and has a significant *high-frequency* component, which implies that its Fourier transform has a significant portion of its mass outside of the ball rB.

The construction and analysis of this function constitutes the technical bulk of the proof. The main difficulty in this step is that even if the Fourier transform $\hat{g}$ of g has some of its L_2 mass on high frequencies, it is not clear that this will also be true for $\widehat{g\varphi} = g \star \mathbf{1}\{B\}$ (note that while

5. Roughly speaking, this is because each function $\mathbf{x} \mapsto f_i(\langle \mathbf{v}_i, \mathbf{x} \rangle)$ is constant in any direction perpendicular to $\mathbf{v}_i$, hence do not have non-zero Fourier components in those directions. In one dimension, this can be seen by the fact that the Fourier transform of the constant 0 function is the Dirac delta function, which equals 0 everywhere except at the origin.

convolving with a Euclidean ball increases the average distance from the origin in the L_1 sense, it doesn't necessarily do the same in the L_2 sense).

We overcome this difficulty by considering a random superposition of indicators of thin shells: Specifically, we consider the function

$$\tilde{g}(\mathbf{x}) = \sum_{i=1}^{N} \epsilon_i g_i(\mathbf{x}), \tag{4}$$

where $\epsilon_i \in \{-1, +1\}$, $N = \text{poly}(d)$, and $g_i(\mathbf{x}) = \mathbf{1}\{\|\mathbf{x}\| \in \Delta_i\}$, where Δ_i are disjoint intervals of width $\mathcal{O}(1/N)$ on values in the range $\Theta(\sqrt{d})$. Note that strictly speaking, we cannot take our hard-to-approximate function g to equal $\tilde{g}$, since $\tilde{g}$ is discontinuous and therefore cannot be expressed by a 3-layer neural network with continuous activations functions. However, since our probability distribution φ^2 can be shown to have bounded density in the support of Eq. (4), we can use a 3-layer network to approximate such a function arbitrarily well with respect to the distribution φ^2 (for example, by computing $\sum_i \epsilon_i g_i(\mathbf{x})$ as above, with each hard indicator function g_i replaced by a Lipschitz function, which differs from g_i on a set with arbitrarily small probability mass). Letting the function g be such a good approximation, we get that if no 2-layer network can approximate the function $\tilde{g}$ in Eq. (4), then it also cannot approximate its 3-layer approximation g.

Let us now explain why the function defined in Eq. (4) gives us what we need. For large N, each g_i is supported on a thin Euclidean shell, hence $g_i\varphi$ is approximately the same as $c_i g_i$ for some constant c_i. As a result, $\tilde{g}(\mathbf{x})\varphi(\mathbf{x}) \approx \sum_{i=1}^{N} \epsilon_i c_i g_i(\mathbf{x})$, so its Fourier transform (by linearity) is $\widehat{\tilde{g}\varphi}(\mathbf{w}) \approx \sum_{i=1}^{N} \epsilon_i c_i \hat{g}_i(\mathbf{w})$. Since g_i is a simple indicator function, its Fourier transform $\hat{g}_i(\mathbf{w})$ is not too difficult to compute explicitly, and involves an appropriate Bessel function which turns out to have a sufficiently large mass sufficiently far away from the origin.

Knowing that each summand g_i has a relatively large mass on high frequencies, our only remaining objective is to find a choice for the signs ϵ_i so that the entire sum will have the same property. This is attained by a random choice of signs: it is an easy observation that given an orthogonal projection P in a Hilbert space H, and any sequence of vectors $v_1, ..., v_N \in H$ such that $|Pv_i| \geq \delta |v_i|$, one has that $\mathbb{E}\left[|P\sum_i \epsilon_i v_i|^2\right] \geq \delta^2 \sum_i |v_i|^2$ when the signs ϵ_i are independent Bernoulli ± 1 variables. Using this observation with P being the projection onto the subspace spanned by functions supported on high frequencies and with the functions $\hat{g}_i$, it follows that there is at least one choice of the ϵ_i's so that a sufficiently large portion of $\tilde{g}$'s mass is on high frequencies.

3. Preliminaries

We begin by defining some of the standard notation we shall use. We let $\mathbb{N}$ and $\mathbb{R}$ denote the natural and real numbers, respectively. Bold-faced letters denote vectors in d-dimensional Euclidean space $\mathbb{R}^d$, and plain-faced letters to denote either scalars or functions (distinguishing between them should be clear from context). L_2 denotes the space of squared integrable functions ($\int_{\mathbf{x}} f^2(\mathbf{x})d\mathbf{x} < \infty$, where the integration is over $\mathbb{R}^d$), and L_1 denotes the space of absolutely integrable functions ($\int_{\mathbf{x}} |f(\mathbf{x})|d\mathbf{x} < \infty$). $\|\cdot\|$ denotes the Euclidean norm, $\langle \cdot, \cdot \rangle_{L_2}$ denotes inner product in L_2 space (for functions f, g, we have $\langle f, g \rangle_{L_2} = \int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}$), $\|\cdot\|_{L_2}$ denotes the standard norm in L_2 space ($\|f\|_{L_2}^2 = \int_{\mathbf{x}} f(\mathbf{x})^2 d\mathbf{x}$), and $\|\cdot\|_{L_2(\mu)}$ denotes the L_2 space norm weighted by a probability measure μ (namely $\|f\|_{L_2(\mu)}^2 = \int f(\mathbf{x})^2 d\mu(\mathbf{x})$). Given two functions f, g, we let fg be shorthand for the

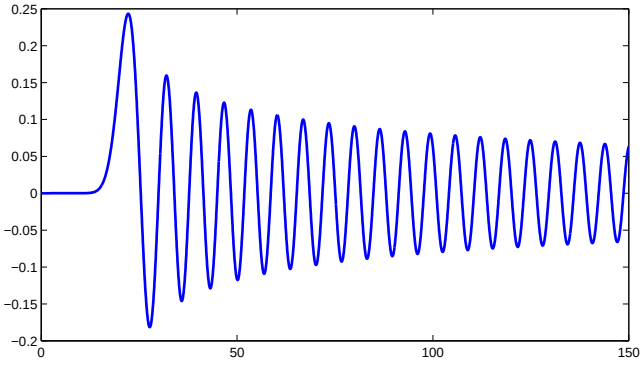

Figure 2: Bessel function of the first kind, $J_{20}(\cdot)$

function $\mathbf{x} \mapsto f(\mathbf{x}) \cdot g(\mathbf{x})$, and $f + g$ be shorthand for $\mathbf{x} \mapsto f(\mathbf{x}) + g(\mathbf{x})$. Given two sets A, B in $\mathbb{R}^d$, we let $A + B = \{a + b : a \in A, b \in B\}$ and $A^C = \{a \in \mathbb{R}^d : a \notin A\}$

Fourier Transform. For a function $f : \mathbb{R} \to \mathbb{R}$, our convention for the Fourier transform is

$$\hat{f}(w) = \int_{\mathbb{R}} \exp(-2\pi i x w) f(x) dx$$

whenever the integral is well defined. This is generalized for $f : \mathbb{R}^d \to \mathbb{R}$ by

$$\hat{f}(\mathbf{w}) = \int_{\mathbb{R}^d} \exp(-2\pi i \langle \mathbf{x}, \mathbf{w} \rangle) f(\mathbf{x}) d\mathbf{x}. \tag{5}$$

Radial Functions. A radial function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is such that $f(\mathbf{x}) = f(\mathbf{x}')$ for any $\mathbf{x}, \mathbf{x}'$ such that $\|\mathbf{x}\| = \|\mathbf{x}'\|$. When dealing with radial functions, which are invariant to rotations, we will somewhat abuse notation and interchangeably use vector arguments $\mathbf{x}$ to denote the value of the function at $\mathbf{x}$, and scalar arguments r to denote the value of the same function for any vector of norm r. Thus, for a radial function $f : \mathbb{R}^d \to \mathbb{R}$, $f(r)$ equals $f(\mathbf{x})$ for any $\mathbf{x}$ such that $\|\mathbf{x}\| = r$.

Euclidean Spheres and Balls. Let $\mathbb{S}^{d-1}$ be the unit Euclidean sphere in $\mathbb{R}^d$, B_d be the d-dimensional unit Euclidean ball, and let R_d be the radius so that $R_d B_d$ has volume one. By standard results, we have the following useful lemma:

Lemma 5 *$R_d = \sqrt{\frac{1}{\pi}} \left(\Gamma\left(\frac{d}{2} + 1\right)\right)^{1/d}$, which is always between $\frac{1}{5}\sqrt{d}$ and $\frac{1}{2}\sqrt{d}$.*

Bessel Functions. Let $J_\nu : \mathbb{R} \mapsto \mathbb{R}$ denote the Bessel function of the first kind, of order ν. The Bessel function has a few equivalent definitions, for example $J_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\nu+1)} \left(\frac{x}{2}\right)^{2m+\nu}$ where $\Gamma(\cdot)$ is the Gamma function. Although it does not have a closed form, $J_\nu(x)$ has an oscillating shape, which for asymptotically large x behaves as $\sqrt{\frac{2}{\pi x}} \cos\left(-\frac{(2\nu+1)\pi}{4} + x\right)$. Figure 3 illustrates the function for $\nu = 20$. In appendix C, we provide additional results and approximations for the Bessel function, which are necessary for our proofs.

4. Proof of Thm. 1

In this section, we provide the proof of Thm. 1. Note that some technical proofs, as well as some important technical lemmas on the structure of Bessel functions, are deferred to the appendix.

4.1. Constructions

As discussed in Sec. 2, our theorem rests on constructing a distribution μ and an appropriate function g, which is easy to approximate (w.r.t. μ) by small 3-layer networks, but difficult to approximate using 2-layer networks. Thus, we begin by formally defining g, μ that we will use.

First, μ will be defined as the measure whose density is $\frac{d\mu}{d\mathbf{x}} = \varphi^2(\mathbf{x})$, where $\varphi(\mathbf{x})$ is the Fourier transform of the indicator of a unit-volume Euclidean ball $\mathbf{1}\{\mathbf{w} \in R_d B_d\}$. Note that since the Fourier transform is an isometry, $\int_{\mathbb{R}^d} \varphi(\mathbf{x})^2 d\mathbf{x} = \int_{\mathbb{R}^d} \mathbf{1}\{\mathbf{w} \in R_d B_d\}^2 \, d\mathbf{w} = 1$, hence μ is indeed a probability measure. The form of φ is expressed by the following lemma:

Lemma 6 *Let $\varphi(\mathbf{x})$ be the Fourier transform of $\mathbf{1}\{\mathbf{w} \in R_d B_d\}$. Then*

$$\varphi(\mathbf{x}) = \left(\frac{R_d}{\|\mathbf{x}\|}\right)^{d/2} J_{d/2}(2\pi R_d \|\mathbf{x}\|).$$

The proof appears in Appendix B.1.

To define our hard-to-approximate function, we introduce some notation. Let $\alpha \geq 1$ and γ be some large numerical constants to be determined later, and set $N = \gamma d^2$, assumed to be an integer (essentially, we need α, γ to be sufficiently large so that all the lemmas we construct below would hold). Consider the intervals

$$\Delta_i = \left[\left(1 + \frac{i-1}{N}\right)\alpha\sqrt{d}\,, \;\; \left(1 + \frac{i}{N}\right)\alpha\sqrt{d}\right] \;\;, \;\; i = 1, 2, \ldots, N.$$

We split the intervals to "good" and "bad" intervals using the following definition:

Definition 7 *Δ_i is a* good interval *(or equivalently, i is good) if for any $x \in \Delta_i$*

$$J_{d/2}^2(2\pi R_d x) \geq \frac{1}{80\pi R_d x}.$$

Otherwise, we say that Δ_i is a bad interval.

For any i, define

$$g_i(x) = \begin{cases} \mathbf{1}\{x \in \Delta_i\} & i \text{ good} \\ 0 & i \text{ bad} \end{cases} \tag{6}$$

By definition of a "good" interval and Lemma 6, we see that g_i is defined to be non-zero, when the value of φ on the corresponding interval Δ_i is sufficiently bounded away from 0, a fact which will be convenient for us later on.

Our proof will revolve around the L_2 function

$$\tilde{g}(\mathbf{x}) = \sum_{i=1}^{N} \epsilon_i g_i(\mathbf{x}),$$

which as explained in Sec. 2, will be shown to be easy to approximate arbitrarily well with a 3-layer network, but hard to approximate with a 2-layer network.

4.2. Key Lemmas

In this subsection, we collect several key technical lemmas on g_i and φ, which are crucial for the main proof. The proofs of all the lemmas can be found in Appendix B.

The following lemma ensures that $\varphi(\mathbf{x})$ is sufficiently close to being a constant on any good interval:

Lemma 8 *If $d \geq 2$, $\alpha \geq c$ and $N \geq c\alpha^{3/2}d^2$ (for some sufficiently large universal constant c), then inside any good interval Δ_i, $\varphi(x)$ has the same sign, and*

$$\frac{\sup_{x\in\Delta_i}|\varphi(x)|}{\inf_{x\in\Delta_i}|\varphi(x)|} \leq 1 + d^{-1/2}.$$

The following lemma ensures that the Fourier transform $\hat{g}_i$ of g_i has a sufficiently large part of its L_2 mass far away from the origin:

Lemma 9 *Suppose $N \geq 100\alpha d^{3/2}$. Then for any i,*

$$\int_{(2R_dB_d)^C} \hat{g_i}^2(\mathbf{w})d\mathbf{w} \;\geq\; \frac{1}{2}\int_{\mathbb{R}^d} \hat{g_i}^2(\mathbf{w})d\mathbf{w},$$

where $\hat{g}_i$ is the Fourier transform of g_i.

The following lemma ensures that $\widehat{g_i\varphi}$ also has sufficiently large L_2 mass far away from the origin:

Lemma 10 *Suppose that $\alpha \geq C$, $N \geq C\alpha^{3/2}d^2$ and $d > C$, where $C > 0$ is a universal constant. Then for any i,*

$$\int_{(2R_dB_d)^C} (\widehat{(g_i\varphi)}(\mathbf{w}))^2 d\mathbf{w} \geq \frac{1}{4}\int_{\mathbb{R}^d} (\varphi(\mathbf{x})g_i(\mathbf{x}))^2 d\mathbf{x}.$$

The following lemma ensures that a linear combination of the g_i's has at least a constant $L_2(\varphi^2)$ probability mass.

Lemma 11 *Suppose that $\alpha \geq c$ and $N \geq c(\alpha d)^{3/2}$ for some sufficiently large universal constant c, then for every choice of $\epsilon_i \in \{-1, +1\}$, $i = 1, \ldots, N$, one has*

$$\int \left(\sum_{i=1}^N \epsilon_i g_i(\mathbf{x})\right)^2 \varphi^2(\mathbf{x})d\mathbf{x} \geq \frac{0.003}{\alpha}.$$

Finally, the following lemma guarantees that the non-Lipschitz function $\sum_{i=1}^N \epsilon_i g_i(\mathbf{x})$ can be approximated by a Lipschitz function (w.r.t. the density φ^2). This will be used to show that $\sum_{i=1}^N \epsilon_i g_i(\mathbf{x})$ can indeed be approximated by a 3-layer network.

Lemma 12 *Suppose that $d \geq 2$. For any choice of $\epsilon_i \in \{-1, +1\}$, $i = 1, \ldots, N$, there exists an N-Lipschitz function f, supported on $[\alpha\sqrt{d}, 2\alpha\sqrt{d}]$ and with range in $[-1, +1]$, which satisfies*

$$\int \left(f(\mathbf{x}) - \sum_{i=1}^N \epsilon_i g_i(\mathbf{x})\right)^2 \varphi^2(\mathbf{x})d\mathbf{x} \;\leq\; \frac{3}{\alpha^2\sqrt{d}}.$$

4.3. Inapproximability of the Function $\tilde{g}$ with 2-Layer Networks

The goal of this section is to prove the following proposition.

Proposition 13 *Fix a dimension d, suppose that $d > C$, $\alpha > C$ and $N \geq C\alpha^{3/2}d^2$ and let k be an integer satisfying*

$$k \leq ce^{cd} \tag{7}$$

with $c, C > 0$ being universal constants. There exists a choice of $\epsilon_i \in \{-1, +1\}$, $i = 1, \ldots, N$, such that the function $\tilde{g}(\mathbf{x}) = \sum_{i=1}^{N} \epsilon_i g_i(\|\mathbf{x}\|)$ has the following property. Let $f : \mathbb{R}^d \to \mathbb{R}$ be of the form

$$f(\mathbf{x}) = \sum_{i=1}^{k} f_i(\langle \mathbf{x}, \mathbf{v}_i \rangle) \tag{8}$$

for $\mathbf{v}_i \in \mathbb{S}^{d-1}$, where $f_i : \mathbb{R} \to \mathbb{R}$ are measurable functions satisfying

$$|f_i(x)| \leq C'(1 + |x|^{\kappa})$$

for constants $C', \kappa > 0$. Then one has

$$\|f - \tilde{g}\|_{L_2(\mu)} \geq \delta/\alpha$$

where $\delta > 0$ is a universal constant.

The proof of this proposition requires a few intermediate steps. In the remainder of the section, we will assume that N, d, α are chosen to be large enough to satisfy the assumptions of Lemma 11 and Lemma 10. In other words we assume that $d > C$ and $N \geq C\alpha^{3/2}d^2$ for a suitable universal constant $C > 0$. We begin with the following:

Lemma 14 *Suppose that d, N are as above. There exists a choice of $\epsilon_i \in \{-1, 1\}$, $i = 1, \ldots, N$ such that*

$$\int_{(2R_dB_d)^C} \left(\left(\widehat{\sum_i \epsilon_i g_i \varphi} \right)(\mathbf{w}) \right)^2 d\mathbf{w} \geq c$$

for a universal constant $c > 0$.

Proof Suppose that each ϵ_i is chosen independently and uniformly at random from $\{-1, +1\}$. It suffices to show that

$$\mathbb{E}\left[\int_{(2R_dB_d)^C} \left(\left(\widehat{\sum_i \epsilon_i g_i \varphi} \right)(\mathbf{w}) \right)^2 d\mathbf{w} \right] \geq c$$

for some universal constant $c > 0$, since that would ensure there exist some choice of $\epsilon_1, \ldots, \epsilon_N$ satisfying the lemma statement. Define $h(\mathbf{w}) = \mathbf{1}\left\{\mathbf{w} \in (2R_dB_d)^C\right\}$ and consider the operator

$$P(g) = \widehat{gh}.$$

This is equivalent to removing low-frequency components from g (in the Fourier domain), and therefore is an orthogonal projection. According to Lemma 10 and isometry of the Fourier transform, we have

$$\|P(g_i\varphi)\|_{L_2}^2 \geq \frac{1}{4}\|g_i\|_{L_2(\mu)}^2 \tag{9}$$

for every good i. Moreover, an application of Lemma 11, and the fact that $\langle g_i, g_j\rangle_{L_2} = 0$ for any $i \neq j$ (as g_i, g_j have disjoint supports) tells us that

$$\sum_{i=1}^{N}\|g_i\|_{L_2(\mu)}^2 = \left\|\sum_{i=1}^{N} g_i\right\|_{L_2(\mu)}^2 \geq c \tag{10}$$

for a universal constant $c > 0$. We finally get,

$$\begin{aligned}
\mathbb{E}\left[\int_{(2R_dB_d)^C}\left(\widehat{\left(\sum_{i=1}^{N}\epsilon_i g_i\varphi\right)}(\mathbf{w})\right)^2 d\mathbf{w}\right] &= \mathbb{E}\left[\int_{\mathbb{R}^d}\left(\left(\sum_{i=1}^{N}\epsilon_i P(g_i\varphi)\right)(\mathbf{x})\right)^2 d\mathbf{x}\right] \\
&= \mathbb{E}\left\|\sum_{i=1}^{N}\epsilon_i P(g_i\varphi)\right\|_{L_2}^2 \\
&= \sum_{i,j=1}^{N}\mathbb{E}[\epsilon_i\epsilon_j]\langle P(g_i\varphi), P(g_j\varphi)\rangle_{L_2} \\
&= \sum_{i=1}^{N}\|P(g_i\varphi)\|_{L_2}^2 \\
&\overset{Eq.\ (9)}{\geq} \frac{1}{4}\sum_{i,j=1}^{N}\|g_i\|_{L_2(\mu)}^2 \\
&\overset{Eq.\ (10)}{\geq} c/4.
\end{aligned}$$

for a universal constant $c > 0$. ■

Claim 15 *Let f be a function such that $f\varphi \in L_2$, and is of the form in Eq. (8). Suppose that the functions f_i are measurable functions satisfying*

$$|f_i(x)| \leq C(1 + |x|^\alpha) \tag{11}$$

for constants $C, \alpha > 0$. Then,

$$\mathrm{Supp}(\widehat{f\varphi}) \subset \bigcup_{i=1}^{k}(\mathrm{Span}\{\mathbf{v}_i\} + R_dB_d) \tag{12}$$

Proof Informally, the proof is based on the convolution-multiplication and linearity principles of the Fourier transform, which imply that if $f = \sum_i f_i$, where each f_i as well as φ have a Fourier transform, then $\widehat{f\varphi} = \sum_i \widehat{f_i\varphi} = \sum_i \hat{f}_i \star \hat{\varphi}$. Roughly speaking, in our case each $\hat{f}_i(\mathbf{x}) = \hat{f}_i(\langle \mathbf{x}, \mathbf{v}_i \rangle)$ (as a function in $\mathbb{R}^d$) is shown to be supported on $\mathrm{Span}\{\mathbf{v}_j\}$, so its convolution with $\hat{\varphi}$ (which is an indicator for the ball $R_d B_d$) must be supported on $\mathrm{Span}\{\mathbf{v}_i\} + R_d B_d$. Summing over i gives the stated result.

Unfortunately, this simple analysis is not formally true, since we are not guaranteed that f_i has a Fourier transform as a function in L_2 (this corresponds to situations where the integral in the definition of the Fourier transform in Eq. (5) does not converge). However, at least for functions satisfying the claim's conditions, the Fourier transform still exists as a generalized function, or *tempered distribution*, over $\mathbb{R}^d$, and using this object we can attain the desired result.

We now turn to provide the formal proof and constructions, starting with a description of tempered distributions and their relevant properties (see (Hunter and Nachtergaele, 2001, Chapter 11) for a more complete survey). To start, let $\mathcal{S}$ denote the space of Schwartz functions [6] on $\mathbb{R}^d$. A tempered distribution μ in our context is a continuous linear operator from $\mathcal{S}$ to $\mathbb{R}$ (this can also be viewed as an element in the dual space $\mathcal{S}^*$). In particular, any measurable function $h : \mathbb{R}^d \mapsto \mathbb{R}$, which satisfies a polynomial growth condition similar to Eq. (11), can be viewed as a tempered distribution defined as

$$\psi \mapsto \langle h, \psi \rangle \;:=\; \int_{\mathbb{R}^d} h(\mathbf{x})\psi(\mathbf{x})d\mathbf{x},$$

where $\psi \in \mathcal{S}$. Note that the growth condition ensures that the integral above is well-defined. The Fourier transform $\hat{h}$ of a tempered distribution h is also a tempered distribution, and defined as

$$\langle \hat{h}, \psi \rangle \;:=\; \langle h, \hat{\psi} \rangle,$$

where $\hat{\psi}$ is the Fourier transform of ψ. It can be shown that this directly generalizes the standard notion of Fourier transforms of functions. Finally, we say that a tempered distribution h is supported on some subset of $\mathbb{R}^d$, if $\langle h, \psi \rangle = 0$ for any function $\psi \in \mathcal{S}$ which vanishes on that subset.

With these preliminaries out of the way, we turn to the setting considered in the claim. Let $\hat{f}_i$ denote the Fourier transform of f_i (possibly as a tempered distribution, as described above), the existence of which is guaranteed by the fact that f_i is measurable and by Eq. (11). We also define, for $\psi : \mathbb{R}^d \to \mathbb{R}$ and $1 \leq i \leq N$, a corresponding function $\psi_i : \mathbb{R} \to \mathbb{R}$ by

$$\psi_i(x) = \psi(x\mathbf{v}_i),$$

and define the tempered distribution μ_i (over Schwartz functions in $\mathbb{R}^d$) as

$$\langle \mu_i, \psi \rangle \;:=\; \langle f_i, \hat{\psi}_i \rangle,$$

which is indeed an element of $\mathcal{S}^*$ by the linearity of the Fourier transform, by the continuity of $\psi \to \psi_i(x)$ with respect to the topology of $\mathcal{S}$ and by the dominated convergence theorem. Finally, define

$$\tilde{f}_i(\mathbf{x}) = f_i(\langle \mathbf{x}, \mathbf{v}_i \rangle).$$

6. This corresponds to infinitely-differentiable functions $\psi : \mathbb{R}^d \mapsto \mathbb{R}$ with super-polynomially decaying values and derivatives, in the sense that $\sup_{\mathbf{x}} \left(x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_d^{\alpha_d} \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdot \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} f(\mathbf{x}) \right) < \infty$ for all indices $\alpha_1, \ldots, \alpha_d$.

Using the fact that[7]

$$\int_{\mathbb{R}^d} \hat{g}(\mathbf{x})d\mathbf{x} = g(0) \tag{13}$$

for any $g \in \mathcal{S}$, recalling that $\mathbf{v}_i$ has unit norm, and letting $\mathbf{v}_i^\perp$ denote the subspace of vectors orthogonal to $\mathbf{v}_i$, we have the following for any $\psi \in \mathcal{S}$:

$$\begin{aligned}
\langle \tilde{f}_i, \hat{\psi} \rangle &= \int_{\mathbb{R}_d} \tilde{f}_i(\mathbf{x})\hat{\psi}(\mathbf{x})d\mathbf{x} \;=\; \int_{\mathbb{R}} \int_{\mathbf{v}_i^\perp} \tilde{f}_i(\mathbf{x}+\mathbf{v}_i y)\hat{\psi}(\mathbf{x}+\mathbf{v}_i y)d\mathbf{x}dy \\
&= \int_{\mathbb{R}} f_i(y) \int_{\mathbf{v}_i^\perp} \hat{\psi}(\mathbf{x}+\mathbf{v}_i y)d\mathbf{x}dy \\
&= \int_{\mathbb{R}} f_i(y) \int_{\mathbf{v}_i^\perp} \int_{\mathbb{R}^d} \psi(\mathbf{w}) \exp(-2\pi i \langle \mathbf{w}, \mathbf{x}+\mathbf{v}_i y\rangle) d\mathbf{w}d\mathbf{x}dy \\
&= \int_{\mathbb{R}} f_i(y) \int_{\mathbf{v}_i^\perp} \int_{\mathbb{R}} \int_{\mathbf{v}_i^\perp} \psi(\mathbf{w}_1+\mathbf{v}_i w_2) \exp(-2\pi i \langle \mathbf{w}_1+\mathbf{v}_i w_2, \mathbf{x}+\mathbf{v}_i y\rangle) d\mathbf{w}_1 dw_2 d\mathbf{x}dy \\
&= \int_{\mathbb{R}} f_i(y) \int_{\mathbb{R}} \exp(-2\pi i w_2 y) \int_{\mathbf{v}_i^\perp} \int_{\mathbf{v}_i^\perp} \psi(\mathbf{w}_1+\mathbf{v}_i w_2) \exp(-2\pi i \langle \mathbf{w}_1, \mathbf{x}\rangle) d\mathbf{w}_1 d\mathbf{x} dw_2 dy \\
&\overset{(13)}{=} \int_{\mathbb{R}} f_i(y) \int_{\mathbb{R}} \exp(-2\pi i w_2 y)\psi(\mathbf{v}_i w_2) dw_2 dy \\
&= \int_{\mathbb{R}} f_i(y)\hat{\psi}(\mathbf{v}_i y)dy \;=\; \int_{\mathbb{R}} f_i(y)\hat{\psi}_i(y)dy \;=\; \langle f_i, \hat{\psi}_i\rangle \;=\; \langle \mu_i, \psi\rangle,
\end{aligned} \tag{14}$$

where the use of Fubini's theorem is justified by the fact that $\psi \in \mathcal{S}$.

We now use the convolution-multiplication theorem (see e.g., (Hunter and Nachtergaele, 2001, Theorem 11.35)) according to which if $f, g \in L_1$ then

$$\widehat{f \star g} = \hat{f}\hat{g}. \tag{15}$$

Using this, we have the following for every $\psi \in \mathcal{S}$:

$$\begin{aligned}
\langle \widehat{\tilde{f}_i \varphi}, \psi \rangle \;&= \langle \tilde{f}_i \varphi, \hat{\psi} \rangle \;=\; \langle \tilde{f}_i, \varphi\hat{\psi} \rangle \\
&\overset{(15)}{=} \langle \tilde{f}_i, \widehat{\hat{\varphi} \star \psi} \rangle \\
&\overset{(14)}{=} \langle \mu_i, \hat{\varphi} \star \psi \rangle \;=\; \langle \mu_i, \mathbf{1}\{R_d B_d\} \star \psi \rangle.
\end{aligned}$$

Based on this equation, we now claim that $\langle \widehat{\tilde{f}_i \varphi}, \psi \rangle = 0$ for any $\psi \in \mathcal{S}$ supported on the complement of $\mathrm{Span}\{\mathbf{v}_i\} + R_d B_d$. This would imply that the tempered distribution $\widehat{\tilde{f}_i \varphi}$ is supported in $\mathrm{Span}\{\mathbf{v}_i\} + R_d B_d$, and therefore $\widehat{f\varphi}$ is supported in $\bigcup_{i=1}^k (\mathrm{Span}\{\mathbf{v}_i\} + R_d B_d)$ (by linearity of the Fourier transform and the fact that $f = \sum_{i=1}^k \tilde{f}_i$). Since the Fourier transform of $f\varphi$ as a tempered distribution coincides with the standard one (as we assume $f\varphi \in L_2$), the result follows.

7. This is because $\int \hat{g}(\mathbf{x})d\mathbf{x} = \int\int g(\mathbf{x}) \exp(-2\pi i \langle \mathbf{x}, \mathbf{w}\rangle) d\mathbf{w}d\mathbf{x} = \int g(\mathbf{x}) \left(\int \exp(-2\pi i \langle \mathbf{x}, \mathbf{w}\rangle) \cdot 1 d\mathbf{w}\right) d\mathbf{x} = \int g(\mathbf{x})\delta(\mathbf{x})d\mathbf{x} = g(0)$, where $\delta(\cdot)$ is the Dirac delta function, which is the Fourier transform of the constant 1 function. See also (Hunter and Nachtergaele, 2001, Chapter 11, Example 11.31).

It remains to prove that $\widehat{f_i\varphi}(\psi) = 0$ for any $\psi \in \mathcal{S}$ supported on the complement of $\mathrm{Span}\{\mathbf{v}_i\} + R_d B_d$. For such ψ, by definition of a convolution, $\mathbf{1}\{R_d B_d\} \star \psi$ is supported on the complement of $\mathrm{Span}\{\mathbf{v}_i\}$. However, μ_i is supported on $\mathrm{Span}\{\mathbf{v}_i\}$ (since if ψ vanishes on $\mathbf{v}_i$, then ψ_i is the zero function, hence $\hat{\psi}_i$ is also the zero function, and $\langle \mu_i, \psi\rangle = \langle f_i, \hat{\psi}_i\rangle = 0$). Therefore, $\langle \mu_i, \mathbf{1}\{R_d B_d\} \star \psi\rangle = 0$, which by the last displayed equation, implies that $\langle \widehat{\tilde{f}_i\varphi}, \psi\rangle = 0$ as required. ■

Lemma 16 *Let q, w be two functions of unit norm in L_2. Suppose that q satisfies*

$$\mathrm{Supp}(q) \subset \bigcup_{j=1}^{k} (\mathrm{Span}\{\mathbf{v}_j\} + R_d B_d) \tag{16}$$

for some $k \in \mathbb{N}$. Moreover, suppose that w is radial and that $\int_{2R_d B_d} w(\mathbf{x})^2 d\mathbf{x} \leq 1 - \delta$ for some $\delta \in [0,1]$. Then

$$\langle q, w\rangle_{L_2} \leq 1 - \delta/2 + k\exp(-cd)$$

where $c > 0$ is a universal constant.

Proof Define $A = (2R_d B_d)^C$ and denote

$$T = \bigcup_{j=1}^{k} (\mathrm{Span}\{\mathbf{v}_j\} + R_d B_d)$$

so that T contains the support of $q(\cdot)$. For each $r > 0$, define

$$h(r) = \frac{Vol_{d-1}(r\mathbb{S}^{d-1} \cap T)}{Vol_{d-1}(r\mathbb{S}^{d-1})},$$

where $\mathbb{S}^{d-1}$ is the Euclidean sphere in $\mathbb{R}^d$. Since T is star shaped, the function $h(r)$ is non-increasing. We claim that there exists a universal constant $c > 0$ such that

$$h(2R_d) \leq k\exp(-cd). \tag{17}$$

Indeed, fix $\mathbf{v} \in \mathbb{S}^{d-1}$ and consider the tube $T_0 = \mathrm{Span}\{\mathbf{v}\} + R_d B_d$. Let $\mathbf{z}$ be a uniform random point in $2R_d\mathbb{S}^{d-1}$. We have by a standard calculation (See e.g., (Sodin, 2007, Section 2))

$$\begin{aligned}
\Pr(\mathbf{z} \in T_0) &= \Pr(\|\mathbf{z}\|^2 - \langle \mathbf{z}, \mathbf{v}\rangle^2 \leq R_d^2) \\
&= \Pr(4R_d^2 - \langle \mathbf{z}, \mathbf{v}\rangle^2 \leq R_d^2) \;=\; \Pr(|\langle \mathbf{z}, \mathbf{v}\rangle| \geq \sqrt{3}R_d) \\
&= \frac{\int_{\sqrt{3}/2}^{1}(1-t^2)^{(d-3)/2}}{\int_0^1 (1-t^2)^{(d-3)/2}} dt \leq \exp(-cd).
\end{aligned}$$

Using a union bound and the definition of h, equation Eq. (17) follows.

Next, define

$$\tilde{q}(\mathbf{x}) = \frac{\int_{\|\mathbf{x}\|\mathbb{S}^{d-1}} q(y) d\mathcal{H}_{d-1}(y)}{Vol_{d-1}(\|\mathbf{x}\|\mathbb{S}^{d-1})}$$

to be the averaging of $q(\cdot)$ with respect to rotations (in the above formula $\mathcal{H}_{d-1}$ denotes the $d-1$ dimensional Hausdorff measure, i.e. the standard measure in $d-1$ dimensions). We have the following: Since $w(\cdot)$ is radial and has unit L_2 norm, and we assume $q(\cdot)$ is supported on T, we have

$$\begin{aligned}
\int_A w(\mathbf{x})q(\mathbf{x})d\mathbf{x} &\overset{(1)}{=} \int_A w(\mathbf{x})\tilde{q}(\mathbf{x})d\mathbf{x} \\
&\overset{(2)}{\leq} \|w\|_{L_2} \|\tilde{q}\mathbf{1}\{A\}\|_{L_2} \\
&\overset{(3)}{=} \sqrt{\int_{2R_d}^{\infty} \tilde{q}(r)^2 Vol_{d-1}(r\mathbb{S}^{d-1})dr} \\
&= \sqrt{\int_{2R_d}^{\infty} Vol_{d-1}(r\mathbb{S}^{d-1}) \left(\frac{1}{Vol_{d-1}(r\mathbb{S}^{d-1})}\int_{r\mathbb{S}^{d-1}\cap T} q(y)d\mathcal{H}_{d-1}(y)\right)^2 dr} \\
&= \sqrt{\int_{2R_d}^{\infty} h(r)^2 Vol_{d-1}(r\mathbb{S}^{d-1}) \left(\frac{1}{Vol_{d-1}(r\mathbb{S}^{d-1}\cap T)}\int_{r\mathbb{S}^{d-1}\cap T} q(y)d\mathcal{H}_{d-1}(y)\right)^2 dr} \\
&\overset{(4)}{\leq} \sqrt{\int_{2R_d}^{\infty} h(r)^2 Vol_{d-1}(r\mathbb{S}^{d-1}) \left(\frac{1}{Vol_{d-1}(r\mathbb{S}^{d-1}\cap T)}\int_{r\mathbb{S}^{d-1}\cap T} q(y)^2 d\mathcal{H}_{d-1}(y)\right) dr} \\
&= \sqrt{\int_{2R_d}^{\infty} h(r)\int_{r\mathbb{S}^{d-1}} q^2(y)d\mathcal{H}_{d-1}(y)dr} \\
&\overset{(5)}{\leq} \sqrt{h(2R_d)}\,\|q\mathbf{1}\{A\}\|_{L_2} \overset{(6)}{\leq} k\exp(-cd/2).
\end{aligned}$$

In the above, (1) follows from $w(\cdot)$ being radial; (2) follows from Cauchy-Schwartz; (3) follows from $w(\cdot)$ having unit L_2 norm; (4) follows from the fact that the term being squared is the expectation of $q(y)$ where y is uniformly distributed in $r\mathbb{S}^{d-1}\cap T$, and we have $(\mathbb{E}_y[q(y)])^2 \leq \mathbb{E}_y[q^2(y)]$ by Jensen's inequality; (5) follows from $h(\cdot)$ being non-increasing; and (6) follows from Eq. (17) and the fact that $q(\cdot)$ has unit L_2 norm.

As a result of these calculations, we have

$$\begin{aligned}
\langle q, w\rangle_{L_2} &= \int_A w(\mathbf{x})q(\mathbf{x})d\mathbf{x} + \int_{A^C} w(\mathbf{x})q(\mathbf{x})d\mathbf{x} \\
&\leq k\exp(-cd/2) + \|q\|_{L_2}\left\|w\mathbf{1}\{A^C\}\right\|_{L_2} = k\exp(-cd/2) + 1\cdot\sqrt{\int_{A^C} w^2(\mathbf{x})d\mathbf{x}} \\
&\leq k\exp(-cd/2) + \sqrt{1-\delta}.
\end{aligned}$$

where we used the assumption that $q(\cdot)$ is unit norm and that $\int_{A^C} w^2(\mathbf{x})d\mathbf{x} = \int_{(2R_dB_d)^C} w^2(\mathbf{x})d\mathbf{x} \leq 1-\delta$. Since $\sqrt{1-\delta} \leq 1-\frac{1}{2}\delta$ for any $\delta\in[0,1]$, the result follows. ■

Proof [Proof of Proposition 13] Define

$$\tilde{g}(\mathbf{x}) = \sum_i \epsilon_i g_i(|\mathbf{x}|)$$

where (ϵ_i) are the signs provided by Lemma 14. According to Lemma 11, we have

$$\|\tilde{g}\|_{L_2(\mu)} \geq c_1/\alpha \tag{18}$$

for a universal constant $c_1 > 0$. Note that the function $\tilde{g}$ is bounded between -1 and 1. Define the function $w = \frac{\widehat{\tilde{g}\varphi}}{\|\tilde{g}\varphi\|_{L_2}}$. By construction (hence according to Lemma 14) we have

$$\begin{aligned}\int_{2R_dB_d} w(\mathbf{x})^2 d\mathbf{x} &= 1 - \frac{\int_{2R_dB_d} \widehat{\tilde{g}\varphi}(\mathbf{x})^2 d\mathbf{x}}{\|\tilde{g}\varphi\|_{L_2}^2} \\ &\leq 1 - \frac{\int_{2R_dB_d} \widehat{\tilde{g}\varphi}(\mathbf{x})^2 d\mathbf{x}}{\|\varphi\|_{L_2}^2} \leq 1 - c_2,\end{aligned}$$

for a universal constant $c_2 > 0$, where in the first inequality we used the fact that $|\tilde{g}(\mathbf{x})| \leq 1$ for all $\mathbf{x}$.

Next, define the function $q = \frac{\widehat{f\varphi}}{\|f\varphi\|_{L_2}}$, where f is an arbitrary function having the form in Eq. (8). Thanks to the assumptions on the functions f_i, we may invoke[8] Claim 15, by which we observe that the functions w, q satisfy the assumptions of Lemma 16. Thus, as a consequence of this lemma we obtain that

$$\langle q, w\rangle_{L_2} \leq 1 - c_2/2 + k\exp(-c_3 d) \tag{19}$$

for a universal constant $c_3 > 0$. Next, we claim that since $\|q\|_{L_2} = \|w\|_{L_2} = 1$, we have for every scalars $\beta_1, \beta_2 > 0$ that

$$\|\beta_1 q - \beta_2 w\|_{L_2} \geq \frac{\beta_2}{2} \|q - w\|_{L_2}. \tag{20}$$

Indeed, we may clearly multiply both β_1 and β_2 by the same constant affecting the correctness of the formula, thus we may assume that $\beta_2 = 1$. It thus amounts to show that for two unit vectors v, u in a Hilbert space, one has that $\min_{\beta>0} \|\beta v - u\|^2 \geq \frac{1}{4}\|v - u\|^2$. We have

$$\begin{aligned}\min_{\beta} \|\beta v - u\|^2 &= \min_{\beta} \left(\beta^2\|v\|^2 - 2\beta\langle v, u\rangle + \|u\|^2\right) \\ &= \min_{\beta} \left(\beta^2 - 2\beta\langle v, u\rangle + 1\right) \\ &= 1 - \langle v, u\rangle^2 = \frac{1}{2}\|v - u\|^2\end{aligned}$$

which in particular implies formula Eq. (20).

Combining the above, and using the fact that q, w have unit L_2 norm, we finally get

$$\begin{aligned}\|f - \tilde{g}\|_{L_2(\mu)} &= \|f\varphi - \tilde{g}\varphi\|_{L_2} = \left\|\widehat{f\varphi} - \widehat{\tilde{g}\varphi}\right\|_{L_2} = \left\|\left(\|f\varphi\|_{L_2}\right) q(\cdot) - \left(\|\tilde{g}\varphi\|_{L_2}\right) w(\cdot)\right\|_{L_2} \\ &\overset{Eq.\ (20)}{\geq} \frac{1}{2}\|q - w\|_{L_2}\|\tilde{g}\varphi\|_{L_2} = \frac{1}{2}\|q - w\|_{L_2}\|\tilde{g}\|_{L_2(\mu)} \\ &\overset{Eq.\ (18)}{\geq} \frac{1}{2}\sqrt{2(1 - \langle q, w\rangle_{L_2})}\frac{c_1}{\alpha} \\ &\overset{Eq.\ (19)}{\geq} \frac{c_1}{2\alpha}\sqrt{2\max(c_2/2 - k\exp(-c_3 d), 0)} \geq \frac{c_1\sqrt{c_2}}{4\alpha}\end{aligned}$$

8. Claim 15 also requires that $f\varphi$ is an L_2 function, but we can assume this without loss of generality: Otherwise, $\|f\varphi\|_{L_2} = \infty$, and since $\tilde{g}\varphi$ is an L_2 function with $\|\tilde{g}\varphi\|_{L_2} < \infty$, we would have $\|f - \tilde{g}\|_{L_2(\mu)}^2 = \|f\varphi - \tilde{g}\varphi\|_{L_2}^2 = \infty$, in which case the proposition we wish to prove is trivially true.

where in the last inequality, we use the assumption in Eq. (7), choosing $c = \min\{c_2/4, c_3\}$. The proof is complete. ■

4.4. Approximability of the Function $\tilde{g}$ with 3-Layer Networks

The next ingredient missing for our proof is the construction of a 3-layer function which approximates the function $\tilde{g} = \sum_{i=1}^{N} \epsilon_i g_i$.

Proposition 17 *There is a universal constant $C > 0$ such that the following holds. Let $\delta \in (0,1)$. Suppose that $d \geq C$ and that the functions g_i are constructed as in Eq. (6). For any choice of $\epsilon_i \in \{-1,+1\}$, $i = 1,\ldots,N$, there exists a function g expressible by a 3-layer network of width at most $\frac{8c_\sigma}{\delta}\alpha^{3/2}Nd^{11/4}+1$, and with range in $[-2,+2]$, such that*

$$\left\| g(\mathbf{x}) - \sum_{i=1}^{N} \epsilon_i g_i(\|\mathbf{x}\|) \right\|_{L_2(\mu)} \leq \frac{\sqrt{3}}{\alpha d^{1/4}} + \delta.$$

The proof of this proposition relies on assumption 1, which ensures that we can approximate univariate functions using our activation function. As discussed before Thm. 1, one can also plug in weaker versions of the assumption (i.e. worse polynomial dependence of the width w on $R, L, 1/\delta$), and get versions of proposition 17 where the width guarantee has worse polynomial dependence on the parameters N, α, d, δ. This would lead to versions of the Thm. 1 with somewhat worse constants and polynomial dependence on the dimension d.

For this proposition, we need a simple intermediate result, in which an approximation for radially symmetric Lipschitz functions in $\mathbb{R}^d$, using assumption 1, is constructed.

Lemma 18 *Suppose the activation function σ satisfies assumption 1. Let f be an L-Lipschitz function supported on $[r,R]$, where $r \geq 1$. Then for any $\delta > 0$, there exists a function g expressible by a 3-layer network of width at most $\frac{2c_\sigma d^2 R^2 L}{\sqrt{r}\delta}+1$, such that*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left| g(\mathbf{x}) - f(\|\mathbf{x}\|) \right| < \delta.$$

Proof Define the $2R$-Lipschitz function

$$l(x) = \min\{x^2, R^2\},$$

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

$$\ell(\mathbf{x}) = \sum_{i=1}^{d} l(x_i) = \sum_{i=1}^{d} \min\{x_i^2, R^2\}$$

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

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

and where the width parameter w is at most $\frac{2c_\sigma dR^2 L}{\sqrt{r}\delta}$. Consequently, the function

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

can be expressed in the form $a + \sum_{i=1}^{w} \alpha_i \sigma(\beta_i x - \gamma_i)$ where $w \leq \frac{2c_\sigma d^2 R^2 L}{\sqrt{r}\delta}$, and it holds that

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left|\tilde{\ell}(\mathbf{x}) - \ell(\mathbf{x})\right| \leq \frac{\sqrt{r}\delta}{L}. \tag{21}$$

Next, define

$$s(x) = \begin{cases} f(\sqrt{x}) & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Since f is L-Lipschitz and supported on $\{x : r \leq x \leq R\}$, it follows that s is $\frac{L}{2\sqrt{r}}$-Lipschitz and supported on the interval $[-R^2, R^2]$. Invoking assumption 1 again, we can construct a function $\tilde{s}(x) = a + \sum_{i=1}^{w} \alpha_i \sigma(\beta_i x - \gamma_i)$ satisfying

$$\sup_{x\in\mathbb{R}} |\tilde{s}(x) - s(x)| < \delta/2, \tag{22}$$

where $w \leq \frac{c_\sigma R^2 L}{\sqrt{r}\delta}$.

Now, let us consider the composition $g = \tilde{s} \circ \tilde{\ell}$, which by definition of $\tilde{s}, \tilde{\ell}$, has the form

$$a + \sum_{i=1}^{w} u_i \sigma\left(\sum_{j=1}^{w} v_{i,j}\sigma\left(\langle \mathbf{w}_{i,j}, \mathbf{x}\rangle + b_{i,j}\right) + c_i\right) \tag{23}$$

for appropriate scalars $a, u_i, c_i, v_{i,j}, b_{i,j}$ and vectors $\mathbf{w}_{i,j}$, and where w is at most

$$\max\left\{\frac{2c_\sigma d^2 R^2 L}{\sqrt{r}\delta}, \frac{c_\sigma R^2 L}{\sqrt{r}\delta}\right\} \;=\; \frac{2c_\sigma d^2 R^2 L}{\sqrt{r}\delta}.$$

Eq. (23) is exactly a 3-layer network (compare to Eq. (2)), except that there is an additional constant term a. However, by increasing w by 1, we can simulate a by an additional neuron $\mathbf{x} \mapsto \frac{a}{\sigma(\sigma(0)+z)} \cdot \sigma(\sigma(\langle \mathbf{0}, \mathbf{x}\rangle) + z)$, where z is some scalar such that $\sigma(\sigma(0) + z) \neq 0$ (note that if there is no such z, then σ is the zero function, and therefore cannot satisfy assumption 1). So, we can write the function g as a 3-layer network (as defined in Eq. (2)), of width at most

$$\frac{2c_\sigma d^2 R^2 L}{\sqrt{r}\delta} + 1.$$

All the remains now is to prove that $\sup_{\mathbf{x}\in\mathbb{R}^d} |g(\mathbf{x}) - f(\|\mathbf{x}\|)| \leq \delta$. To do so, we note that for any $\mathbf{x} \in \mathbb{R}^d$, we have

$$\begin{aligned}
|g(\mathbf{x}) - f(\|\mathbf{x}\|)| &= \left|\tilde{s}(\tilde{\ell}(\mathbf{x})) - f(\|\mathbf{x}\|)\right| \\
&\leq \left|\tilde{s}(\tilde{\ell}(\mathbf{x})) - s(\tilde{\ell}(\mathbf{x})\right| + \left|s(\tilde{\ell}(\mathbf{x}) - s(\ell(\mathbf{x}))\right| + |s(\ell(\mathbf{x})) - f(\|\mathbf{x}\|)| \\
&= \left|\tilde{s}(\tilde{\ell}(\mathbf{x})) - s(\tilde{\ell}(\mathbf{x})\right| + \left|s(\tilde{\ell}(\mathbf{x}) - s(\ell(\mathbf{x}))\right| + \left|f(\sqrt{\ell(\mathbf{x})}) - f(\|\mathbf{x}\|)\right|.
\end{aligned}$$

Let us consider each of the three absolute values:

- The first absolute value term is at most $\delta/2$ by Eq. (22).
- The second absolute value term, since s is $\frac{L}{2\sqrt{r}}$-Lipschitz, is at most $\frac{L}{2\sqrt{r}}|\tilde{\ell}(\mathbf{x}) - \ell(\mathbf{x})|$, which is at most $\delta/2$ by Eq. (21).
- As to the third absolute value term, if $\|\mathbf{x}\|^2 \leq R^2$, then $\ell(\mathbf{x}) = \|\mathbf{x}\|^2$ and the term is zero. If $\|\mathbf{x}\|^2 > R^2$, then it is easy to verify that $\ell(\mathbf{x}) \geq R^2$, and since f is continuous and supported on $[r, R]$, it follows that $f(\sqrt{\ell(\mathbf{x})} = f(\|\mathbf{x}\|) = 0$, and again, we get a zero.

Summing the above, we get that $|g(\mathbf{x}) - f(\|\mathbf{x}\|)| \leq \frac{\delta}{2} + \frac{\delta}{2} + 0 = \delta$ as required. ■

We are now ready to prove Proposition 17, which is essentially a combination of Lemmas 12 and 18.

Proof [Proof of Proposition 17] First, invoke Lemma 12 to obtain an N-Lipschitz function h with range in $[-1, +1]$ which satisfies

$$\left\| h(\mathbf{x}) - \sum_{i=1}^{N} \epsilon_i g_i(\mathbf{x}) \right\|_{L_2(\mu)} = \sqrt{\int_{\mathbb{R}^d} \left(\tilde{f}(\mathbf{x}) - \sum_{i=1}^{N} \epsilon_i g_i(\mathbf{x}) \right)^2 \varphi^2(\mathbf{x}) d\mathbf{x}} \leq \frac{\sqrt{3}}{\alpha d^{1/4}}. \tag{24}$$

Next, we use Lemma 18 with $R = 2\alpha\sqrt{d}$, $r = \alpha\sqrt{d}$, $L = N$ to construct a function g expressible by a 3-layer network of width at most $\frac{8c_\sigma}{\delta}\alpha^{3/2}Nd^{11/4} + 1$, satisfying $\sup_{\mathbf{x}\in\mathbb{R}^d} |g(\mathbf{x}) - h(\mathbf{x})| \leq \delta$. This implies that $\|g - h\|_{L_2(\mu)} \leq \delta$, and moreover, that the range of g is in $[-1-\delta, 1+\delta] \subseteq [-2, +2]$ (since we assume $\delta < 1$). Combining with Eq. (24) and using the triangle inequality finishes the proof. ■

4.5. Finishing the Proof

We are finally ready to prove our main theorem.

Proof [Proof of Theorem 1] The proof is a straightforward combination of propositions 13 and 17 (whose conditions can be verified to follow immediately from the assumptions used in the theorem). We first choose $\alpha = C$ and $N = \lceil C\alpha^{3/2}d^2 \rceil$ with the constant C taken from the statement of Proposition 13. By invoking this proposition we obtain signs $\epsilon_i \in \{-1, 1\}$ and a universal constant $\delta_1 > 0$ for which any function f expressed by a bounded-size 2-layer network satisfies

$$\|\tilde{g} - f\|_{L_2(\mu)} \geq \delta_1, \tag{25}$$

where $\tilde{g}(\mathbf{x}) = \sum_{i=1}^{N} \epsilon_i g_i(\|\mathbf{x}\|)$. Next, we use Proposition 17 with $\delta = \min\{\delta_1/2, 1\}$ to approximate $\tilde{g}$ by a function g expressible by a 3-layer network of width at most

$$\frac{16c_\sigma}{\delta}\alpha^{3/2}Nd^{11/4} + 1 = \frac{16c_\sigma}{\delta}C^{3/2}\lceil C^{5/2}d^2 \rceil d^{11/4} + 1 \leq C' c_\sigma d^{19/4},$$

(where C' is a universal constant depending on the universal constants C, δ_1), so that

$$\|\tilde{g} - g\|_{L_2(\mu)} \leq \delta \leq \delta_1/2. \tag{26}$$

Combining Eq. (25) and Eq. (26) with the triangle inequality, we have that $\|f - g\|_{L_2(\mu)} \geq \delta_1/2$ for any 2-layer function f. The proof is complete. ■

Acknowledgments

OS is supported in part by an FP7 Marie Curie CIG grant, the Intel ICRI-CI Institute, and Israel Science Foundation grant 425/13. We thank James Martens and the anonymous COLT 2016 reviewers for several helpful comments.

References

Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

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

N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. *arXiv preprint arXiv:1509.05009*, 2015.

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

C. Debao. Degree of approximation by superpositions of a sigmoidal function. *Approximation Theory and its Applications*, 9(3):17–28, 1993.

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

DLMF. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.10 of 2015-08-07, 2015. URL http://dlmf.nist.gov/.

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

Loukas Grafakos and Gerald Teschl. On fourier transforms of radial functions and distributions. *Journal of Fourier Analysis and Applications*, 19(1):167–179, 2013.

András Hajnal, Wolfgang Maass, Pavel Pudlák, Mário Szegedy, and György Turán. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46(2):129–154, 1993.

J. Håstad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 6–20. ACM, 1986.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

John K. Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Co., Inc., River Edge, NJ, 2001.

Ilia Krasikov. Approximations for the bessel and airy functions with an explicit error term. *LMS Journal of Computation and Mathematics*, 17(01):209–225, 2014.

W. Maass, G. Schnitger, and E. Sontag. A comparison of the computational power of sigmoid and boolean threshold circuits. In V. P. Roychowdhury, K. Y. Siu, and A. Orlitsky, editors, *Theoretical Advances in Neural Computation and Learning*, pages 127–151. Springer, 1994.

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

James Martens. Private Communication, 2015.

James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted boltzmann machines. In *NIPS*, pages 2877–2885, 2013.

G. F Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2924–2932, 2014.

I. Parberry. *Circuit complexity and neural networks*. MIT press, 1994.

R. Pascanu, G. Montufar, and Y. Bengio. On the number of inference regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv*, 1312, 2013.

B. Rossman, R. Servedio, and L.-Y. Tan. An average-case depth hierarchy theorem for boolean circuits. In *FOCS*, 2015.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J.L. McClelland, editors, *Parallel distributed Processing*, volume 1. MIT Press, 1986.

A. Shpilka and A. Yehudayoff. Arithmetic circuits: A survey of recent results and open questions. *Foundations and Trends® in Theoretical Computer Science*, 5(3–4):207–388, 2010.

Sasha Sodin. Tail-Sensitive Gaussian Asymptotics for Marginals of Concentrated Measures in High Dimension. In Vitali D. Milman and Gideon Schechtman, editors, *Geometric Aspects of Functional Analysis*, volume 1910 of *Lecture Notes in Mathematics*, pages 271–295. Springer Berlin Heidelberg, 2007.

M. Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

Appendix A. Approximation Properties of the ReLU Activation Function

In this appendix, we prove that the ReLU activation function satisfies assumption 1, and also prove bounds on the Lipschitz parameter of the approximation and the size of the required parameters. Specifically, we have the following lemma:

Lemma 19 *Let $\sigma(z) = \max\{0, z\}$ be the ReLU activation function, and fix $L, \delta, R > 0$. Let $f : \mathbb{R} \to \mathbb{R}$ which is constant outside an interval $[-R, R]$. There exist scalars $a, \{\alpha_i, \beta_i\}_{i=1}^{w}$, where $w \leq 3\frac{RL}{\delta}$, such that the function*

$$h(x) = a + \sum_{i=1}^{w} \alpha_i \sigma(x - \beta_i) \tag{27}$$

is L-Lipschitz and satisfies

$$\sup_{x \in \mathbb{R}} \left| f(x) - h(x) \right| \leq \delta. \tag{28}$$

Moreover, one has $|\alpha_i| \leq 2L$ and $w \leq 3\frac{RL}{\delta}$.

Proof If one has $2RL < \delta$, then the results holds trivially because we may take the function h to be the 0 function (with width parameter $w = 0$). Otherwise, we must have $R \geq \delta/2L$, so by increasing the value of R by a factor of at most 2, we may assume without loss of generality that there exists an integer m such that $R = m\delta/L$.

Let h be the unique piecewise linear function which coalesces with f on points of the form δ/Li, $i \in \mathbb{Z} \cap [-m, m]$, is linear in the intervals $(w\delta/L, (w+1)\delta/L)$ and is constant outside $[-R, R]$. Since f is L-Lipschitz, equation Eq. (28) holds true. It thus suffices to express h as a function having the form Eq. (27). Let $\beta_i = i\delta/L$, choose $a = h(-R)$ and set

$$\alpha_i = h'(\beta_i + \tfrac{\delta}{2L}) - h'(\beta_i - \tfrac{\delta}{2L}), \quad -m \leq i \leq m.$$

Then clearly equation Eq. (27) holds true. Moreover, we have $|\alpha_i| \leq 2L$, which completes the proof.

■

Appendix B. Technical Proofs

B.1. Proof of Lemma 6

By definition of the Fourier transform,

$$\varphi(\mathbf{x}) = \int_{\mathbf{w}:\|\mathbf{w}\| \leq R_d} \exp(-2\pi i \mathbf{x}^\top \mathbf{w}) d\mathbf{w}.$$

Since $\varphi(\mathbf{x})$ is radial (hence rotationally invariant), let us assume without loss of generality that it equals $r\mathbf{e}_1$, where $r = \|\mathbf{x}\|$ and $\mathbf{e}_1$ is the first standard basis vector. This means that the integral becomes

$$\int_{\mathbf{w}:\|\mathbf{w}\| \leq R_d} \exp(-2\pi i r w_1) d\mathbf{w} = \int_{w_1=-R_d}^{R_d} \exp(-2\pi i r w_1) \left(\int_{w_2 \ldots w_d : \sum_{j=2}^{d} w_j^2 \leq R_d^2 - w_1^2} dw_2 \ldots dw_d \right) dw_1.$$

The expression inside the parenthesis is simply the volume of a ball of radius $\left(R_d^2 - w_1^2\right)^{1/2}$ in $\mathbb{R}^{d-1}$. Letting V_{d-1} be the volume of a unit ball in $\mathbb{R}^{d-1}$, this equals

$$\int_{w_1=-R_d}^{R_d} \exp(-2\pi i r w_1) \left(V_{d-1}(R_d^2 - w_1^2)^{\frac{d-1}{2}}\right) dw_1.$$

Performing the variable change $z = \arccos(w_1/R_d)$ (which implies that as w_1 goes from $-R_d$ to R_d, z goes from π to 0, and also $R_d \cos(z) = w_1$ and $-R_d \sin(z) dz = dw_1$), we can rewrite the integral above as

$$\begin{aligned} &V_{d-1} \int_{z=0}^{\pi} \left(R_d^2 - R_d^2 \cos^2(z)\right)^{\frac{d-1}{2}} \exp(-2\pi i r R_d \cos(z)) R_d \sin(z) dz \\ &= V_{d-1} R_d^d \int_{z=0}^{\pi} \sin^d(z) \exp\left(-2\pi i r R_d \cos(z)\right) dz. \end{aligned}$$

Since we know that this integral must be real-valued (since we're computing the Fourier transform $\varphi(\mathbf{x})$, which is real-valued and even), we can ignore the imaginary components, so the above reduces to

$$V_{d-1} R_d^d \int_{z=0}^{\pi} \sin^d(z) \cos\left(2\pi r R_d \cos(z)\right) dz. \tag{29}$$

By a standard formula for Bessel functions (see Equation 10.9.4. in DLMF), we have

$$J_{d/2}(x) = \frac{(x/2)^{d/2}}{\pi^{1/2} \Gamma\left(\frac{d+1}{2}\right)} \int_0^{\pi} \sin^d(z) \cos(x \cos(z)) dz,$$

which by substituting $x = 2\pi r R_d$ and changing sides, implies that

$$\int_0^{\pi} \sin^d(z) \cos(2\pi r R_d \cos(z)) dz = \frac{\pi^{1/2} \Gamma\left(\frac{d+1}{2}\right)}{(\pi r R_d)^{d/2}} J_{d/2}(2\pi r R_d).$$

Plugging this back into Eq. (29), we get the expression

$$V_{d-1} R_d^{d/2} \frac{\pi^{1/2} \Gamma\left(\frac{d+1}{2}\right)}{(\pi r)^{d/2}} J_{d/2}(2\pi r R_d).$$

Plugging in the explicit formula $V_{d-1} = \frac{\pi^{(d-1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}$, this simplifies to

$$\left(\frac{R_d}{r}\right)^{d/2} J_{d/2}(2\pi R_d r).$$

Recalling that this equals $\varphi(x)$ where $\|\mathbf{x}\| = r$, the result follows.

B.2. Proof of Lemma 8

By Lemma 6,

$$\varphi(x) = \left(\frac{R_d}{x}\right)^{d/2} J_{d/2}(2\pi R_d x).$$

Moreover, using the definition of a good interval, and the fact that the maximal value in any interval is at most $2\alpha\sqrt{d}$, we have

$$|J_{d/2}(2\pi R_d x)| \geq \frac{1}{\sqrt{80\pi R_d x}} \geq \frac{1}{\sqrt{160\pi R_d \alpha\sqrt{d}}}. \tag{30}$$

Since x (in any interval) is at least $\alpha\sqrt{d}$, then $J_{d/2}(2\pi R_d x)$ is $2\pi R_d$-Lipschitz in x by Lemma 20. Since the width of each interval only $\frac{\alpha\sqrt{d}}{N}$, Eq. (30) implies that $J_{d/2}(2\pi R_d x)$ (and hence $\varphi(x)$) does not change signs in the interval, provided that $N > 2\sqrt{160}\left(\pi\alpha R_d\sqrt{d}\right)^{3/2}$. Recalling that $R_d \leq \frac{1}{2}\sqrt{d}$, this is indeed satisfied by the lemma's conditions.

Turning to the second part of the lemma, assuming $\varphi(x)$ is positive without loss of generality, and using the Lipschitz property of $J_{d/2}(\cdot)$ and Eq. (30), we have

$$\begin{aligned}
\frac{\sup_{x\in\Delta_i}\varphi(x)}{\inf_{x\in\Delta_i}\varphi(x)} &\leq \frac{\sup_{x\in\Delta_i}\left(\frac{R_d}{x}\right)^{d/2}}{\inf_{x\in\Delta_i}\left(\frac{R_d}{x}\right)^{d/2}} \cdot \frac{\sup_{x\in\Delta_i} J_{d/2}(2\pi R_d x)}{\inf_{x\in\Delta_i} J_{d/2}(2\pi R_d x)} \\
&\leq \left(\frac{\sup_{x\in\Delta_i} x}{\inf_{x\in\Delta_i} x}\right)^{d/2} \cdot \frac{\inf_{x\in\Delta_i} J_{d/2}(2\pi R_d x) + \frac{2\pi R_d\alpha\sqrt{d}}{N}}{\inf_{x\in\Delta_i} J_{d/2}(2\pi R_d x)} \\
&\leq \left(\frac{\inf_{x\in\Delta_i} x + \frac{\alpha\sqrt{d}}{N}}{\inf_{x\in\Delta_i} x}\right)^{d/2} \left(1 + \frac{2\pi R_d\alpha\sqrt{d}}{N}\sqrt{80\pi R_d\alpha\sqrt{d}}\right) \\
&\leq \left(1 + \frac{\alpha\sqrt{d}}{N\alpha\sqrt{d}}\right)^{d/2} \left(1 + \frac{2\sqrt{80}(\pi\alpha R_d\sqrt{d})^{3/2}}{N}\right) \\
&\leq \left(1 + \frac{1}{N}\right)^{d/2} \left(1 + \frac{2\sqrt{80}(\pi\alpha d/2)^{3/2}}{N}\right),
\end{aligned}$$

which is less than $1 + d^{-1/2}$ provided that $N \geq c\alpha^{3/2}d^2$ for some universal constant c.

B.3. Proof of Lemma 9

The result is trivially true for a bad interval i (where g_i is the 0 function, hence both sides of the inequality in the lemma statement are 0), so we will focus on the case that i is a good interval.

For simplicity, let us denote the interval Δ_i as $[\ell, \ell+\delta]$, where $\delta = \frac{1}{N}$ and ℓ is between $\alpha\sqrt{d}$ and $2\alpha\sqrt{d}$. Therefore, the conditions in the lemma imply that $\delta \leq \frac{1}{50d\ell}$. Also, we drop the i subscript and refer to g_i as g.

Since, g is a radial function, its Fourier transform is also radial, and is given by

$$\hat{g}(\mathbf{w}) = \hat{g}(\|\mathbf{w}\|) = 2\pi\int_{s=0}^{\infty} g(s)\left(\frac{s}{\|\mathbf{w}\|}\right)^{d/2-1} J_{d/2-1}(2\pi s\|\mathbf{w}\|)s\,ds,$$

(see for instance Grafakos and Teschl (2013), section 2, and references therein). Using this formula, and switching to polar coordinates (letting A_d denote the surface area of a unit sphere in $\mathbb{R}^d$), we

have the following:

$$\begin{aligned}\int_{2R_dB_d}\hat{g}^2(\mathbf{w})d\mathbf{w} &= \int_{r=0}^{2R_d}A_dr^{d-1}\hat{g}^2(r)dr \\ &= \int_{r=0}^{2R_d}A_dr^{d-1}\left(2\pi\int_{s=0}^{\infty}g(s)\left(\frac{s}{r}\right)^{d/2-1}J_{d/2-1}(2\pi sr)s\,ds\right)^2dr \\ &= 4\pi^2A_d\int_{r=0}^{2R_d}r\left(\int_{s=0}^{\infty}g(s)s^{d/2}J_{d/2-1}(2\pi sr)\,ds\right)^2dr \\ &= 4\pi^2A_d\int_{r=0}^{2R_d}r\left(\int_{s=\ell}^{\ell+\delta}s^{d/2}J_{d/2-1}(2\pi sr)\,ds\right)^2dr. \end{aligned} \tag{31}$$

By Lemma 20, $|J_{d/2-1}(x)|\leq 1$, hence Eq. (31) can be upper bounded by

$$\begin{aligned}&4\pi^2A_d\int_{r=0}^{2R_d}r\left(\int_{s=\ell}^{\ell+\delta}s^{d/2}ds\right)^2dr \;\leq\; 4\pi^2A_d\int_{r=0}^{2R_d}r\left(\delta(\ell+\delta)^{d/2}\right)^2dr \\ &\leq 4\pi^2A_d\delta^2(\ell+\delta)^d\int_{r=0}^{2R_d}r\,dr \;=\; 8\pi^2A_d\delta^2(\ell+\delta)^dR_d^2.\end{aligned}$$

Overall, we showed that

$$\int_{2R_dB_d}\hat{g}^2(\mathbf{w})d\mathbf{w} \;\leq\; 8\pi^2R_d^2A_d\delta^2(\ell+\delta)^d. \tag{32}$$

Let us now turn to consider $\int\hat{g}^2(\mathbf{w})d\mathbf{w}$, where the integration is over all of $\mathbf{w}\in\mathbb{R}^d$. By isometry of the Fourier transform, this equals $\int g^2(\mathbf{x})d\mathbf{x}$, so

$$\int\hat{g}^2(\mathbf{w})d\mathbf{w} \;=\; \int_{\mathbb{R}^d}g^2(\mathbf{x})d\mathbf{x} \;=\; \int_{r=0}^{\infty}A_dr^{d-1}g^2(r)dr \;=\; \int_{r=\ell}^{\ell+\delta}A_dr^{d-1}dr \;\geq\; A_d\delta\ell^{d-1}.$$

Combining this with Eq. (32), we get that

$$\frac{\int_{2R_dB_d}\hat{g}^2(\mathbf{w})d\mathbf{w}}{\int_{\mathbb{R}^d}\hat{g}^2(\mathbf{w})d\mathbf{w}} \;\leq\; \frac{8\pi^2R_d^2A_d\delta^2(\ell+\delta)^d}{A_d\delta\ell^{d-1}} \;=\; 8\pi^2R_d^2\ell\delta\left(1+\frac{\delta}{\ell}\right)^d.$$

Since we assume $\delta\leq\frac{1}{50d\ell}$, and it holds that $\left(1+\frac{1}{50d}\right)^d\leq\exp(1/50)$ and $R_d\leq\frac{1}{2}\sqrt{d}$ by Lemma 5, the above is at most

$$2\pi^2d\ell\delta\left(1+\frac{1}{50d}\right)^d \;\leq\; 2\pi^2d\ell\delta\exp(1/50) \;\leq\; 2\pi^2\frac{1}{50}\exp(1/50) < \frac{1}{2}.$$

Overall, we showed that $\frac{\int_{2R_dB_d}\hat{g}^2(\mathbf{w})d\mathbf{w}}{\int_{\mathbb{R}^d}\hat{g}^2(\mathbf{w})d\mathbf{w}}\leq\frac{1}{2}$, and therefore

$$\frac{\int_{(2R_dB_d)^C}\hat{g}^2(\mathbf{w})d\mathbf{w}}{\int_{\mathbb{R}^d}\hat{g}^2(\mathbf{w})d\mathbf{w}} \;=\; \frac{\int_{\mathbb{R}^d}\hat{g}^2(\mathbf{w})d\mathbf{w}-\int_{2R_dB_d}\hat{g}^2(\mathbf{w})d\mathbf{w}}{\int_{\mathbb{R}^d}\hat{g}^2(\mathbf{w})d\mathbf{w}} \;\geq\; 1-\frac{1}{2} \;=\; \frac{1}{2}$$

as required.

B.4. Proof of Lemma 10

The result is trivially true for a bad interval i (where g_i is the 0 function, hence both sides of the inequality in the lemma statement are 0), so we will focus on the case that i is a good interval.

Define $a = \sup_{x \in \Delta_i} \varphi(x)$. Using Lemma 8, we have that $\varphi(x)$ does not change signs in the interval Δ_i. Suppose without loss of generality that it is positive. Moreover, by the same lemma we have that

$$|\varphi(x) - a| \leq d^{-1/2} a, \;\; \forall x \in \Delta_i$$

Consequently, we have that

$$\begin{aligned} \int_{(2R_d B_d)^C} (\widehat{((\varphi - a) g_i)}(\mathbf{w}))^2 d\mathbf{w} \;&\leq \int_{\mathbb{R}^d} (\widehat{((\varphi - a) g_i)}(\mathbf{w}))^2 d\mathbf{w} \\ &= \int_{\mathbb{R}^d} ((\varphi - a) g_i(x))^2 dx \\ &\leq d^{-1} \int_{\mathbb{R}^d} (a g_i(x))^2 dx. \end{aligned} \tag{33}$$

Next, by choosing the constant C to be large enough, we may apply Lemma 9, which yields that

$$\int_{(2R_d B_d)^C} (\widehat{(a g_i)}(\mathbf{w}))^2 d\mathbf{w} \geq \frac{1}{2} \int_{\mathbb{R}^d} (a g_i(x))^2 dx. \tag{34}$$

By the triangle inequality, we have that for two vectors u, v in a normed space, one has $\|v\|^2 \geq \|u\|^2 - 2\|v\|\|v - u\|$. This teaches us that

$$\begin{aligned} \int_{(2R_d B_d)^C} (\widehat{(g_i \varphi)}(\mathbf{w}))^2 d\mathbf{w} \;&\geq \int_{(2R_d B_d)^C} (\widehat{(a g_i)}(\mathbf{w}))^2 d\mathbf{w} \\ &\quad - 2\sqrt{\int_{(2R_d B_d)^C} (\widehat{(a g_i)}(\mathbf{w}))^2 d\mathbf{w}} \sqrt{\int_{(2R_d B_d)^C} (\widehat{((\varphi - a) g_i)}(\mathbf{w}))^2 d\mathbf{w}} \\ &\overset{Eq.\ (33)}{\geq} \int_{(2R_d B_d)^C} (\widehat{(a g_i)}(\mathbf{w}))^2 d\mathbf{w} - 2d^{-1/2} \int_{\mathbb{R}^d} (a g_i(x))^2 dx \\ &\overset{Eq.\ (34)}{\geq} \frac{1}{2}(1 - 4d^{-1/2}) \int_{\mathbb{R}^d} (a g_i(x))^2 dx \geq \frac{1}{4} \int_{\mathbb{R}^d} (\varphi(x) g_i(x))^2 dx. \end{aligned}$$

B.5. Proof of Lemma 11

Since the g_i for different i have disjoint supports (up to measure-zero sets), the integral in the lemma equals

$$\int \sum_{i=1}^N (\epsilon_i g_i(\mathbf{x}))^2 \varphi^2(\mathbf{x}) d\mathbf{x} \;=\; \int \sum_{i=1}^N g_i^2(\mathbf{x}) \varphi^2(\mathbf{x}) d\mathbf{x} \;=\; \int_{\mathbf{x}: \|\mathbf{x}\| \in \text{good } \Delta_i} \varphi^2(\mathbf{x}) d\mathbf{x},$$

where we used the definition of g_i. Switching to polar coordinates (letting A_d be the surface area of the unit sphere in $\mathbb{R}^d$), and using the definition of φ from Lemma 6, this equals

$$A_d \int_{r \in \text{good } \Delta_i} r^{d-1} \varphi^2(r) dr \;=\; A_d \int_{r \in \text{good } \Delta_i} \frac{R_d^d}{r} J_{d/2}^2(2\pi R_d r) dr$$

Recalling that $A_d = \frac{d\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ and that $R_d^d = \pi^{-d/2}\Gamma\left(\frac{d}{2}+1\right)$ by Lemma 5, this equals

$$d\int_{r\in\text{good }\Delta_i} \frac{J_{d/2}^2(2\pi R_d r)}{r} dr. \tag{35}$$

We now claim that for any $r \in [\alpha\sqrt{d}, 2\alpha\sqrt{d}]$ (that is, in any interval),

$$J_{d/2}^2(2\pi R_d r) \geq \frac{1}{40\pi R_d r} \quad \Longrightarrow \quad r \in \text{ good } \Delta_i, \tag{36}$$

which would imply that we can lower bound Eq. (35) by

$$d\int_{\alpha\sqrt{d}}^{2\alpha\sqrt{d}} \frac{J_{d/2}^2(2\pi R_d r)}{r} \mathbf{1}\left\{J_{d/2}^2(2\pi R_d r) \geq \frac{1}{40\pi R_d r}\right\} dr. \tag{37}$$

To see why Eq. (36) holds, consider an r which satisfies the left hand side. The width of its interval is at most $\frac{\alpha\sqrt{d}}{N}$, and by Lemma 20, $J_{d/2}(2\pi R_d r)$ is at most $2\pi R_d$-Lipschitz in r. Therefore, for any other r' in the same interval as r, it holds that

$$\left|J_{d/2}(2\pi R_d r')\right| \geq \sqrt{\frac{1}{40\pi R_d r}} - \frac{2\pi R_d \alpha\sqrt{d}}{N},$$

which can be verified to be at least $\sqrt{\frac{1}{80\pi R_d r}}$ by the condition on N in the lemma statement, and the facts that $r \leq 2\alpha\sqrt{d}, R_d \leq \frac{1}{2}\sqrt{d}$. As a result, $J_{d/2}^2(2\pi R_d r') \geq \frac{1}{80\pi R_d r}$ for any r' in the same interval as r, which implies that r is in a good interval.

We now continue by taking Eq. (37), and performing the variable change $x = 2\pi R_d r$, leading to

$$d\int_{2\pi R_d\alpha\sqrt{d}}^{4\pi R_d\alpha\sqrt{d}} \frac{J_{d/2}^2(x)}{x} \mathbf{1}\left\{J_{d/2}^2(x) \geq \frac{1}{20x}\right\} dx.$$

Applying Lemma 23 with $\beta = 2\pi R_d\alpha/\sqrt{d}$ (which by Lemma 5, is between $2\pi\alpha/5$ and $\pi\alpha$, hence satisfies the conditions of Lemma 23 if α is large enough), this is at least

$$d\frac{0.005}{\beta d} \geq \frac{0.005}{2\pi\alpha/5} \geq \frac{0.003}{\alpha},$$

from which the lemma follows.

B.6. Proof of Lemma 12

For any i, define

$$\check{g}_i(x) = \begin{cases} \max\{1, N\text{dist}(x, \Delta_i^C)\} & i \text{ good} \\ 0 & i \text{ bad} \end{cases}$$

where $\text{dist}(x, \Delta_i^C)$ is the distance of x from the boundaries of Δ_i. Note that for bad i, this is the same as $g_i(x)$, whereas for good i, it is an N-Lipschitz approximation of $g_i(x)$.

Let $f(\mathbf{x}) = \sum_{i=1}^N \epsilon \check{g}(\mathbf{x})$, and note that since the support of $\check{g}_i$ are disjoint, f is also N Lipschitz. With this definition, the integral in the lemma becomes

$$\int \left(\sum_{i=1}^N \epsilon_i (\check{g}_i(\mathbf{x}) - g_i(\mathbf{x})) \right)^2 \varphi^2(\mathbf{x}) d\mathbf{x}.$$

Since the support of $\check{g}_i(\mathbf{x}) - g_i(\mathbf{x})$ is disjoint for different i, this equals

$$\int \sum_{i=1}^N (\check{g}_i(\mathbf{x}) - g_i(\mathbf{x}))^2 \varphi^2(\mathbf{x}) d\mathbf{x} \;=\; \sum_{i=1}^N \int (\check{g}_i(\mathbf{x}) - g_i(\mathbf{x}))^2 \varphi^2(\mathbf{x}) d\mathbf{x}.$$

Switching to polar coordinates (using A_d to denote the surface area of the unit sphere in $\mathbb{R}^d$), and using the definition of φ from Lemma 6, this equals

$$\sum_{i=1}^N \int_0^\infty A_d r^{d-1} (\check{g}_i(r) - g_i(r))^2 \varphi^2(r) dr \;=\; \sum_{i=1}^N \int_0^\infty A_d \frac{R_d^d}{r} (\check{g}_i(r) - g_i(r))^2 J_{d/2}^2(2\pi R_d r) dr.$$

Using the definition of R_d from Lemma 5, and the fact that $A_d = \frac{d\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$, this equals

$$\sum_{i=1}^N \int_0^\infty \frac{d}{r} (\check{g}_i(r) - g_i(r))^2 J_{d/2}^2(2\pi R_d r) dr.$$

Now, note that by definition of $\check{g}_i, g_i$, their difference $|\check{g}_i(r) - g_i(r)|$ can be non-zero (and at most 1) only for r belonging to two sub-intervals of width $\frac{1}{N}$ within the interval Δ_i (which itself lies in $[\alpha\sqrt{d}, 2\alpha\sqrt{d}]$). Moreover, for such r (which is certainly at least $\alpha\sqrt{d}$), we can use Lemma 22 to upper bound $J_{d/2}^2(2\pi R_d r)$ by $\frac{1.3}{\alpha d}$. Overall, we can upper bound the sum of integrals above by

$$\sum_{i=1}^N \frac{d}{\alpha\sqrt{d}} \cdot \frac{2}{N} \cdot \frac{1.3}{\alpha d} \;<\; \frac{3}{\alpha^2\sqrt{d}}.$$

Appendix C. Technical Results On Bessel functions

Lemma 20 *For any $\nu \geq 0$ and x, $|J_\nu(x)| \leq 1$. Moreover, for any $\nu \geq 1$ and $x \geq 3\nu$, $J_\nu(x)$ is 1-Lipschitz in x.*

Proof The bound on the magnitude follows from equation 10.14.1 in DLMF.

The derivative of $J_\nu(x)$ w.r.t. x is given by $-J_{\nu+1}(x) + (\nu/x)J_\nu(x)$ (see equation 10.6.1 in DLMF). Since $|J_{\nu+1}(x)|$ and $|J_\nu(x)|$, for $\nu \geq 1$, are at most $\frac{1}{\sqrt{2}}$ (see equation 10.14.1 in DLMF), we have that the magnitude of the derivative is at most $\frac{1}{\sqrt{2}} \left|1 + \frac{\nu}{x}\right| \leq \frac{1}{\sqrt{2}} \left(1 + \frac{1}{3}\right) < 1$. ■

To prove the lemmas below, we will need the following explicit approximation result for the Bessel function $J_{d/2}(x)$, which is an immediate corollary of Theorem 5 in Krasikov (2014), plus some straightforward approximations (using the facts that for any $z \in (0, 0.5]$, we have $\sqrt{1-z^2} \geq 1 - 0.3z$ and $0 \leq z \arcsin(z) \leq 0.6z$):

Lemma 21 (Krasikov (2014)) *If $d \geq 2$ and $x \geq d$, then*

$$\left| J_{d/2}(x) - \sqrt{\frac{2}{\pi c_{d,x} x}} \cos\left(-\frac{(d+1)\pi}{4} + f_{d,x} x \right) \right| \;\leq\; x^{-3/2},$$

where

$$c_{d,x} = \sqrt{1 - \frac{d^2-1}{4x^2}} \quad , \quad f_{d,x} = c_{d,x} + \frac{\sqrt{d^2-1}}{2x} \arcsin\left(\frac{\sqrt{d^2-1}}{2x} \right).$$

Moreover, assuming $x \geq d$,

$$1 \geq c_{d,x} \geq 1 - \frac{0.15\, d}{x} \geq 0.85$$

and

$$1.3 \geq 1 + \frac{0.3\, d}{x} \geq f_{d,x} \geq 1 - \frac{0.15\, d}{x} \geq 0.85$$

Lemma 22 *If $d \geq 2$ and $r \geq \sqrt{d}$, then*

$$J_{d/2}^2(2\pi R_d r) \leq \frac{1.3}{r\sqrt{d}}.$$

Proof Using Lemma 21 (which is justified since $r \geq \sqrt{d}$ and $R_d \geq \frac{1}{5}\sqrt{d}$ by Lemma 5), the fact that $\cos$ is at most 1, and the assumption $d \geq 2$,

$$\begin{aligned}
\left| J_{d/2}(2\pi R_d r) \right| &\leq \sqrt{\frac{2}{\pi \cdot 0.85 \cdot 2\pi R_d r}} + (2\pi R_d r)^{-3/2} \\
&= \frac{1}{\sqrt{2\pi R_d r}} \left(\sqrt{\frac{2}{0.85\pi}} + \frac{1}{2\pi R_d r} \right) \\
&\leq \sqrt{\frac{5}{2\pi\sqrt{d} r}} \left(\sqrt{\frac{2}{0.85\pi}} + \frac{5}{2\pi\sqrt{d}\sqrt{d}} \right) \\
&\leq \sqrt{\frac{5}{2\pi\sqrt{d} r}} \left(\sqrt{\frac{2}{0.85\pi}} + \frac{5}{4\pi} \right)
\end{aligned}$$

Overall, we have that

$$J_{d/2}^2(2\pi R_d r) \leq \frac{5}{2\pi r\sqrt{d}} \left(\sqrt{\frac{2}{0.85\pi}} + \frac{5}{4\pi} \right)^2 \;\leq\; \frac{1.3}{r\sqrt{d}}.$$

■

Lemma 23 *For any $\beta \geq 1, d \geq 2$ such that $\beta d \geq 127$, it holds that*

$$\int_{\beta d}^{2\beta d} \frac{J_{d/2}^2(x)}{x} \cdot \mathbf{1}\left\{ J_{d/2}^2(x) \geq \frac{1}{20x} \right\} dx \;\geq\; \frac{0.005}{\beta d}.$$

Proof For any $a, b \geq 0$, we have $a \cdot \mathbf{1}\{a \geq b\} \geq a - b$. Therefore,

$$\begin{aligned}
&\int_{\beta d}^{2\beta d} \frac{1}{x} \cdot J_{d/2}^2(x) \cdot \mathbf{1}\left\{J_{d/2}^2(x) \geq \frac{1}{20x}\right\} dx \\
&\geq \int_{\beta d}^{2\beta d} \frac{1}{x} \cdot \left(J_{d/2}^2(x) - \frac{1}{20x}\right) dx \\
&= \int_{\beta d}^{2\beta d} \frac{1}{x} J_{d/2}^2(x) dx - \frac{1}{20} \int_{\beta d}^{2\beta d} \frac{1}{x^2} dx \\
&= \int_{\beta d}^{2\beta d} \frac{1}{x} J_{d/2}^2(x) dx - \frac{1}{40\beta d}.
\end{aligned}$$

We now wish to use Lemma 21 and plug in the approximation for $J_{d/2}(x)$. To do so, let $a = J_{d/2}(x)$, let b be its approximation from Lemma 21, and let $\epsilon = x^{-3/2}$ the bound on the approximation from the lemma. Therefore, we have $|a - b| \leq \epsilon$. This implies

$$a^2 \geq b^2 - (2|b| + \epsilon)\epsilon, \tag{38}$$

which follows from

$$b^2 - a^2 = (b + a)(b - a) \leq (|b| + |a|)|b - a| \leq (|b| + |b| + \epsilon)\epsilon = (2|b| + \epsilon)\epsilon.$$

Eq. (38) can be further simplified, since by definition of b and Lemma 21,

$$|b| \leq \sqrt{\frac{2}{\pi c_{d,x} x}} \leq \sqrt{\frac{2}{\pi \cdot 0.85 \cdot x}} \leq \frac{1}{\sqrt{x}}.$$

Plugging this back into Eq. (38), plugging in the definition of a, b, and recalling that $c_{d,x} \leq 1$ and $x \geq d \geq 2$, we get that

$$\begin{aligned}
J_{d/2}^2(x) &\geq \frac{2}{\pi c_{d,x} x} \cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x} x\right) - \left(\frac{2}{\sqrt{x}} + x^{-3/2}\right) x^{-3/2} \\
&\geq \frac{2}{\pi x} \cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x} x\right) - 3x^{-2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_{\beta d}^{2\beta d} \frac{1}{x} J_{d/2}^2(x) dx \\
&\geq \frac{2}{\pi} \int_{\beta d}^{2\beta d} \frac{1}{x^2} \cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x} x\right) dx - 3 \int_{\beta d}^{2\beta d} x^{-3} dx \\
&= \frac{2}{\pi} \int_{\beta d}^{2\beta d} \frac{1}{x^2} \cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x} x\right) dx - \frac{9}{8\beta^2 d^2}.
\end{aligned}$$

To compute the integral above, we will perform a variable change, but first lower bound the integral in a more convenient form. A straightforward calculation (manually or using a symbolic computation toolbox) reveals that

$$\frac{\partial}{\partial x}\left(f_{d,x} x\right) = \sqrt{1 - \frac{d^2 - 1}{4x^2}},$$

which according to Lemma 21, equals $c_{d,x}$, which is at most 1. Using this and the fact that $f_{d,x} \geq 0.85$ by the same lemma ,

$$\begin{aligned}
&\int_{\beta d}^{2\beta d} \frac{1}{x^2}\cos^2\left(-\frac{(d+1)\pi}{4}+f_{d,x}x\right)dx \\
&\quad\geq \int_{\beta d}^{2\beta d} \frac{1}{x^2}\cos^2\left(-\frac{(d+1)\pi}{4}+f_{d,x}x\right)\left(\frac{\partial}{\partial x}\left(f_{d,x}x\right)\right)dx \\
&\quad\geq \int_{\beta d}^{2\beta d} \frac{0.85^2}{(f_{d,x}x)^2}\cos^2\left(-\frac{(d+1)\pi}{4}+f_{d,x}x\right)\left(\frac{\partial}{\partial x}\left(f_{d,x}x\right)\right)dx
\end{aligned}$$

Using the variable change $z = f_{d,x}x$, and the fact that $1.3 \geq f_{d,x} \geq 0.85$, the above equals

$$0.85^2\int_{f_{d,\beta d}\beta d}^{f_{d,2\beta d}2\beta d} \frac{1}{z^2}\cos^2\left(-\frac{(d+1)\pi}{4}+z\right)dz \;\geq\; 0.85^2\int_{1.3\beta d}^{1.7\beta d} \frac{1}{z^2}\cos^2\left(-\frac{(d+1)\pi}{4}+z\right)dz$$

We now perform integration by parts. Note that $\cos^2\left(-\frac{(d+1)\pi}{4}+z\right) = \frac{\partial}{\partial z}\left(\frac{z}{2}+\frac{1}{4}\sin\left(-\frac{(d+1)\pi}{2}+2z\right)\right)$, and $\sin$ is always bounded by 1, hence

$$\begin{aligned}
&\int_{1.3\beta d}^{1.7\beta d} \frac{1}{z^2}\cos^2\left(-\frac{(d+1)\pi}{4}+z\right)dz \\
&\quad= \frac{\frac{z}{2}+\frac{1}{4}\sin\left(-\frac{(d+1)\pi}{2}+2z\right)}{z^2}\Big|_{1.3\beta d}^{1.7\beta d} + 2\int_{1.3\beta d}^{1.7\beta d}\frac{\frac{z}{2}+\frac{1}{4}\sin\left(-\frac{(d+1)\pi}{2}+2z\right)}{z^3}dz \\
&\quad\geq \left(\frac{1}{2z}+\frac{\sin\left(-\frac{(d+1)\pi}{2}+2z\right)}{4z^2}\right)\Big|_{1.3\beta d}^{1.7\beta d} + \int_{1.3\beta d}^{1.7\beta d}\left(\frac{1}{z^2}-\frac{1}{2z^3}\right)dz \\
&\quad= \left(\frac{1}{2z}+\frac{\sin\left(-\frac{(d+1)\pi}{2}+2z\right)}{4z^2}\right)\Big|_{1.3\beta d}^{1.7\beta d} + \left(-\frac{1}{z}+\frac{1}{4z^2}\right)\Big|_{1.3\beta d}^{1.7\beta d} \\
&\quad= \left(-\frac{1}{2z}+\frac{1+\sin\left(-\frac{(d+1)\pi}{2}+2z\right)}{4z^2}\right)\Big|_{1.3\beta d}^{1.7\beta d} \\
&\quad= \left(-\frac{1}{2z}\right)\Big|_{1.3\beta d}^{1.7\beta d} + \left(\frac{1+\sin\left(-\frac{(d+1)\pi}{2}+2z\right)}{4z^2}\right)\Big|_{1.3\beta d}^{1.7\beta d} \\
&\quad\geq \left(\frac{0.09}{\beta d}\right) \quad - \quad \frac{1+1}{4(1.3\beta d)^2} \\
&\quad= \frac{1}{\beta d}\left(0.09-\frac{1}{3.38\beta d}\right).
\end{aligned}$$

Concatenating all the lower bounds we attained so far, we showed that

$$\begin{aligned}\int_{\beta d}^{2\beta d} & \frac{1}{x} \cdot J_{d/2}^2(x) \cdot \mathbf{1}\left\{J_{d/2}^2(x) \geq \frac{1}{9x}\right\} \, dx \\ & \geq -\frac{1}{40\beta d} - \frac{9}{8\beta^2 d^2} + \frac{2}{\pi} 0.85^2 \frac{1}{\beta d}\left(0.09 - \frac{1}{3.38\beta d}\right) \\ & \geq \frac{1}{\beta d}\left(0.015 - \frac{1.27}{\beta d}\right).\end{aligned}$$

If $\beta d \geq 127$, this is at least $\frac{0.005}{\beta d}$, from which the lemma follows. ■