Journal of Machine Learning Research 23 (2022) 1-56 Submitted 9/21; Revised 3/22; Published 3/22

Depth separation beyond radial functions

Luca Venturi VENTURI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, USA

Samy Jelassi SJELASSI@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540, USA

Tristan Ozuch OZUCH@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02142, USA

Joan Bruna BRUNA@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences and Center for Data Science
New York University
New York, NY 10011, USA

Editor: Julien Mairal

Abstract

High-dimensional depth separation results for neural networks show that certain functions can be efficiently approximated by two-hidden-layer networks but not by one-hidden-layer ones in high-dimensions. Existing results of this type mainly focus on functions with an underlying radial or one-dimensional structure, which are usually not encountered in practice. The first contribution of this paper is to extend such results to a more general class of functions, namely functions with piece-wise oscillatory structure, by building on the proof strategy of (Eldan and Shamir, 2016). We complement these results by showing that, if the domain radius and the rate of oscillation of the objective function are constant, then approximation by one-hidden-layer networks holds at a poly(d) rate for any fixed error threshold.

The mentioned results show that one-hidden-layer networks fail to approximate high-energy functions whose Fourier representation is spread in the frequency domain, while they succeed at approximating functions having a sparse Fourier representation. However, the choice of the domain represents a source of gaps between these positive and negative approximation results. We conclude the paper focusing on a compact approximation domain, namely the sphere $\mathbb{S}^{d-1}$ in dimension d, where we provide a characterization of both functions which are efficiently approximable by one-hidden-layer networks and of functions which are provably not, in terms of their Fourier expansion.

Keywords: Neural networks, Depth separation

1. Introduction

Learning in high-dimensions is a challenging task for computational, statistical and approximation reasons. Even in the classic supervised learning setup, current empirical successes of deep learning algorithms remain largely out of reach for existing theories, despite phenomenal recent progress. Amongst the algorithmic aspects enabling this success, depth remains a major non-negotiable element. Depth in structured neural networks such as convolutional neural networks provides a multiscale processing of information, but more generally it defines an intricate function class with powerful approximation biases.

Understanding the benefits of depth for approximating certain functions of interest represents a long-standing problem. The classic result of the universal approximation theorem ensures approximation by neural networks of any continuous function, but it focuses on *shallow* (that is, one-hidden-layer) models and does not provide any approximation rates. The seminal work (Barron, 1993) provides quadratic approximation rates by shallow networks under a condition of sparsity of the Fourier transform.

Recent works (Eldan and Shamir, 2016; Daniely, 2017) suggest that this property (sparsity of Fourier transform) is essentially necessary in order to recover polynomial approximation rates, by constructing examples of deep networks which are spread in direction and away from zero in the frequency regime, and by showing that these function can not be efficiently approximated by a shallow counterpart. These depth-separation phenomena occur in the high-dimensional regime, where approximation by neural networks of standard Sobolev spaces is cursed (see e.g. (Maiorov and Meir, 2000)). On the other hand, proofs of such high-dimensional depth-separation phenomena are currently limited to radial functions, that is of the form $f(\mathbf{x}) = \varphi(\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2)$. In this work we extend the results just cited.

We describe rates of approximation by one-hidden-layer networks in terms of the number of units N of the network, by looking at the Fourier representation of the function to be approximated. We consider two types of approximation rate, inspired by the work (Safran et al., 2019): (i) the rate of approximation is polynomial in both the input dimension d and the error estimation ϵ, that is $N \simeq \text{poly}(d, \epsilon^{-1})$ – we refer to this rate of approximation as *universal* approximation (ii) for any fixed error threshold ϵ, the number of units N needed for approximation of approximation depends at most polynomially on d, that is $N \simeq \text{poly}(d)$ for any fixed error threshold ϵ – we refer to this rate of approximation as *fixed-threshold* approximation. We distinguish two fundamentally different regimes of approximation: relative to a heavy-tailed, unbounded data distribution, or relative to a concentrated distribution. Whereas the former captures the most general setup, the latter is motivated by practical machine learning applications. Our contributions are as follows.

- First, we consider a class of two-hidden-layer networks exhibiting piece-wise oscillatory behavior, namely functions of the form

$$f_{r,\mathbf{w},\mathbf{v}} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{2\pi i r\,(\mathbf{v}^T\mathbf{x} + \mathbf{w}^T\mathbf{x}_+)} \ .$$

  In section 3, we show that, under appropriately heavy-tailed data distributions, approximation at a rate $N \simeq \text{poly}(d)$ cannot hold (unconditionally on the weights of the approximant network), as long as the rate of oscillations r grows faster than d. On the other hand, $f_{r,\mathbf{w},\mathbf{v}}$ can be universally approximated (that is, at a rate $\text{poly}(d, \epsilon^{-1})$)

by a two-hidden-layer network with any practical activation of choice. The proof of this result (Theorem 4) extends the main idea introduced by the results of Eldan and Shamir (Eldan and Shamir, 2016) beyond the radial case.

- In section 4, we show that the poly(d)-oscillatory aspect and the heavy-tailed data distributions are necessary in the depth-separation result mentioned above. More specifically, we show that any deep network, with $O(1)$-bounded weights and $O(1)$-Lipschitz activation, can be fixed-threshold approximated by one-hidden-neural networks over a compact set of radius $O(1)$ (Theorem 11). This extends an equivalent result in (Safran et al., 2019), from the class of radial functions to the one of deep neural networks with Hölder activations.

- Aforementioned depth separation results consider functions whose Fourier representation is spread in high frequencies. On the other hand, universal approximation results often require the function to be approximated to be, in some sense, sparse in the Fourier domain. Unfortunately, there are currently many gaps between these two types of results, one of them being the definition of approximation domain. In order to reduce the gap between the two results above, we consider approximation on a fixed compact domain, namely the unit sphere $\mathbb{S}^{d-1}$, where Fourier analysis can be done using spherical harmonics. We individuate two conditions on the spherical harmonics decomposition of a function $f \in C(\mathbb{S}^{d-1})$. The first is a sparsity condition on the decomposition, which we show to be sufficient to prove universal approximation (that is, at a rate $N \simeq \text{poly}(d, \epsilon^{-1})$) of f by one-hidden-layer networks. The second is a high-energy spreadness condition on the spherical harmonics decomposition of f, which we show to imply that universal approximation of f by one-hidden-layer networks cannot hold. This is the content of section 5, of which the main results are summarized in section 5.2.

1.1 Related works

There is a huge literature of approximation results for neural networks. Early approximation results provided upper and lower bounds on the approximation of some functional spaces such as Sobolev spaces (Maiorov and Meir, 2000) or L^p spaces (Pinkus, 1999) by neural networks. For high input dimensions d, such results hold for functions with smoothness proportional to d, or require an approximation rate that scales as $N \sim \epsilon^{-d}$ (see e.g. (Petersen, 2020; Gühring et al., 2020) for a review), where N denotes the number of units of the network and ϵ the error threshold.

In more recent years, quite a few works pointed out the benefits of deep networks versus their shallow counterparts from the point of view of approximation rates. For example, this has been shown for sawtooth function (Telgarsky, 2016), functions with positive curvature (Liang and Srikant, 2016; Yarotsky, 2017; Safran and Shamir, 2017), functions with a compositional structure (Poggio et al., 2017), piecewise smooth functions (Petersen and Voigtlaender, 2018), Gaussian mixture models (Jalali et al., 2019), polynomials (Rolnick and Tegmark, 2017), or model reduction models (Rim et al., 2020). The result of (Telgarsky, 2016) has been further generalized using a notion of periodicity (Chatziafratis et al., 2019). It must be noticed that most of the cited works show depth separation that is inde-

pendent of the dimension d and that increases exponentially with the depth of the network. Another line of works (Eldan and Shamir, 2016; Daniely, 2017; Safran et al., 2019) on the other hand shows depth separation exponential in the dimension d, between networks with one and two hidden layers. This is the framework of this work. It was also shown recently that depth separation results between fixed depths greater than this are arguably difficult to prove (Vardi and Shamir, 2020; Vardi et al., 2021).

This depth-width trade-off has been analyzed through different lens than approximation capabilities, such as classification capabilities (Malach and Shalev-Shwartz, 2019), exact representability (Arora et al., 2016), Betti numbers (Bianchini and Scarselli, 2014), number of linear regions (Pascanu et al., 2013; Montufar et al., 2014; Raghu et al., 2017; Hanin and Rolnick, 2019a,b), trajectory lengths (Raghu et al., 2017), globale curvature (Poole et al., 2016) or topological entropy (Bu et al., 2020). In essence, all these results state that networks expressivity improve exponentially as we increase the depth. Another related question is whether depth-separation holds from a learnability (therefore, not solely approximation) point of view as well (Malach and Shalev-Shwartz, 2019; Malach et al., 2021). In this work we focus on approximation and we consider the Fourier representation as a complexity measure. This is the approach followed by e.g. (Eldan and Shamir, 2016; Daniely, 2017), which construct examples of deep neural networks, whose Fourier energy is exponentially higher than those of shallow neural networks with a moderate number of units.

On the other hand, sparsity of the Fourier transform has been used to show polynomial rates of approximation of functions by neural networks (Klusowski and Barron, 2018; Ongie et al., 2019; Bresler and Nagaraj, 2020). In the last part of the paper, we show that an equivalent condition can be described in terms of spherical harmonics decomposition.

2. Preliminaries

2.1 Neural networks

For $L \geq 1$, we call an L-hidden-layer feed-forward neural network a function

$$f : \mathbf{x} \in \mathbb{R}^d \to \mathbf{x}^{(L+1)}(\mathbf{x}) \in \mathbb{C}^{d_{L+1}} \ , \tag{1}$$

where $\mathbf{x}^{(L)}$ is defined by recursion by $\mathbf{x}^{(0)}(\mathbf{x}) = \mathbf{x}$,

$$\mathbf{x}^{(k)}(\mathbf{x}) = \sigma^{(k)}(\mathbf{A}^{(k)}\mathbf{x}^{(k-1)}(\mathbf{x})) \text{ for } k \in [L] \quad \text{and} \quad \mathbf{x}^{(L+1)}(\mathbf{x}) = \mathbf{A}^{(L+1)}\mathbf{x}^{(L)}(\mathbf{x}) \ ,$$

where

$$\begin{aligned} \mathbf{A}^{(k)} &= [\mathbf{a}_1^{(k)}| \cdots |\mathbf{a}_{d_k}^{(k)}]^T \in \mathbb{R}^{d_k \times d_{k-1}} \quad \text{for } k \in [L], \\ \mathbf{A}^{(L+1)} &= [\mathbf{a}_1^{(L+1)}| \cdots |\mathbf{a}_{d_{L+1}}^{(L+1)}]^T \in \mathbb{C}^{d_{L+1} \times d_L} \end{aligned}$$

(with $d_0 = d$) and $\sigma^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ are *activation* functions, that is $\left(\sigma^{(k)}(\mathbf{x})\right)_i = \sigma_i^{(k)}(x_i)$ for some function $\sigma_i^{(k)} : \mathbb{R} \to \mathbb{R}$. A neural network is therefore a sequence of sums and compositions of *ridge* functions, that is functions of the form $\mathbf{x} \mapsto \sigma(\mathbf{w}^T\mathbf{x})$. In the following, unless specified, we only consider neural networks (or, more simply, networks) as defined in

(1). Most of the times we will deal with real-valued networks, that is $\mathbf{A}^{(L+1)} \in \mathbb{R}^{d_{L+1} \times d_L}$. We say that a network has activation σ if $\sigma_i^{(k)}(x) = \sigma(x + b_i^k)$ for some bias term $b_i^k \in \mathbb{R}$ for all k, i. We refer to the function

$$\mathbf{x} \in \mathbb{R}^{d_{k-1}} \mapsto \sigma^{(k)}(\mathbf{A}^{(k)}\mathbf{x}) \in \mathbb{R}^{d_k}$$

as k-th *hidden (or inner) layer* of *width* d_k, for $k \in [L]$, while we refer to the linear function defined by $\mathbf{A}^{(L+1)}$ as the last (or $L+1$-th) layer. We refer to the value $W(f) \doteq \max_{k \in [L]} d_k$ as width of the network f and to the vectors $\mathbf{a}_i^{(k)}$ as *weights* (of the k-th layer), for all k, i. A basic complexity measure for neural network (1) is given by the total number of units, or *size*:

$$N(f) \doteq \sum_{k=1}^{L} d_k \ .$$

The number of layers $L(f) = L$ is also a relevant measure of complexity, which we refer to as *depth*. Finally, in the following we sometimes require a control on the value of the weights; such controls are expressed in terms of norm p of the weights, that is

$$m_p(f) \doteq \max_{k,i} \|\mathbf{a}_{k,i}\|_p \ ,$$

for some $p \in [1, \infty]$.

2.2 Neural network approximation rates

We measure the approximation error between two functions $f, g : \Omega \subseteq \mathbb{R}^d \to \mathbb{C}$ in terms of the $L^2(\mu)$ norm (with respect to a probability measure or density μ)

$$\|f - g\|_{\mu,2}^2 \doteq \int_\Omega |f(\mathbf{x}) - g(\mathbf{x})|^2 \, d\mu(\mathbf{x}) \ ,$$

or L^∞ norm

$$\|f - g\|_{\Omega,\infty} \doteq \sup_{\mathbf{x} \in \Omega} |f(\mathbf{x}) - g(\mathbf{x})| \ .$$

Notice that a (uniform) L^2 lower bound implies a L^∞ one, and viceversa for an upper bound. The focus of this chapter is to establish upper and lower bounds for approximation of certain function classes by shallow neural networks, in high dimensions d. We distinguish two different approximation regimes of interest. In the following we will denote by $\mathcal{F}_N^\sigma$ the space of one-hidden-layer neural networks $f_N : \mathbb{R}^d \to \mathbb{R}$ with width at most N and activation σ (where the dimension d is inferred from the context); similarly, we will denote by $\mathcal{F}_N$ the space of one-hidden-layer neural networks with width at most N and any (continuous) activation.

Definition 1 *We say that a sequence* $\left\{ f^{(d)} : \Omega_d \subseteq \mathbb{R}^d \to \mathbb{C} \right\}_{d \geq 2}$ *is* universally approximable *by one-hidden-layer networks with activation* σ *if it is approximable at a* $\mathrm{poly}(d, \epsilon^{-1})$ *rate; that is if there exists some constants* $\alpha > 0$ *and* $\beta > 0$ *such that it holds*

$$\left\| f^{(d)} - f_N \right\|_{\Omega_d,\infty} \leq \epsilon$$

for some one-hidden-layer $f_N \in \mathcal{F}_N^\sigma$ *satisfying* $N + m_\infty(f_N) \leq \alpha(d\epsilon^{-1})^\beta$.

Definition 2 *We say that* $\left\{f^{(d)}\right\}_d$ *is* fixed-threshold approximable *if for any* $\epsilon \in (0,1)$ *it is* ϵ*-approximable at a* poly(d) *rate; that is if for any* $\epsilon > 0$ *there exists some constants* $\alpha > 0$ *and* $\beta > 0$ *such that it holds*

$$\left\| f^{(d)} - f_N \right\|_{\Omega_d,\infty} \leq \epsilon$$

for some one-hidden-layer $f_N \in \mathcal{F}_N^\sigma$ *satisfying* $N + m_\infty(f_N) \leq \alpha d^\beta$.

These approximation schemes were introduced in (Safran et al., 2019). To ensure significance of the approximation rates, in the following upper and lower bounds are stated for objective functions $f^{(d)}$ normalized such that $\|f^{(d)}\|_2 \leq 1$ or $\|f^{(d)}\|_\infty \leq 1$.

Notice that some of the inapproximability results shown below are for target networks with complex values. Although, one can obtain equivalent results with a real-valued target network by simply taking the real (or imaginary) part of such complex-valued target networks.

2.3 Activation assumptions

Finally, the results in the next sections generally hold for activations satisfying the following assumptions, which are satisfied by common activation such as the ReLU $\mathrm{ReLU}(x) = x_+$ or the sigmoid $\mathrm{sigmoid}(x) = (1 + e^{-x})^{-1}$ (Eldan and Shamir, 2016). Most of the results can be easily generalized to hold under less strict conditions, but we take these assumptions for sake of simplicity.

Assumption 1 *Given an activation* $\sigma : \mathbb{R} \to \mathbb{R}$*, there exist constants* ι_σ *and* ν_σ *such that*

1. *it is* ι_σ*-Lipschitz and* $\sigma(0) \leq \iota_\sigma$*;*
2. *for any L-Lipschitz function* $f : \mathbb{R} \to \mathbb{R}$ *constant outside of an interval* $[-R, R]$ *and any* $\epsilon > 0$ *there exits* $f_N \in \mathcal{F}_N^\sigma$ *with* $\|f - f_N\|_\infty \leq \epsilon$ *such that* $N + w_\infty(f_N) \leq \nu_\sigma L R \epsilon^{-1}$.

Notice that this assumption implies that, given a (deep) neural network f with poly(d) weights and activations satisfying Assumption 1, then we are always able to replace the activations in f by any other activation satisfying Assumption 1, by paying an at most polynomial cost. This is formalized in the following lemma.

Lemma 3 *Let* $\left\{f^{(d)} : K_d \subset \mathbb{R}^d \to \mathbb{C}\right\}_d$ *be neural networks with activations satisfying Assumption 1 and such that* $N(f^{(d)}) + w_\infty(f^{(d)}) + \mathrm{diam}(K^{(d)}) \leq \mathrm{poly}(d)$*; also let* σ *be any activation function satisfying Assumption 1. Then the sequence* $\left\{f^{(d)}\right\}_d$ *is universally approximable by networks (of the same depth as* $f^{(d)}$*) with activation* σ.

2.4 Notation

We introduce notation we use throughout the rest of the paper. We denote scalar valued variables as lowercase non-bold; vector valued variables as lowercase bold; matrix and tensor valued variables and multivariate random variables (r.v.'s) as uppercase bold. Given a vector $\mathbf{v} \in \mathbb{R}^d$, we denote its components as v_k; given a matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$, we denote its columns as $\mathbf{w}_k$. For a matrix $\mathbf{W}$, we denote by $\|\mathbf{W}\|_{F,p}$ its entrywise p-norm, by $\|\mathbf{W}\|_{p,q}$ its (p,q) operator norm (that is $\|\mathbf{W}\|_{p,q} = \max_{\|\mathbf{y}\|_p=1} \|\mathbf{W}\mathbf{y}\|_q$) and by $\|\mathbf{W}\|_p$ its p operator

norm (that is $\|\mathbf{W}\|_p = \|\mathbf{W}\|_{p,p}$). We denote by $\mathbb{S}^{d-1} \subset \mathbb{R}^n$ the $(d-1)$-dimensional sphere $\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\right\}$ and by $B^d_{r,p}$ the ℓ^p ball of radius r in $\mathbb{R}^d$, that is $\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_p \leq r\right\}$. We denote by $L^p(\Omega)$, $L^p(\mu)$, $L^p(\varphi)$ the spaces of functions $f : \Omega \to \mathbb{R}$ which are p-integrable with respect to the Lebesgue measure, the measure μ or the density φ, respectively. The respective norms (and scalar products for $p = 2$) are denoted by $\|f\|_{\zeta,p}$ ($\langle f, g\rangle_\zeta$) for $\zeta \in \{\Omega, \mu, \varphi\}$; we simply write $\|f\|_p$ when the measure is clear from the context. For a finite signed Borel measure μ, we denote its total variation as $\|\mu\|_1$. Finally, we denote by $\hat{f}$ or $\mathscr{F}(f)$ (resp. $\check{f}$ or $\mathscr{F}^*(f)$) the Fourier transform (resp. the inverse Fourier transform) of f (meant in the following in the sense of tempered distributions).

3. A depth separation example

Our starting point for the study of depth-separation is to consider a generic data distribution μ with adversarial properties against shallow approximations. In the seminal work (Eldan and Shamir, 2016), Eldan and Shamir establish an unconditional (with no restrictions on the norms of the weights of the network) depth-separation result by considering a density μ in $\mathbb{R}^d$ with tails $\mu(\|\mathbf{x}\|_2) \simeq \|\mathbf{x}\|_2^{-(d+1)/2}$ and a radial function $f^{(d)}(\mathbf{x}) = h_d(\|\mathbf{x}\|_2)$ with $h_d : \mathbb{R} \to \mathbb{R}$ a carefully chosen oscillating function with compact support. The proof in (Eldan and Shamir, 2016) reveals the limitations of shallow neural networks at approximating high-dimensional functions via a powerful harmonic analysis insight, that is particularly convenient in the setting of radial functions. In this section, we show that their proof strategy can be extended to include more diverse function classes, namely those arising naturally from ReLU networks. Specifically, we consider networks of the form

$$f_{r,\mathbf{w},\mathbf{v}} : \mathbf{x} \in \mathbb{R}^d \mapsto \sigma_r\left(\mathbf{v}^T\mathbf{x} + \mathbf{w}^T\mathbf{x}_+\right) \tag{2}$$

where $\mathbf{x}_+$ denotes the element-wise ReLU activation, $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^d$ and $\sigma_r(t) = e^{2\pi i r t}$. We are thus considering a function which is piece-wise oscillatory, with constant envelope $|f_{r,\mathbf{w},\mathbf{v}}(\mathbf{x})| = 1$, and where the frequency of oscillations is controlled by r. The main result of this section can be summarized as follows.

Theorem 4 (Informal) *Assume that $\|\mathbf{w}\|_2 = \Theta(1)$, $\|\mathbf{v}\|_2 = O(1)$ and that $r = \Theta(d^k)$ for some $k \geq 2$. Then there exists a (low-decay) product measure μ on $\mathbb{R}^d$ such that the function $f_{r,\mathbf{w},\mathbf{v}}$ is universally approximable by two-hidden-layer networks but it is not fixed-threshold approximable by one-hidden-layer networks.*

3.1 The lower bound

Let $\psi \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ with $\|\psi\|_2 = 1$, and such that its Fourier transform $\hat{\psi}$ is compactly supported in $[-K, K]$, for some $K > 0$. Assume also that

$$\|\psi\|_1 < \sqrt{2/K}\ . \tag{3}$$

The condition ensure that the density ψ is sufficiently spread away from zero (see Remark 8). Our first objective is to establish depth separation for the approximation of $f_{r,\mathbf{w},\mathbf{v}}$ under the L^2 metric defined by the probability density φ^2, where $\varphi : \mathbf{x} \in \mathbb{R}^d \mapsto \prod_{j=1}^d \psi(x_j)$.

Theorem 5 *Let $f^{(d)} = f_{r_d,\mathbf{w}_d,\mathbf{v}_d}$, for some $r_d \in \mathbb{R}$, $\mathbf{w}_d, \mathbf{v}_d \in \mathbb{R}^d$. For a fixed $\gamma > 0$, define*

$$\tau_d \doteq \sup_{S\subseteq[d]} \|\mathbf{v}_d + \mathbf{w}_{d,S}\|_\infty , \quad \Omega_d \doteq \{j \in [d] : r_d|w_{d,j}| \geq \gamma d^2\} \quad \textit{and} \quad \eta_d \doteq \frac{|\Omega_d|}{d} ,$$

where $\mathbf{w}_{d,S} \in \mathbb{R}^d$ is defined by $w_{d,S,i} = w_i \mathbb{1}\{i \in S\}$. Assume that

(i) oscillations grow polynomially, that is $\tau_d \cdot r_d = \Theta(d^k)$ for some constant $k > 0$;

(ii) the vectors $\mathbf{w}_d$ are sufficiently spread, that is $\eta_d \geq \eta$ for some $\eta > 0$ independent of d;

(iii) the density φ^2 is sufficiently spread, i.e. $2K\|\psi\|_1^2 < 2^{2\eta}$.

Then there exists a constant $\alpha \in (0, 1)$ (independent of d) such that for all d it holds

$$\inf_{f_N\in\mathcal{F}_N} \|f^{(d)} - f_N\|^2_{\varphi^2,2} \geq 1 - N \cdot \alpha^d \cdot O(d^{k+1}) . \tag{4}$$

Notice that this lower bound is unconditional on the weights of the neurons $m_\infty(f_N)$.

The proof follows a similar strategy as in the work (Eldan and Shamir, 2016). The approximation error can be expressed in the Fourier domain as

$$\|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|^2_{\varphi^2,2} = \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} \cdot \varphi - f_N \cdot \varphi\|_2^2 = \|\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi} - \hat{f_N} * \hat{\varphi}\|_2^2 .$$

Thanks to the assumptions, the target function $f_{r_d,\mathbf{w}_d,\mathbf{v}_d}$ satisfies a key property, namely that its Fourier transform has its energy sufficiently spread in the high-frequencies, after the convolution by $\hat{\varphi}$. Such frequency spread is caused by the shattering of the first ReLU layer, which effectively creates $\Theta(2^{\eta d})$ different frequencies. The piece-wise structure arising from the ReLU can be handled in the Fourier domain by the Hilbert transform of the function ψ, which has sufficient decay thanks to the assumptions. Noticing that $\|\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi}\|_2 = 1$, this is formalized in the following.

Lemma 6 (Informal) *It holds that*

$$\left|\left(\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi}\right)(\boldsymbol{\xi})\right| \lesssim 2^{-\eta d}\|\varphi\|_1\|\boldsymbol{\xi}\|_\infty^{-1} \qquad \textit{for } \|\boldsymbol{\xi}\|_\infty \gtrsim \mathrm{poly}(d) .$$

On the other hand, since $\hat{\varphi}$ is compactly supported and the Fourier transform of a single-unit network is localised in a frequency ray, the Fourier transform of $f_{r_d,\mathbf{w}_d,\mathbf{v}_d} \cdot \varphi$ is localised in a union of N tubes, of the form $T_{\boldsymbol{\alpha}} = \mathrm{span}(\{\boldsymbol{\alpha}\}) + [-K, K]^d$. This implies that

$$\inf_{f_N\in\mathcal{F}_N} \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|^2_{\varphi^2,2} \geq \inf_{f_N\in\mathcal{T}_{(N)}} \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|^2_{\varphi^2,2}$$

where $\mathcal{T}_{(N)}$ denotes the set of L^2 functions such that their Fourier transform is supported on the union of N tubes $T_{\boldsymbol{\alpha}_1}, \dots, T_{\boldsymbol{\alpha}_N}$ as above, for some arbitrary $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N \in \mathbb{R}^d$. Thanks to Plancherel's identity, and since $\|f_{r_d,\mathbf{w}_d,\mathbf{v}_d}\|_{\varphi^2,2} = 1$, it further holds that

$$\inf_{f_N\in\mathcal{T}_{(N)}} \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|^2_{\varphi^2,2} \geq 1 - N \cdot \sup_{\boldsymbol{\alpha}\in\mathbb{S}^{d-1}} \left\| \mathbb{1}_{T_{\boldsymbol{\alpha}}} \cdot \left(\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi}\right)\right\|_2^2 ,$$

where $\mathbb{1}_{T_{\boldsymbol{\alpha}}}$ denotes the indicator function of $T_{\boldsymbol{\alpha}}$. Lemma 6 can then be used to show that such projections are exponentially (in d) small, which implies equation (11). The detailed proof is deferred to section A.1.

Remark 7 *Theorem 5 asks for two main conditions to hold. First, the magnitude of oscillations of the objective function (parametrised by r_d) must grow at least polynomially with d, similarly to the assumptions in the works (Eldan and Shamir, 2016) and (Daniely, 2017). Second, the data distribution μ with density φ^2 should be heavy-tailed, in order for its Fourier transform to be sufficiently localised. When r_d does not grow fast enough with d, the energy starts piling up at the low frequencies, creating an important roadblock to establish approximation lower-bounds, and leaving open the possibility of efficient shallow approximation. Similarly, when μ concentrates too quickly, the proof strategy also fails, due to the fact that in that case $\hat{\varphi}$ is too spread in the Fourier domain, creating full overlap of the energies.*

Remark 8 *The admissibility condition* (3) *is necessary since $\eta \leq 1$ by definition. Notice that*

$$1 = \|\psi\|_2^2 = \|\hat{\psi}\|_2^2 \leq (2K)\|\hat{\psi}\|_\infty^2 \leq (2K)\|\psi\|_1^2$$

and therefore condition (3) *can be considered as a requirement on ψ not being too concentrated in the origin. The choice $\psi(t) = \sqrt{3/2}\, sinc^2(\pi t)$ corresponds to $K = 1$, $\|\psi\|_1 = \sqrt{3/2}$ and $\|\psi\|_2 = 1$, which verifies (3). In that case, from condition (ii) we need $\eta > \frac{\log_2 3}{2} \approx 0.79$. However, the choice $\psi(t) = Csinc(\pi t)$ (the equivalent separable version of the of density considered in (Eldan and Shamir, 2016)) is not admissible, since ψ is not in L^1. The lower bound may be optimized by finding compactly supported windows with an optimal L^1 to L^2 ratio of their Fourier transforms.*

Remark 9 *The theorem considers a separable ReLU transform $\mathbf{x} \mapsto \mathbf{x}_+$, combined with a separable data distribution μ with density φ^2. One could expect a similar lower bound to apply in the more general case of a layer of the form $\mathbf{x} \mapsto (\mathbf{U}\mathbf{x} + \mathbf{b})_+$, $\mathbf{U} \in \mathbb{R}^{d' \times d}, \mathbf{b} \in \mathbb{R}^{d'}$. Such general case replaces the Hilbert transform of ψ with the Fourier transform of indicators of convex polytopes, which has been used in the context of ReLU networks to characterize spectral properties (Rahaman et al., 2019).*

Example 1 *We give an explicit example of a family of function $\left\{f^{(d)} : \mathbb{R}^d \to \mathbb{R}\right\}$ which satisfy the assumptions of Theorem 5. Consider the functions*

$$f^{(d)}(\mathbf{x}) = \exp\left(2\pi i d^2 \sum_{k=1}^{d} \max\{0, x_k\}\right).$$

Then, if μ_d is the product probability measure defined by the density in Remark 8, that is

$$\mu_d(d\mathbf{x}) = \prod_{k=1}^{d} \left[\frac{3}{2}\mathrm{sinc}^4(\pi x_k)\, dx_k\right],$$

then it holds that

$$\inf_{f_N \in \mathcal{F}_N} \left\| f_N(\mathbf{x}) - f^{(d)}(\mathbf{x}) \right\|_{\mu_d,2}^2 \geq 1 - 1300N \cdot d^2 \cdot (0.75)^d .$$

For example, this implies that

$$\inf_{f_N \in \mathcal{F}_N} \left\| f_N(\mathbf{x}) - f^{(d)}(\mathbf{x}) \right\|_{\mu_d,2} \geq \frac{1}{2}$$

unless

$$N \geq \frac{1.3^d}{10^4 d^3} \, .$$

The numbers are obtained by explicitly tracking the constant in the proof of Theorem 5 (see section A.1 for more details). Finally, notice that the functions $f^{(d)}$ are not radial. Indeed, they show different behaviour over each orthant, thanks to the ReLU layer. On the other hand, radial functions would behave equally over any orthant. A similar reasoning would generally hold for radiality in certain directions of the input.

3.2 The upper bound

According to the definition of neural networks we gave in section 2.1, the function $f_{r,\mathbf{w},\mathbf{v}}$ is naturally a two-hidden-layer neural network. Although, while there are cases of sinusoidal activations being used in practice, activations such as ReLU or sigmoid are more relevant to practical applications. The following theorem, proved in section A.2, shows that we can efficiently represent the function $f_{r,\mathbf{w},\mathbf{v}}$ in the hypothesis of the Theorem 5 as a two-hidden-layer neural network with fixed activation, such as the ReLU or the sigmoid. The main technical difference with Lemma 3 is that the result is proved for approximation w.r.t. the probability measure with density φ^2 introduced above.

Theorem 10 *Let σ be an activation satisfying Assumption 1. Assume that there exists a constant $k \geq 1$ such that $m_\infty(f_{r_d,\mathbf{v}_d,\mathbf{w}_d}) \leq O(d^k)$ and assume that ψ is such that $|\psi(x)| = O(|x|^{-1})$. Then, for every $\epsilon > 0$, there exists $f_N \in \mathcal{F}_N^\sigma$ with*

$$N + m_\infty(f_N) \leq O\Big(d^{2(1+k)}\epsilon^{-3/2}\Big) \quad \textit{such that} \quad \|f_N - f_{r_d,\mathbf{w}_d,\mathbf{v}_d}\|^2_{\varphi^2,2} \leq \epsilon \, .$$

Theorems 5 and 10 therefore estabilish a depth separation result. If $f^{(d)} = f_{r_d,\mathbf{w}_d,\mathbf{v}_d}$ are defined with $r_d, \mathbf{w}_d, \mathbf{v}_d$ satisfying the assumptions of both theorems (that is, they satisfy assumptions (i)-(ii)-(iii) of Theorem 5 with $\tau_d \cdot r_d = \Theta(d^k)$), then Theorem 5 says that $\left\{f^{(d)}\right\}_d$ is not fixed-threshold approximable by one-hidden-layer networks, while Theorem 10 says that the sequence is universally approximable by two-hidden-layer networks with a fixed activation satisfying Assumption 1. For example, the family of functions considered in Example 1 satisfies such assumptions.

We thus identify two key aspects responsible for such depth separation: heavy-tailed data and oscillations growing with dimension. In the next sections we want to understand how necessary these two conditions are. The next section shows that if these two condition do not hold anymore, then a lower bound such as the one in Theorem 5 is not achievable; more specifically we show that the objective function is fixed-threshold approximable by one-hidden-layer networks.

4. Approximation of deep networks by shallow ones

In this section, we show that any deep neural network f (which include the target functions considered in the previous section) can be approximated by shallow ones at a rate which is polynomial in d, as long as the rate of oscillation in the inner layers of f is constant in d and the metric is concentrated in a ball of constant radius. We start by reporting the

result in a general form for two-hidden-layer networks and we discuss some consequences and extensions afterwards.

Consider a family of two-hidden-layers neural network $\{f^{(d)} : K_d \subset \mathbb{R}^d \to \mathbb{C}\}$ of the form

$$f^{(d)} : \mathbf{x} \in \mathbb{R}^d \mapsto \boldsymbol{\gamma}_d^T \mathbf{g}\big(\mathbf{W}_d^T \mathbf{h}\big(\mathbf{U}_d^T \mathbf{x}\big)\big) \in \mathbb{C} \ , \tag{5}$$

where $\mathbf{h} = \mathbf{h}^{(d)} : \mathbb{R}^{p_d} \to \mathbb{R}^{p_d}$ and $\mathbf{g} = \mathbf{g}^{(d)} : \mathbb{R}^{o_d} \to \mathbb{R}^{o_d}$ are, respectively, component-wise 1-Lipschitz and $(1, \alpha)$-Holder[1] activation functions, and $\mathbf{U}_d \in \mathbb{R}^{d \times p_d}$, $\mathbf{W}_d \in \mathbb{R}^{p_d \times o_d}$, $\boldsymbol{\gamma}_d \in \mathbb{C}^{o_d}$. We wish to approximate $f^{(d)}$ by one-hidden-layer neural networks with a given activation.

Theorem 11 *Assume that* $\mathrm{diam}(K_d) = O(1)$ *and that the networks* $f^{(d)}$ *have* ℓ^1 *bounded weights, that is* $m_1(f^{(d)}) = O(1)$. *Then, for every activation* σ *satisfying Assumption 1.2 and every* $\epsilon \in (0,1)$ *it holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f^{(d)} - f_N^\sigma\|_{K,\infty} \leq \epsilon \quad \text{for some } N \leq \exp\Big(O\Big(\epsilon^{-1-2/\alpha} \log(p_d/\epsilon)\Big)\Big) \ .$$

Moreover, f_N^σ *can be chosen such that* $m_\infty(f_N^\sigma) \leq (1+N^2) = \exp\big(O\big(\epsilon^{-1-2/\alpha} \log(p_d/\epsilon)\big)\big)$.

The proof is constructive and based on the following observation. Consider the case where $o_d = 1$, $\gamma_d = 1$, $p_d = p$ and $g(x) = x^r$ some positive integer r. If $h_k(x) = e^{ix}$ for all $k \in [p]$, then the function $f = f^{(d)}$ at (5) has form

$$f(\mathbf{x}) = \left(\sum_{k=1}^N w_k e^{i\mathbf{u}_k^T \mathbf{x}}\right)^r$$

for some $w \in \mathbb{R}^N$, $\mathbf{u}_k \in \mathbb{R}^d$, where $N = p$. By expanding the power we can write

$$f(\mathbf{x}) = \sum_{j_1 + \cdots + j_N = r} \binom{r}{j_1 \cdots j_N} \Big(w_1^{j_1} \cdots w_N^{j_N}\Big) e^{i\left(\sum_{h=1}^N j_h \mathbf{w}_h\right)^T \mathbf{x}} \ ,$$

that is a formulation of f as a one-hidden-layer network with activation $\sigma_1(t) = e^{2\pi i t}$ (in the following we refer to this type of networks as *shallow Fourier networks*) and a number of units that scales as N^r. Since both polynomials and trigonometric polynomials are universal approximators, with well known convergence rates, in the general case one can proceed as follows. Each of the non-linearities applied to the first hidden layer can be approximated by a trigonometric polynomial at a polynomial rate on the interval of interest. Similarly, every non-linearity applied to the second hidden layer can be approximated by a polynomial at a linear (in the degree of the polynomial) rate on the interval of interest. Assuming for simplicity that both rates behave as ϵ^{-1}, where $\epsilon > 0$ denotes the approximation error, the composition of the two approximation following the structure of the target network results in a shallow Fourier network (that is with activation $\sigma_1(t) = e^{2\pi i t}$) whose size N behaves, roughly speaking, as

$$N \simeq \Theta\big(p\epsilon^{-2}\big)^{\epsilon^{-1}} \ .$$

1. We say that a function $g : \mathbb{R} \to \mathbb{R}$ is $(1, \alpha)$-Holder if it holds that $|g(x) - g(y)| \leq |x - y|^\alpha$ for all $x, y \in \mathbb{R}$.

Moreover, it is also possible to control the value of the coefficients appearing in the final approximation. With this, we can approximate each summand in the shallow Fourier network by a one-hidden-layer network with activation σ with a controlled number of units, thanks to Assumption 1.2. A more detailed statement and a formal proof are reported in appendix B.

In essence, in the Theorem 11, we show that it is possible to fixed-threshold approximate a two-hidden-layer neural network with constant(d) oscillations at a poly(d) rate over a compact set of constant(d) radius. On the other hand, it easy to show that it is also possible to obtain approximation at a poly(ϵ^{-1}) rate (see section B.7), for fixed d. Finally, existing results in the literature (see (Safran et al., 2019)) show that universal approximation is not possible, the counterexample being essentially a radial function.

Interestingly, the upper bound in Theorem 11 does not depend on the number of units in the second layer of the objective function. This parameter is *hidden* in the control we impose on the ℓ^1 norm of the objective weights. The proof technique of this upper bound highlights how the difficulty of approximating at poly(d, ϵ^{-1}) rate stems from the high-energy of the second layer, which requires the shallow network used for approximation to have a (potentially) exponential (in d) number of directions. Notice that the lower bound in Theorem 5 actually tells that the function is not fixed-threshold approximable. High oscillations in the lower bound (4) essentially ensure that an exponential (in d) number of neurons are necessary. An open question is then whether a low-decaying measure is, in general, necessary for such a result to hold.

Expanding on the proof technique above, it is possible to extend the result of Theorem 11 to approximation of L-hidden-layers networks by shallow ones, which gives a rate scaling as $\exp(O(\epsilon^{-L}\log(p/\epsilon)))$.

Theorem 12 *Let $f^{(d)}$ as in (1), with $O(1)$-Lipschitz activations, first hidden layer width $d_1 = p_d$, depth $L_d = L$ and bounded weights, that is $m_1(f^{(d)}) = O(1)$. Then for every $\epsilon > 0$ there exists a shallow Fourier network $f_N \in \mathcal{F}_N^\sigma$ with*

$$N \leq \left(p_d \cdot O\left(1 + \frac{1}{\epsilon^2}\right)\right)^{O(L)\left(1+\frac{1}{\epsilon}\right)^{L-1}} \quad \textit{such that} \quad \left\|f^{(d)} - f_N\right\|_{B^d_{1,\infty},\infty} \leq \epsilon \ .$$

See section B.6 for a formal statement and its proof. While it has been shown that generic $O(1)$-Lipschitz function can not be (computably) represented by neural networks with $N \simeq \text{poly}(d)$ units (Vardi et al., 2021), an interesting related follow-up conjecture is whether our result can be generalized to any generic $O(1)$-Lipschitz function which is poly(d)-computable. Notice that this is dependent on the choice of the uniform norm to measure the approximation error. For example, it has been shown that a rate $N \simeq \text{poly}(d)$ is achievable for approximation in the L^2 norm with the uniform measure (Hsu et al., 2021).

Finally, notice that the approximation rate shown in Theorem 11 and Theorem 12 are actually polynomial in the size p_d of the first hidden layer of $f^{(d)}$ rather than in the input dimension d. Although, up to choosing a worse (yet constant) exponent in ϵ, we can replace p_d by d in the statement, by considering the function as a $(L+1)$-hidden-layer network, where the first layer is the identity.

4.1 Two cases of interest

Theorem 11 allows to recover, for any fixed threshold $\epsilon > 0$, a poly(d) rate for the approximation of $f_{r,\mathbf{w},\mathbf{v}}$ by one-hidden-layer networks and it can be seen as a generalization of Theorem 1 in (Safran et al., 2019). This is the content of the following corollaries.

Corollary 13 (Radial functions) *Let $f^{(d)}(\mathbf{x}) = \varphi_d(\|\mathbf{x}\|_2)$, where $\varphi_d : [-1,1] \to \mathbb{R}$ are 1-Lipschitz, and $K_d = B^d_{1,2}$. Then, for any $\epsilon \in (0,1)$ it holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \left\| f_N^\sigma - f^{(d)} \right\|_{K_d,\infty} \leq \epsilon \quad \text{for some } N \leq \exp\big(O\big(\epsilon^{-5}\log(d/\epsilon)\big)\big) \ .$$

Moreover, f_N^σ can be chosen so that $m_\infty(f_N^\sigma) \leq \exp\big(O\big(\epsilon^{-5}\log(d/\epsilon)\big)\big)$.

Consider the functions $f^{(d)} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{i\mathbf{w}_d^T(\mathbf{U}_d\mathbf{x})_+}$ for some $\mathbf{w}_d \in \mathbb{R}^{p_d}$, $\mathbf{U}_d \in \mathbb{R}^{p_d \times d}$. This is a more general version of the function $f_{r,\mathbf{w},\mathbf{v}}$ considered in section 3. If the weights are bounded, that is $m_1(f^{(d)}) = O(1)$, then Theorem 11 implies the following.

Corollary 14 (Shallow approximation of (2)) *If $r_d = O(1)$ and $K_d = B^d_{r_d,2}$, for any $\epsilon \in (0,1)$ it holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f_N^\sigma - f^{(d)}\|_{K_d,\infty} \leq \epsilon \quad \text{for some } N \leq \exp\big(O\big(\epsilon^{-2}\log(p_d/\epsilon)\big)\big) \ .$$

Moreover, f_N^σ can be chosen so that $m_\infty(f_N^\sigma) \leq \exp\big(O\big(\epsilon^{-2}\log(p_d/\epsilon)\big)\big)$.

Although the result of Corollary 14 is established for approximation in the uniform norm over the unit ball, it is not difficult to extend it to a result in L^2 over a measure that concentrated over a compact set of constant (in d) radius, such as a normalized Gaussian. A formal statement of this fact, along with the proof, is reported in section B.5. Compared with the result of section 3, Corollary 14 implies the following. The function $f^{(d)}$ can be approximated, at a poly(d) rate over a compact set of constant radius if its weights $\mathbf{w}_d, \mathbf{U}_d$ are uniformly bounded. On the other hand, if the norm of the weights grows polynomially in d, then approximation at a poly(d) rate is not possible, under a polynomially slow decaying measure. An open question is whether approximation at a poly(d) rate is possible if only one of these two conditions hold, that is if either (1) the norm of the weights is constant but the measure is polynomially slow decaying or (2) the measure is concentrated over a compact set of constant radius but the norm of the weights grows (at least) polynomially.

5. Approximation by shallow networks: a spherical harmonics analysis

As already discussed, difficulties in approximating functions in high dimension by shallow networks appear when the function has a Fourier transform spread in a (exponential) number of directions in (polynomial) high energy. On the other hand, the presence of only one of these two conditions is not enough to prevent efficient approximability. While the previous results highlight this, the lower bound presented in Theorem 5 applies to a specific choice of error measure, with (polynomially) slowly decaying tails.

In this section, we aim to disentagle the role of the measure tail and understand how the Fourier representation can tell whether a function is efficiently approximable by a one-hidden-layer network or not. In particular, we focus on approximation results for functions defined over the $(d-1)$-dimensional sphere $\mathbb{S}^{d-1}$, for which a rich literature of Fourier analysis is available.

First, we give a sufficient condition on the target function in terms of its spherical harmonics decomposition to be not efficiently approximable by shallow one-hidden-layer networks. This condition captures a slowly decaying and sufficiently spread spherical harmonic expansion. We also show that certain symmetry properties imply this condition. On the other hand, one may ask if a reverse statement holds. In this direction, building on existing theory, we provide a sufficient condition for approximation by one-hidden-layer networks.

5.1 Spherical harmonics decomposition

Let $d \geq 2$ and S^{d-1} (S when the dimension is clear from the context) be the uniform measure over $\mathbb{S}^{d-1}$. The spherical harmonics are a particular orthonormal basis for $L^2(S)$. They consists of

$$\bigcup_{k=0}^{\infty} \operatorname{span}\left(\left\{Y_{k,i}^d\right\}_{i=1}^{N_k^d}\right) = \bigcup_{k=0}^{\infty} H_k^d$$

where $Y_{k,i}^d$ is a restriction to $\mathbb{S}^{d-1}$ of an homogeneous harmonic polynomial of degree k. The projection operator over H_k^d is given by

$$\mathcal{P}_k^d : f \in L^2(S) \mapsto f_k \doteq \sum_{i=1}^{N_k^d} \langle f, Y_{k,i}^d \rangle Y_{k,i}^d \ .$$

Similarly, $\mathcal{P}_I$ denotes the operator $\oplus_{i\in I}\mathcal{P}_i^d$, for any $I \subseteq \mathbb{N}$. The function f_k is referred to as the degree k spherical harmonic component of the function f. Since the spherical harmonic form an orthonormal basis of L_S^2, it holds that $f = \sum_{k=0}^{\infty} f_k$ and $\|f\|_2^2 = \sum_{k=0}^{\infty} \|f_k\|_2^2$ for every $f \in L^2(S)$, where $\|\cdot\|_2$ denotes the norm in $L^2(S)$. As spherical harmonics decomposition can be seen as a generalization of Fourier series to dimensions $d \geq 3$, in the following we refer to the spherical harmonics decomposition of a function as its Fourier representation, interchangeably. The operator $\mathcal{P}_k$ can be associated with a kernel given by

$$\sum_{i=1}^{N_k^d} Y_{k,i}^d(\mathbf{x}) \overline{Y_{k,i}^d(\mathbf{y})} = N_k^d P_k^d\big(\mathbf{x}^T\mathbf{y}\big)$$

where

$$N_k^d = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!} = \Theta\left(\sqrt{\frac{k+d}{kd}}\frac{(k+d)^{k+d}}{k^k d^d}\frac{d^2}{(k+d)^2}\right)$$

is the dimension of H_k^d and P_k^d is the $((d-2)/2)$-Gegenbauer polynomial defined as

$$P_k^d(x) = k!\,\Gamma\left(\frac{d-1}{2}\right)\sum_{j=0}^{\lfloor k/2\rfloor}(-1)^j \frac{(1-x^2)^j x^{k-2j}}{4^j j!(k-2j)!\Gamma\big(j+\frac{d-1}{2}\big)} \ .$$

Let ω_d be the Lebesgue area of the sphere:

$$\omega_d = \omega_{d-1}\frac{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} = \Theta\left(\frac{(2\pi e)^{d/2}}{d^{d/2-1/2}}\right) = \Theta\left(\sqrt{d}\left(\frac{2\pi e}{d}\right)^{d/2}\right) .$$

The polynomials $\{(N_k^d)^{1/2}P_k^d\}_{k\geq 0}$ form a basis of orthonormal polynomials for $L^2(\mu_d)$, where μ_d is the probability measure on $[-1,1]$ defined by

$$d\mu_d(t) = \alpha_d(1-t^2)^{(d-3)/2}\,dt ,$$

where $\alpha_d = \omega_{d-1}/\omega_d = \Theta(\sqrt{d})$. Notice that, given a function $f \in L^2(S)$, it holds

$$f_k(\mathbf{x}) = N_k^d \int_{\mathbb{S}^{d-1}} f(\mathbf{y})P_k^d\big(\mathbf{x}^T\mathbf{y}\big)\,dS(\mathbf{y}) .$$

Moreover, if the function f only depends on a linear projection of the input, the Funk-Hecke formula holds.

Theorem 15 (Funk-Hecke formula) *For every $\sigma : [-1,1] \to \mathbb{C}$ such that $\mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sigma(x_1)$ is in $L^2(S)$, and for every $\mathbf{w} \in \mathbb{S}^{d-1}$, it holds that*

$$\int_{\mathbb{S}^{d-1}} \sigma(\mathbf{w}^T\mathbf{x})P_k^d(\boldsymbol{\xi}^T\mathbf{x})\,dS(\mathbf{x}) = \lambda_k P_k^d(\boldsymbol{\xi}^T\mathbf{w})$$

where $\lambda_k = \langle\sigma, P_k^d\rangle_{\mu_d}$.

Functions of the form

$$\mathbf{x} \in \mathbb{S}^{d-1} \mapsto \alpha P_k^d(\mathbf{w}^T\mathbf{x})$$

for some $\alpha \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{S}^{d-1}$, are called zonal harmonics. By the Funk-Hecke formula it follows that

$$\int_{\mathbb{S}^{d-1}} P_k^d(\mathbf{w}^T\mathbf{x})P_k^d(\mathbf{v}^T\mathbf{x})\,dS(\mathbf{x}) = (N_k^d)^{-1}P_k^d(\mathbf{w}^T\mathbf{v})$$

for any $\mathbf{w}, \mathbf{v} \in \mathbb{S}^{d-1}$. This implies that H_k^d has an RKHS structure with kernel K given by

$$K(\mathbf{v},\mathbf{w}) \doteq N_k^d P_k^d(\mathbf{v}^T\mathbf{w}) .$$

In particular, zonal harmonics actually span H_k^d. Moreover, it can be shown that there exists $\mathbf{w}_1, \ldots, \mathbf{w}_{N_k^d} \in \mathbb{S}^{d-1}$ such that $H_k^d = \operatorname{span}(\{P_k^d(\mathbf{w}_i^T\cdot)\}_{i=1}^{N_k^d})$ (Efthimiou and Frye, 2014, Theorem 4.13). For these facts and more details about spherical harmonics we refer to the books (Atkinson and Han, 2012; Dai and Xu, 2013).

5.2 Concentration and spreadness in H_k^d and main results

Intuitively, one can say function $f \in C(\mathbb{S}^{d-1})$ is concentrated over $\mathbb{S}^{d-1}$ if there is an area $\Omega \subset \mathbb{S}^{d-1}$ such that the mass of f is concentrated over Ω. On the other hand one could say that f is spread if it assumes non-negligible values uniformly over the sphere. The

spreadness/concentration of the function f can be quantified by looking at ratios of the type

$$\ell_{q,p}(f) \doteq \frac{\|f\|_q}{\|f\|_p}$$

for $1 \leq p < q \leq \infty$. Since the norms above are with respect to a probability measure, it holds that $\ell_{q,p} \geq 1$. Intuitively, the closest this ratio is to 1, the more spread is the function. On the other hand, the largest this ratio, the more concentrated the function is. Consider the case of a function $f_k \in H_k^d$. Then, it holds that

$$\ell_{\infty,2}(f_k) \leq \sqrt{N_k^d}$$

The equality is attained for functions of the type $f_k(\mathbf{x}) = \alpha P_k^d(\mathbf{w}^T\mathbf{x})$ for some $\alpha \in \mathbb{C}$ and $\mathbf{w} \in \mathbb{S}^{d-1}$, i.e. zonal harmonics. In this sense, zonal harmonics could be considered as the most concentrated functions in H_k^d. A similar inequality can be shown for the quantity $\ell_{2,1}$: it holds that

$$\ell_{2,1}(f_k) \leq \sqrt{N_k^d} \tag{6}$$

for $f_k \in H_k^d$. Nevertheless, in this case, zonal harmonics do not attain equality; the inequality is actually not tight; a more detail discussion on this quantity is reported in section 5.4.

Thanks to the Funk-Hecke formula, it holds that a one-hidden-layer $f_N \in \mathcal{F}_N$, with hidden layer weights given by $\mathbf{w}_1, \dots, \mathbf{w}_N$, satisfies

$$\mathcal{P}_k^d f_N = \sum_{j=1}^{d} \alpha_j P_k^d(\mathbf{w}_j^T\mathbf{x})$$

for some $\boldsymbol{\alpha} \in \mathbb{C}^N$. In other words, its Fourier representation is concentrated along N directions. According to the remarks above, this implies that if the width N is relatively small, the Fourier components of the neural network f_N are relatively concentrated in space. One would then expect that such concentration can be used to determine whether a function can be approximated efficiently by a one-hidden-layer neural network or not. In the next sections, we show that this is indeed the case. Let $f \in C(\mathbb{S}^{d-1})$; assuming that $\|f_k^{(d)}\|_2 \simeq \text{poly}(d, k^{-1})$, the results can be informally summarized as follows:

- If the spherical components of f are (exponentially) spread in $\ell_{\infty,2}$ sense, that is, for example,

  $$\ell_{\infty,2}(f_k) \lesssim \epsilon^k \cdot \sqrt{N_k^d} = \epsilon^k \cdot \sup_{g \in H_k^d} \ell_{\infty,2}(g) \quad \text{for some } \epsilon \in (0,1)$$

  then f is provably not universally approximable by one-hidden-layer networks.

- If the spherical components of f are (polynomially) concentrated in $\ell_{2,1}$ sense, that is, for example,

  $$\ell_{2,1}(f_k) \gtrsim \text{poly}(d^{-1}, k^{-1})\sqrt{N_k^d}$$

  then f is universally approximable by one-hidden-layer networks.

Notice that, on the other hand, if $\|f_k\|_2$ decreases exponentially fast then universal approximation follows, and similarly if $\|f_k\|_2$ decreases exponentially slower than a power of k^{-1} then universal approximation can not hold. The first of the two conditions above expresses concentration of the Fourier decomposition, while the second expresses spreadness of the same. We notice at least two gaps between the two conditions. The first one is the expression of the concentration phenomena: one is with respect to $\ell_{\infty,2}$, while the other one is with respect to $\ell_{2,1}$. Second, the two regimes above do not include many other possible ones. For example, we suspect the existence of a regime which prevents universal approximability but allows for fixed-threshold one, a topic worth of future study. These results are properly formalized, stated and discussed in section 5.3 and section 5.4, respectively.

5.3 Inapproximability of functions with spread Fourier representation

As discussed above, one-hidden-layer functions have a *zonal* structure. In more detail, if $h(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$ for some $\mathbf{w} \in \mathbb{S}^{d-1}$ and $b \in \mathbb{R}$, then it is easy to see that

$$h_k(\mathbf{x}) = s_k \|h_k\|_2 \sqrt{N_k^d} P_k^d(\mathbf{w}^T\mathbf{x})$$

with $s_k \in \{\pm 1\}$. In particular, it follows that $\|h_k\|_\infty = |h_k(\pm\mathbf{w})| = (N_k^d)^{1/2}\|h_k\|_2$. This can be interpreted by saying that the Fourier components of single neurons are most concentrated (along the neuron direction) in space. Therefore, it is natural to expect that functions with spread Fourier decomposition are difficult to approximate by neural networks. The proposition below formalizes this fact. The proof follows a technique similar to the one used in (Daniely, 2017) (see Remark 17 for a comparison) and essentially upper bounds the scalar product between the objective function and the network.

Proposition 16 *Let $\left\{f^{(d)}\right\}_d$ a sequence of functions such that $f^{(d)} \in C(\mathbb{S}^{d-1})$ and $M > 0$. Assume that for every d there exists $I_d \subseteq \mathbb{N}$ such that*

1. *It holds that $\|f^{(d)}\|_2 \leq O(d^M) \cdot \|P_{I_d} f^{(d)}\|_2$;*
2. *There exists a non-negative sequence $\{c_{d,k}\}_{k \in I_d}$ such that $\|f_k^{(d)}\|_\infty \leq c_{d,k}\sqrt{N_k^d}\|f^{(d)}\|_2$ for all $k \in I_d$ and such that $\left(\sum_{k\in I_d} c_{d,k}^2\right)^{1/2} \leq \epsilon^{d^\alpha} \cdot O(d^M)$ for some $\epsilon \in (0,1)$ and $\alpha > 0$.*

Moreover, assume that $\|f^{(d)}\|_\infty = O(1)$ and $\|f^{(d)}\|_2 = \Omega(d^{-M})$. Then the sequence $\left\{f^{(d)}\right\}_{d>2}$ is not universally approximable by one-hidden-neural networks.

The proof of Proposition 16 is reported in section C.1. We discuss a few particular cases where the assumptions of Proposition 16 hold. Let $\left\{f^{(d)}\right\}_{d>2}$ be a sequence of functions $f^{(d)} \in C(\mathbb{S}^{d-1})$.

Example 2 (Constant control on $\ell_{\infty,2}$) *Assume that assumption 1 in Proposition 16 holds with $I_d = \left\{k \in \mathbb{N} : k \geq d^2\right\}$ and that $\|f^{(d)}\|_2 = \Omega(d^{-M})$ for some constant $M > 0$. If it holds that*

$$\ell_{\infty,2}(f_k^{(d)}) \leq \bar{\ell}$$

for all $k \geq d^2$ for some constant $\bar{\ell} \geq 1$, then it is easy to check that Proposition 16 holds. This condition could be thought as the spherical harmonic components of the function $f^{(d)}$ being uniformly spread for high energy ($k \geq d^2$). Indeed assumption 2 holds with

$$c_{d,k} \doteq \frac{\bar{\ell}}{\sqrt{N_k^d}} \frac{\|f_k^{(d)}\|_2}{\|f^{(d)}\|_2}$$

since

$$\sum_{k=d^2}^{\infty} c_{d,k}^2 \leq \frac{\bar{\ell}^2}{N_{d^2}^d} = O(d^{3-d}) \ .$$

This is similar to the condition used in (Daniely, 2017), discussed in the remark below.

Remark 17 *Daniely (Daniely, 2017) showed a depth-separation result using a result similar to Proposition 16. The difference in this case is that the author considers functions defined on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. Although, since $L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) = L^2(\mathbb{S}^{d-1}) \otimes L^2(\mathbb{S}^{d-1})$, the space $L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})$ admits a decomposition in spherical harmonics*

$$L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) = \sum_{j,k=0}^{\infty} H_j^d \otimes H_k^d \ .$$

In particular, Daniely considers functions of the type

$$f^{(d)} : (\mathbf{x}, \mathbf{y}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto h^{(d)}(\mathbf{x}^T \mathbf{y})$$

for some $h^{(d)} \in C([-1,1])$. Such functions belong to $\sum_{k=0}^{\infty} H_k^d \otimes H_k^d$ and satisfy

$$\ell_{\infty,2}(f_{k,k}^{(d)}) \leq \bar{\ell} \cdot \left(N_k^d\right)^{1/2} = \bar{\ell} \cdot \left(N_k^d\right)^{-1/2} \cdot \ell_{k,k}^*$$

where $\ell_{k,k}^* = \max_{f \in H_k^d \otimes H_k^d} \ell_{\infty,2}(f)$. The equation above resembles condition 2 in Proposition 16, since it implies that

$$\left\| f_{k,k}^{(d)} \right\|_\infty \leq \frac{\bar{\ell}}{\sqrt{N_k^d}} \frac{\|f_{k,k}^{(d)}\|_2}{\|f^{(d)}\|_2} \cdot \ell_{k,k}^* \cdot \|f^{(d)}\|_2$$

and since

$$c_d \doteq \left[\sum_{k \geq k_d} \left(\frac{\bar{\ell}}{\sqrt{N_k^d}} \frac{\|f_{k,k}^{(d)}\|_2}{\|f^{(d)}\|_2} \right)^2 \right]^{1/2} \leq \frac{\bar{\ell}}{\sqrt{N_{k_d}^d}}$$

which, for $k_d \geq d^2$ implies that $c_d \lesssim d^3 \, 2^{-d}$. The proof is then concluded by choosing $I_d = \{(k,k) : k \geq k_d\}$, since (using the same notations as in the proof of Proposition 16), it holds

$$\|f_N - f^{(d)}\|_2^2 \geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \sum_{(j,j) \in I_d} \sum_{i=1}^{N} \left(\ell_{j,j}^*\right)^{-1} |u_i| \|f_{j,j}^{(d)}\|_\infty \|f_{j,j}^{\sigma_i, \mathbf{w}_i}\|_2$$

which is an equivalent of formula (44).

Example 3 *Assume that assumption 1 in Proposition 16 holds with* $I_d = \left\{k \in \mathbb{N} : k \geq \rho d^\beta\right\}$ *for some* $\rho > 0$, $\beta > 0$ *and that* $\|f^{(d)}\|_2 = \Omega(d^{-M})$ *for some constant* $M > 0$. *If it holds that*

$$\ell_{\infty,2}(f_k^{(d)}) \leq \epsilon^k \cdot O(d^M) \cdot \sqrt{N_k^d}$$

for all $k \geq \rho d^\beta$ *for some constant* $M > 0$, *then Proposition 16 holds, since*

$$\sum_{k=\rho d^\beta}^{\infty} \epsilon^k = \frac{\epsilon^{\rho d^\beta}}{1-\epsilon} \ .$$

This condition could also be thought as the spherical harmonic components of the function $f^{(d)}$ *being uniformly spread for high energy (*$k \geq d^2$*), although in this case the spreadness is required to increase exponentially, as the degree increases, with respect to the maximum concetration achievable (that is* $(N_k^d)^{1/2}$*).*

Example 4 (Invariant functions) *Finally, we show that certain symmetry assumptions can imply energy spreadness. Consider the case of a sign-invariant function* $f \in C(\mathbb{S}^{d-1})$, *that is such that* $f(\boldsymbol{\epsilon} \circ \mathbf{x}) = f(\mathbf{x})$ *for every* $\boldsymbol{\epsilon} \in \{\pm 1\}^d$ *and* $\mathbf{x} \in \mathbb{S}^{d-1}$.

Lemma 18 *Let* $f \in C(\mathbb{S}^{d-1})$ *be a sign-invariant function. If*

$$\|f_k\|_\infty = \sup_{\boldsymbol{\epsilon} \in \{\pm 1\}^d} |f_k(\boldsymbol{\epsilon})| \tag{7}$$

for some $k \geq 16d^2$ *then it holds*

$$\|f_k\|_\infty \leq 2 \cdot 2^{-d/2} \sqrt{N_k^d} \|f_k\|_2 \ .$$

Proof *[Proof] Notice that since* f *is sign-invariant, so is* f_k. *Consider the function*

$$P : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto 2^{-d} N_k^d \sum_{\boldsymbol{\epsilon} \in \{\pm 1\}^d} P_k^d(\boldsymbol{\epsilon}^T \mathbf{x}) \ .$$

The function P *satisfies* $\|P\|_2 \leq 2 \cdot 2^{-d/2} \sqrt{N_k^d}$ *(see Lemma 46). Let* $\boldsymbol{\epsilon} \in \{\pm 1\}^d$. *Then it holds*

$$\|f_k\|_\infty = |f_k(\boldsymbol{\epsilon})| = |\langle f_k, P \rangle| \leq \|P\|_2 \|f_k\|_2 \leq 2 \cdot 2^{-d/2} \sqrt{N_k^d} \|f_k\|_2 \ .$$

This concludes the proof. ■

The statement of the above lemma therefore says that if f *is sign-invariant and achieves maximum energy in a specific frequency then it satisfies Assumption 2 from Proposition 16. Under polynomial decay of* $\|f_k\|_2$, *it should be possible to relax the condition* (7) *to ask for the frequency* $\mathbf{w}^{(k)} \in [0, \infty)^d$ *such that* $\|f_k\|_\infty = \left|f_k(\mathbf{w}^{(k)})\right|$ *to satisfy*

$$\inf_{j \in [d]} \left|w_j^{(k)}\right| \geq \text{poly}(d^{-1}) \ .$$

5.4 Efficient approximation under a sparsity condition of the spherical harmonics decomposition

Works by Barron (Barron, 1993; Klusowski and Barron, 2018) essentially show that efficient approximation holds under a sparsity condition on the Fourier transform of the function to approximate; more specifically, for $f \in L^1(\mathbb{R}^d)$, the rate of (uniform) approximation is controlled by the quantity $\int_{\mathbb{R}^d} \|\mathbf{w}\|_1^2 |\hat{f}(\mathbf{w})| \, d\mathbf{w}$. In this section we show that an equivalent control can be determined for approximation on the sphere, in terms of spherical harmonics decomposition. For technical reason, the result is estabilished for functions in $\hat{H}^d \doteq H_1^d \oplus \bigoplus_{k=1}^{\infty} H_{2k}^d$ (which correspond to the space of function in L_S^2 whose odd part is linear) and mainly for ReLu activation. We briefly discuss extensions to different activation functions in Remark 24. Consider the space of homogeneous one-hidden-layer neural networks with ReLU activations:

$$\mathcal{F}_N^{\mathrm{ReLU},0} = \left\{ f : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sum_{k=1}^{N} u_k \big(\mathbf{w}_k^T \mathbf{x}\big)_+ : \mathbf{u} \in \mathbb{R}^N, \mathbf{w}_k \in \mathbb{S}^{d-1} \right\} .$$

Since

$$\big(\mathbf{w}^T \mathbf{x}\big)_+ = \frac{1}{2} \big|\mathbf{w}^T \mathbf{x}\big| + \frac{1}{2} \big(\mathbf{w}^T \mathbf{x}\big) \ ,$$

every function in $\mathcal{F}_N^{\mathrm{ReLU},0}$ is the sum of a linear function with an even one. In other words, $\mathcal{F}_N^{\mathrm{ReLU},0} \subset \hat{H}^d$. Since any linear function belongs to $\mathcal{F}_2^{\mathrm{ReLU},0}$, it is equivalent to consider the problem of approximating even functions by homogeneous one-hidden-layer neural networks with activation $\mathrm{abs}(x) = |x|$, that is, elements of the space

$$\mathcal{F}_N^{\mathrm{abs},0} = \left\{ f : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sum_{k=1}^{N} u_k \big|\mathbf{w}_k^T \mathbf{x}\big| : \mathbf{u} \in \mathbb{R}^N, \mathbf{w}_k \in \mathbb{S}^{d-1} \right\} .$$

To study this, consider the corresponding functional space

$$\mathcal{H}^1 \doteq \{h_\pi : \pi \text{ is a signed even Radon measure}\}$$

where h_π is defined to be the function

$$h_\pi : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \int_{\mathbb{S}^{d-1}} \big|\mathbf{w}^T \mathbf{x}\big| \, d\pi(\mathbf{w}) \ .$$

The space $\mathcal{H}^1$ is a Banach space endowed with the norm $\gamma_1(h) = \inf_{h \,:\, h = h_\pi} \|\pi\|_1$. As discussed in the introduction, the space $\mathcal{H}^1$ consists of functions which are efficiently approximable by one-hidden-layer networks. More formally, the following holds.

Theorem 19 (Bourgain et al. (1989)) *Let $f \in \mathcal{H}^1$. Then it holds that*

$$\inf_{f_N \in \mathcal{F}_N^{\mathrm{abs},0}} \|f - f_N\|_\infty \leq c \frac{\gamma_1(f)}{N^{1/3}}$$

where $c > 0$ is a numerical constant. Moreover, f_N satisfying the bound can be chosen to satisfy $\gamma_1(f_N) \leq \gamma_1(f)$.

The question of interest can now be transposed to: which functions $f \in C(\mathbb{S}^{d-1})$ have a (polynomially) small norm $\gamma_1(f)$? One way to approach this problem is by the so-called Blaschke–Levy operator. Consider the transformation

$$T\varphi = \int_{\mathbb{S}^{d-1}} \left|\mathbf{x}^T\mathbf{y}\right| \varphi(\mathbf{y})\, dS(\mathbf{y})$$

for functions $\varphi \in C(\mathbb{S}^{d-1})$. T can be described in terms of spherical harmonics (Rubin, 1998) as

$$T\varphi = \sum_{k\geq 0 \text{ even}} \sigma_k \varphi_k \quad \text{where} \quad \sigma_k = \frac{(-1)^{1+k/2}}{2\pi} \frac{\Gamma((k-1)/2)\Gamma(d/2)}{\Gamma((k+d+1)/2)} \ .$$

In particular, it holds that the functional T is an automorphism of $C^\infty_{even}(\mathbb{S}^{d-1})$ (the set of even function in $C^\infty(\mathbb{S}^{d-1})$) (Rubin, 1998) . Clearly, its inverse can be defined in terms of spherical harmonics by

$$T^{-1} : \varphi \in C^\infty_{even}(\mathbb{S}^{d-1}) \mapsto \sum_{k\geq 0 \text{ even}} \sigma_k^{-1} \varphi_k \ .$$

The following is immediate.

Proposition 20 *For any $\varphi \in C^\infty_{even}(\mathbb{S}^{d-1})$ it holds that $\varphi \in \mathcal{H}^1$ and*

$$\gamma_1(\varphi) = \left\| T^{-1}\varphi \right\|_1 \ .$$

Using these results, we can proceed similarly to the work (Ongie et al., 2019) and obtain the following.

Proposition 21 *Let $f \in C(\mathbb{S}^{d-1})$ even. It holds that $f \in \mathcal{H}^1$ if and only if*

$$\sup_{\varphi \in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty \leq 1} \langle T^{-1}\varphi, f\rangle < \infty \ . \tag{8}$$

In this case,

$$\gamma_1(f) = \sup_{\varphi \in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty \leq 1} \langle T^{-1}\varphi, f\rangle \ .$$

The proof of Proposition 21 is reported in section C.3. Functions that satisfy equation (8) include all even functions in $C^{d+2}(\mathbb{S}^{d-1})$ if d is even and all even functions in $C^{d+3}(\mathbb{S}^{d-1})$ if d is odd (Weil, 1976). This is inline with existing results that show approximability by neural networks for functions whose regularity is proportional to the dimension d (e.g. (Maiorov and Meir, 2000)).

Given $f \in C(\mathbb{S}^{d-1})$ even, the condition of Proposition 21 is implied by the (weak) convergence (as $N \to \infty$) of the series

$$S_N f = \sum_{k=0}^{N} \sigma_{2k}^{-1} f_{2k}$$

to a finite signed measure π. In this case $f = h_\pi$. In particular, a stronger condition is convergence in $L^1(S)$. This is implied if it holds that

$$\sum_{k\geq 0 \text{ even}} |\sigma_k|^{-1}\|f_k\|_1 < \infty \ . \tag{9}$$

Notice that, instead, the series converges in L^2_S if and only if

$$\sum_{k\geq 0 \text{ even}} \sigma_k^2\|f_k\|_2^2 < \infty \ .$$

This is equivalent to asking that $f \in \mathcal{H}^2$, the RKHS given by the kernel function

$$k : (\mathbf{x}, \mathbf{y}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto \int_{\mathbb{S}^{d-1}} \left|\mathbf{x}^T\mathbf{w}\right|\left|\mathbf{w}^T\mathbf{y}\right| dS(\mathbf{w}) \ .$$

Since in this case $\mathcal{H}^2$ can be described as

$$\mathcal{H}_2 \doteq \left\{h_\pi : \pi \text{ is a signed even Radon measure with an } L^2_S \text{ density}\right\} \ ,$$

it is clear that $\mathcal{H}^1 \subset \mathcal{H}^2$. We refer to (Bach, 2017) for more details about these statements. On the other hand, notice that the condition (9) is potentially much stronger than simply asking for $f \in \mathcal{H}^1$.

Example 5 (Highly concentrated function) *Some computations show that*

$$|\sigma_k|^{-1} \leq \Theta\left(d^{3/4}k^2\sqrt{N_k^d}\right) \ . \tag{10}$$

Using these observations it is then straightforward to prove the following.

Proposition 22 *Let $\left\{f^{(d)}\right\}_d$ a sequence of even functions in $C(\mathbb{S}^{d-1})$. Assume that there exist some constant $M, N > 0$ constant such that*

$$\sqrt{N_k^d}\|f_k^{(d)}\|_1 \leq O\big(k^M d^N\big) \cdot \|f_k^{(d)}\|_2 \quad \textit{and} \quad \sum_{k=0}^{\infty} k^{M+2}\|f_k^{(d)}\|_2 = O(d^N) \ .$$

Then the sequence $\left\{f^{(d)}\right\}_{d\geq 2}$ is universally approximable by the space $\mathcal{F}_N^{\text{abs},0}$.

Proof *[Proof] By Proposition 20 and equation* (10) *above we get that*

$$\gamma_1(f^{(d)}) \leq \sum_{k\geq 0 \ even} |\sigma_k|^{-1}\|f_k^{(d)}\|_1 \leq \Theta(d^{N+3/4}) \sum_{k\geq 0 \ even} k^{2+M}\|f_k^{(d)}\|_2 \leq O(d^{3/4+2N}) \ .$$

The application of Theorem 19 concludes the proof. ■

The proposition above requires essentially two conditions to hold. First, that the energy of the functions decreases fast enough (yet polynomially in k and d). The second condition is that the Fourier components of the function are concentrated enough, that is they are

polynomially close to the bound (6). *We remark that this condition is infact pretty strong; it requires the function f to be band-limited. According to (Dai et al., 2016), it holds that*

$$\|f_k^{(d)}\|_2 \leq C(d) k^{\frac{d-2}{4}} \|f_k^{(d)}\|_1 \ ,$$

for some function $C(d)$. Then $f^{(d)}$ would satisfy

$$\sqrt{N_k^d} \leq \mathrm{poly}(k,d) \frac{\|f_k^{(d)}\|_2}{\|f_k^{(d)}\|_1} \leq \mathrm{poly}(k,d) k^{\frac{d-2}{4}}$$

Since $\sqrt{N_k^d} \geq c(d) k^{\frac{d-2}{2}}$ for some $c(d)$, this implies that $k^{\frac{d-2}{4}} \mathrm{poly}(k^{-1}) \leq H(d)$ for some function $H(d)$. It follows that k must satisfy $k \leq K(d)$ for some $K(d)$. Although, the rate of the function $K(d)$ does not follow from (Dai et al., 2016); we conjecture that $K(d)$ behaves as a power of d.

Example 6 (High energy zonal harmonics) *The properties discussed in this section indicate that high-energy only does not yield not-universal-approximability. As an 'extreme' case, consider the case of a zonal harmonic $f(\mathbf{x}) \doteq P_k^d(\mathbf{w}^T\mathbf{x})$, for $\mathbf{x}, \mathbf{w} \in \mathbb{S}^{d-1}$ where $\mathbf{w}$ is fixed. Notice that $\|f\|_\infty = 1$. It holds that*

$$\gamma_1(f) = \frac{\|f_k\|_1}{|\sigma_k|} \leq O(k^2 d^{3/4}) \sqrt{N_k^d} \|f_k\|_1 \leq O(k^2 d^{3/4}) \left\| \sqrt{N_k^d} f_k \right\|_2 = O(k^2 d^{3/4}) \ ,$$

which implies universal approximability by Theorem 19. Similarly, polynomial combinations of zonal harmonics can be well approximated, as expected.

Remark 23 (Ridge functions) *For a single neuron network $f(\mathbf{x}) = |\mathbf{w}^T\mathbf{x}|$, it holds $\|f\|_\infty = 1$ and $\|f\|_2 = d^{-1/2}$. The spherical components of f are given by*

$$f_k(\mathbf{x}) = N_k^d \Big[\Big(T \Big[P_k^d(\mathbf{x}^T \cdot) \Big] \Big)(\mathbf{w}) \Big] = (\sigma_k N_k^d) P_k^d(\mathbf{w}^T\mathbf{x}) \ .$$

In particular, it holds

$$1 = \gamma_1(f) = \left\| \sum_{k \geq 0 \ even} \sigma_k^{-1} f_k \right\|_1 = \left\| \sum_{k \geq 0 \ even} N_k^d P_k^d(\mathbf{w}^T \cdot) \right\|_1 \ .$$

Therefore, understanding how tight (or strong) condition (9) *is highly correlated with understanding convergence of the series $\sum_{k \geq 0 \ even} N_k^d \left\| P_k^d(\mathbf{w}^T \cdot) \right\|_1$, or equivalently, computing $\left\| P_k^d \right\|_{\mu_d, 1}$.*

Remark 24 *While the result of this section mainly concern approximation by homogeneous one-hidden-layer networks with the ReLU (or absolute value) activation, they can easily be extended to any other activation satisfying Assumption 1, under the same assumptions. Moreover, notice that, thanks to Theorem 19, universal approximation by $\mathcal{F}_N^{\mathrm{ReLU},0}$ is equivalent to universal approximation by $\mathcal{H}_1 \oplus H_1^d$.*

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

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

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

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

Jean Bourgain, Joram Lindenstrauss, and Vitali Milman. Approximation of zonoids by zonotopes. *Acta mathematica*, 162(1):73–141, 1989.

Guy Bresler and Dheeraj Nagaraj. Sharp representation theorems for relu networks with precise dependence on depth. *arXiv preprint arXiv:2006.04048*, 2020.

Kaifeng Bu, Yaobo Zhang, and Qingxian Luo. Depth-width trade-offs for neural networks via topological entropy. *arXiv preprint arXiv:2010.07587*, 2020.

John Charles Burkill. *Lectures on approximation by polynomials*, volume 16. Tata Institute of Fundamental Research, 1959.

Vaggos Chatziafratis, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Depth-width trade-offs for relu networks via sharkovsky's theorem. *arXiv preprint arXiv:1912.04378*, 2019.

Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*, volume 23. Springer, 2013.

Feng Dai, Han Feng, and Sergey Tikhonov. Reverse hölder's inequality for spherical harmonics. *Proceedings of the American Mathematical Society*, 144(3):1041–1051, 2016.

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

Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions*. World Scientific, 2014.

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

Ingo Gühring, Mones Raslan, and Gitta Kutyniok. Expressivity of deep neural networks. *arXiv preprint arXiv:2007.04759*, 2020.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. *arXiv preprint arXiv:1901.09021*, 2019a.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems*, pages 359–368, 2019b.

Daniel Hsu, Clayton Sanford, Rocco A Servedio, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. *arXiv preprint arXiv:2102.02336*, 2021.

Shirin Jalali, Carl Nuzman, and Iraj Saniee. Efficient deep learning of gmms. *arXiv preprint arXiv:1902.05707*, 2019.

Jason M Klusowski and Andrew R Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

Shiyu Liang and Rayadurgam Srikant. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

VE Maiorov and Ron Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics*, 13(1): 79–103, 2000.

Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? In *Advances in Neural Information Processing Systems*, pages 6426–6435, 2019.

Eran Malach, Gilad Jehudai, Shai Shalev-Shwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks. *arXiv preprint arXiv:2102.00434*, 2021.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

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

Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Philipp Christian Petersen. Neural network theory, 2020. URL `http://pc-petersen.eu/Neural_Network_Theory.pdf`.

Allan Pinkus. Approximation theory of the mlp model. *Acta Numerica 1999: Volume 8*, 8: 143–195, 1999.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org, 2017.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

Donsub Rim, Luca Venturi, Joan Bruna, and Benjamin Peherstorfer. Depth separation for reduced deep networks in nonlinear model reduction: Distilling shock waves in nonlinear hyperbolic problems. *arXiv preprint arXiv:2007.13977*, 2020.

Theodore J Rivlin. *An introduction to the approximation of functions*. Courier Corporation, 1981.

David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. *arXiv preprint arXiv:1705.05502*, 2017.

Boris Rubin. Inversion of fractional integrals related to the spherical radon transform. *journal of functional analysis*, 157(2):470–487, 1998.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2979–2987. JMLR. org, 2017.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: What is actually being separated? *arXiv preprint arXiv:1904.06984*, 2019.

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

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

Gal Vardi, Daniel Reichman, Toniann Pitassi, and Ohad Shamir. Size and depth separation in approximating natural functions with neural networks. *arXiv preprint arXiv:2102.00314*, 2021.

Wolfgang Weil. Centrally symmetric convex bodies and distributions. *Israel Journal of Mathematics*, 24(3):352–367, 1976.

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

Appendix A. Proofs of depth-separation results

A.1 Proof of Theorem 5

The proof of the lower bound follows the same strategy as (Eldan and Shamir, 2016). For sake of simplicity in the following we remove the dimension d from the following notations: $\mathbf{w}_d = \mathbf{w}$ and $\mathbf{v}_d = \mathbf{v}$. In the following we always assume $d \geq 3$. Let $S \subseteq [d]$ a subset and let $\mathbf{I}_S$ be the truncated identity matrix defined as

$$\mathbf{I}_S := \sum_{s \in S} \mathbf{e}_s \mathbf{e}_s^\top .$$

Moreover, define the function $H_S(\mathbf{x})$ as

$$H_S(\mathbf{x}) \doteq \prod_{i: i \in S} \mathbf{1}_{x_i > 0} \prod_{j: j \in [d] \setminus S} \mathbf{1}_{x_j \leq 0} .$$

Lastly, for a subset $S \subseteq [d]$, let $\mathbf{v}_S := \mathbf{v} + \mathbf{I}_S \mathbf{w}$ and define the function $\sigma_{r,S}(\mathbf{x}) := \sigma_r(\mathbf{v}_S^T \mathbf{x})$. Therefore, the expression of $f_{r_d,\mathbf{w},\mathbf{v}}$ can be rewritten as:

$$f_{r_d,\mathbf{w},\mathbf{v}}(\mathbf{x}) = \sum_{S \subseteq [d]} g_S(\mathbf{x}) = \sum_{S \subseteq [d]} H_S(\mathbf{x}) \sigma_{r_d,S}(\mathbf{x})$$

where $g_S(\mathbf{x}) := H_S(\mathbf{x}) \sigma_{r_d,S}(\mathbf{x})$. Let the space of N-units one-hidden-layer networks be

$$\mathcal{F}_N = \left\{ f_N : \mathbf{x} \in \mathbb{R}^r \mapsto \sum_{k=1}^{N} \sigma_k(\mathbf{a}_k^T \mathbf{x}) : \mathbf{a}_k \in \mathbb{R}^d, \, \sigma_k \text{ are 1-Lipschitz activations} \right\}.$$

Assume that

(A1) it holds that $\tau_d \cdot r_d \geq \beta d^k$ for some constant $k \geq 1$;

(A2) it holds that $\eta > \log_2\left(\|\psi\|_1 \sqrt{K/2} \right)$

Then, for large enough d, it holds

$$\inf_{f \in \mathcal{F}_N} \|f_{r_d,\mathbf{w},\mathbf{v}} - f\|_\varphi^2 \geq 1 - N \left(2^{1-2\eta} K \|\psi\|_1^2 \right)^d O(d \cdot \tau_d \cdot r_d) , \tag{11}$$

where we denote

$$\|g\|_\varphi^2 \doteq \int_{\mathbb{R}^d} |g(\mathbf{x})|^2 \varphi^2(\mathbf{x}) \, d\mathbf{x}$$

for $g \in L^2_{\varphi^2}$. In particular, if $N \simeq \text{poly}(d)$, then the error (11) tends to 1 as $d \to \infty$.

To show equation (11), we proceed as follows. Let $\mathcal{F} = \{\widehat{f\varphi} : f \in \mathcal{F}_1\}$, and denote by $F := \widehat{\varphi \cdot f_{r_d,\mathbf{w},\mathbf{v}}} = \hat{f}_{r_d,\mathbf{w},\mathbf{v}} * \hat{\varphi}$. Since $\hat{\varphi}$ has compact support in $[-K, K]^d$ and the Fourier transform of a one-unit shallow network $f(\mathbf{x}) = \sigma(\mathbf{x}^T \mathbf{a})$ has support in the line $\{\boldsymbol{\xi} : \boldsymbol{\xi} = \alpha \mathbf{a}, \, \alpha \in \mathbb{R}\}$, it follows that any function in $\mathcal{F}$ is supported in a tube $T = \{\boldsymbol{\xi} : \boldsymbol{\xi} = \alpha \mathbf{a} + [-K, K]^d, \, \alpha \in \mathbb{R}\}$ of radius K. For each tube T of radius K, we consider $\mathcal{T}_T = \{\phi \in L^2 : \text{ supp}(\phi) \subseteq T\}$ and

$$\kappa \doteq \sup_{T \text{ tube of radius } K} \|P_{\mathcal{T}_T}(F)\|_2 ,$$

where $P_{\mathcal{T}_T}(F) = \text{argmin}_{h\in\mathcal{T}_T}\|h-F\|_2^2$. We claim that

$$\inf_{f\in\mathcal{F}_N}\|f_{r_d,\mathbf{w},\mathbf{v}}-f\|_\varphi^2 \geq 1-N\kappa^2\ . \tag{12}$$

Indeed, given $f\in\mathcal{F}_N$, denote by $T_1,\dots T_N$ the associated N tubes, and by $\mathcal{T}_{T_1,\dots T_N} = \bigoplus_{k\in[N]}\mathcal{T}_{T_k}$ the corresponding subspace spanned by $\mathcal{T}_{T_k}$, $k\in[N]$. Then, by using the isometry of the Fourier transform, we have that

$$\begin{aligned}\inf_{f\in\mathcal{F}_N}\|f-f_{r_d,\mathbf{w},\mathbf{v}}\|_\varphi^2 &= \inf_{f\in\mathcal{F}_N}\|\widehat{f\varphi}-F\|_2^2\\ &\geq \inf_{T_1,\dots T_N}\inf_{h\in\mathcal{T}_{T_1,\dots T_N}}\|h-F\|_2^2\\ &= \inf_{T_1,\dots T_N}\|P_{\mathcal{T}_{T_1,\dots T_N}}(F)-F\|_2^2\\ &= \inf_{T_1,\dots T_N}(\|F\|_2^2-\|P_{\mathcal{T}_{T_1,\dots T_N}}(F)\|_2^2)\ .\end{aligned} \tag{13}$$

Now, observe that $\sup_{T_1,\dots,T_N}\|P_{\mathcal{T}_{T_1,\dots T_N}}(F)\|_2^2 \leq N\sup_T\|P_{\mathcal{T}_T}(F)\|_2^2$. Equation (13) therefore becomes

$$\inf_{f\in\mathcal{F}_N}\|f-f_{r_d,\mathbf{w},\mathbf{v}}\|_\varphi^2 \quad\geq\quad \|F\|_2^2 - N\sup_T\|P_{\mathcal{T}_T}(F)\|_2^2\ ,$$

which proves (12) by plugging in the definition of κ and recalling that $\|F\|_2^2 = \|f_{r_d,\mathbf{w},\mathbf{v}}\|_\varphi^2 = 1$ by Parseval. To establish (11), it is therefore sufficient to prove that

$$\kappa^2 \leq \left(\|\psi\|_1^2 2^{1-2\eta}K\right)^d O(d\cdot\tau_d\cdot r_d)\ . \tag{14}$$

The rest of the proof will be devoted to establishing a sufficiently sharp upper bound for $\|P_{\mathcal{T}_T}(F)\|_2$. Observe that $P_{\mathcal{T}_T}(F)$ is simply obtained by setting to zero all frequencies of F outside T. We start by computing an upper bound on $|F(\boldsymbol{\xi})|$. We claim the following.

Lemma 25 *It holds that*

$$|F(\boldsymbol{\xi})| \quad\leq\quad \frac{\|\varphi\|_1}{2^d}\sum_{S\subseteq[d]}\prod_{j=1}^d \min\left(1,\frac{2K}{\pi(|\xi_j-\xi_{S,j}|-K)_+}\right)\ .$$

Let $D(\boldsymbol{\xi}) \doteq \sum_S D_S(\boldsymbol{\xi})$, with $D_S(\boldsymbol{\xi}) \doteq \prod_{j=1}^d \min\left(1,\frac{2K}{\pi(|\xi_j-\xi_{S,j}|-K)_+}\right)$, so that from Lemma 25 we have

$$|F(\boldsymbol{\xi})| \leq 2^{-d}\|\varphi\|_1 D(\boldsymbol{\xi})\ . \tag{16}$$

Recall that $\tau_d = \sup_{S\in[d]}\|\mathbf{v}_S\|_\infty$. Given $\boldsymbol{\xi}$ non-zero, we claim the following.

Lemma 26 *It holds that*

$$D(\boldsymbol{\xi}) \leq C_{K,\gamma}2^{d(1-\eta)}\min\left\{1, 2K(\pi(\|\boldsymbol{\xi}\|_\infty - r_d\tau_d - K)_+)^{-1}\right\}\ , \tag{17}$$

where $C_{K,\gamma} = 2\exp\left(\sqrt{\frac{8K}{\pi\gamma}}\right)$.

Now, pick any arbitrary non-zero direction $\boldsymbol{\nu}$ such that $\|\boldsymbol{\nu}\|_\infty = 1$. Let

$$T = \{\boldsymbol{\xi} : \inf_{\alpha\in\mathbb{R}} \|\boldsymbol{\xi} - \alpha\boldsymbol{\nu}\|_\infty \leq K\} \tag{18}$$

denote the tube of radius K in the direction $\boldsymbol{\nu}$. It holds that

$$\int_T D(\boldsymbol{\xi})^2 d\boldsymbol{\xi} \;=\; \underbrace{\int_{T\cap\{\|\boldsymbol{\xi}\|_\infty \leq 2\tau_d r_d\}} D(\boldsymbol{\xi})^2 d\boldsymbol{\xi}}_{t_1} + \underbrace{\int_{T\cap\{\|\boldsymbol{\xi}\|_\infty > 2\tau_d r_d\}} D(\boldsymbol{\xi})^2 d\boldsymbol{\xi}}_{t_2} \;.$$

In order to control the two terms t_1 and t_2, we use the following lemma to upper bound the measure of a ℓ_∞-cylinder.

Lemma 27 *Let T be an ℓ_∞-tube of radius K as defined in* (18)*. If μ denotes the d-dimensional Lebesgue measure, then*

$$\mu\left(T \cap [-R, R]^d\right) \leq 8e^2(d-1)(K+R)(2K)^{d-1} \;. \tag{20}$$

Moreover, if $g : \mathbb{R} \to \mathbb{R}$ is in $L^1(\mathbb{R})$ and non-increasing, then

$$\int_{T\cap\{\|\boldsymbol{\xi}\|_\infty > R\}} g(\|\boldsymbol{\xi}\|_\infty)\, d\boldsymbol{\xi} \leq 4e^2(d-1)(2K)^{d-1} \int_{R-K(2+3/(d-1))}^{\infty} g(u)du \;, \tag{21}$$

as long as $R > K(2 + 3/(d-1))$.

From (17) and (20), the first term of (19) can be bounded as

$$\begin{aligned} t_1 &\leq 8e^2 C_{K,\gamma}^2 2^{2d(1-\eta)+(d-1)} K^{d-1}(d-1)(K + 2\tau_d r_d) \\ &\leq D_{K,\gamma}^{(1)} \cdot d \cdot (\tau_d r_d)\Big(2^{2(1-\eta)+1}K\Big)^d \end{aligned} \tag{22}$$

for $D_{K,\gamma}^{(1)} = 16e^2K^{-1}C_{K,\gamma}^2$ and d large enough, such that $2\tau_d r_d \geq K$. Similarly, using (21), the second term t_2 in turn can be bounded as

$$\begin{aligned} t_2 &\leq 8e^2\pi^{-2}C_{K,\gamma}^2 d\Big(2^{2(1-\eta)+1}K\Big)^d \int_{2\tau_d r_d - K(2+3/(d-1))} (u - \tau_d r_d - K)^{-2}\, du \\ &= 8e^2\pi^{-2}KC_{K,\gamma}^2 d\Big(2^{2(1-\eta)+1}K\Big)^d (\tau_d r_d - 3K(1 + 1/(d-1)))^{-1} \\ &\leq D_{K,\gamma}^{(2)} \cdot d \cdot \Big(2^{2(1-\eta)+1}K\Big)^d \;, \end{aligned} \tag{23}$$

for $D_{K,\gamma}^{(2)} = 16e^2\pi^{-2}C_{K,\gamma}^2$ and and d large enough, such that $\tau_d r_d \geq 10K$. Thus, collecting (22) and (23) and using (16), we obtain

$$\begin{aligned} \int_T |F(\boldsymbol{\xi})|^2 d\boldsymbol{\xi} &\leq \|\varphi\|_1^2 \cdot 2^{-2d}(t_1 + t_2) \\ &\leq d \cdot \|\varphi\|_1^2 \big(2^{1-2\eta}K\big)^d \Big(D_{K,\gamma}^{(1)}\tau_d r_d + D_{K,\gamma}^{(2)}\Big) \\ &\leq D_{K,\gamma} \cdot d \cdot \|\varphi\|_1^2 \big(2^{1-2\eta}K\big)^d \max(1, \tau_d r_d) \;, \end{aligned}$$

where

$$D_{K,\gamma} \doteq D^{(1)}_{K,\gamma} + D^{(2)}_{K,\gamma} = 32 \exp\left(2 + \sqrt{\frac{8K}{\pi\gamma}}\right)\left(\pi^{-2} + K^{-1}\right) .$$

It follows that

$$\|P_{\mathcal{T}_T}(F)\|_2^2 = \int_T |F(\boldsymbol{\xi})|^2 d\boldsymbol{\xi} \leq D_{K,\gamma} \cdot (d \cdot \tau_d \cdot r_d) \cdot \left(\|\psi\|_1^2 \, 2^{1-2\eta} K\right)^d ,$$

as long as $d \geq \left[\beta^{-1} \max(1, 10K)\right]^{1/k}$ (where β and k satisfy $\tau_d r_d \geq \beta d^k$). We have just established (14), and this concludes the proof of the theorem. In the remaining part of this section we prove the auxiliary lemmas used above.

Proof [Proof of Lemma 25] We start by computing $\hat{f}_{r_d,\mathbf{w},\mathbf{v}}$. From the definition of σ_r, it follows that

$$\hat{\sigma}_{r,S}(\boldsymbol{\xi}) = \delta(\boldsymbol{\xi} - r\mathbf{v}_S) ,$$

which combined with the definition of H yields

$$\hat{f}_{r_d,\mathbf{w},\mathbf{v}}(\boldsymbol{\xi}) = \sum_{S\subseteq[d]} \Big(\hat{H}_S * \hat{\sigma}_{r_d,S}\Big)(\boldsymbol{\xi}) = \sum_{S\subseteq[d]} \hat{H}_S(\boldsymbol{\xi} - r_d\mathbf{v}_S) .$$

Let $\boldsymbol{\xi}_S \doteq r_d\mathbf{v}_S$. It holds that

$$\begin{aligned} F(\boldsymbol{\xi}) &= \int_{\mathbb{R}^d} \hat{f}_{r_d,\mathbf{w},\mathbf{v}}(\boldsymbol{\nu})\hat{\varphi}(\boldsymbol{\xi} - \boldsymbol{\nu})\, d\boldsymbol{\nu} = \sum_{S\subseteq[d]} \int_{\mathbb{R}^d} \hat{H}_S(\boldsymbol{\nu} - \boldsymbol{\xi}_S)\hat{\varphi}(\boldsymbol{\xi} - \boldsymbol{\nu})\, d\boldsymbol{\nu} \\ &= \sum_{S\subseteq[d]} \underbrace{\int_{\mathbb{R}^d} \hat{H}_S(\boldsymbol{\nu})\hat{\varphi}(\boldsymbol{\xi} - \boldsymbol{\xi}_S - \boldsymbol{\nu})\, d\boldsymbol{\nu}}_{\doteq F_S(\boldsymbol{\xi} - \boldsymbol{\xi}_S)} . \end{aligned} \tag{24}$$

We can now bound each term F_S separately. It holds that

$$F_S(\boldsymbol{\xi}) = \int \hat{H}_S(\boldsymbol{\nu})\hat{\varphi}(\boldsymbol{\xi} - \boldsymbol{\nu}) d\boldsymbol{\nu} = \int H_S(\mathbf{x}) e^{2i\pi\boldsymbol{\xi}^T\mathbf{x}} \varphi(\mathbf{x}) d\mathbf{x} \; = \prod_{j=1}^{d} F_j(\xi_j) \tag{25}$$

where

$$F_j(t) = \int_{\mathbb{R}} \mathbb{1}\{\epsilon_j x > 0\} e^{2i\pi t x} \psi(x)\, dx , \tag{26}$$

with $\epsilon_j = \pm 1$. Assume without loss of generality that $\epsilon_j = 1$. Observe that $F_j = \check{Q}$, where

$$Q(u) = \mathbb{1}\{u > 0\}\psi(u) .$$

Since $\psi \in L^1(\mathbb{R})$ and its Fourier transform $\hat{\psi}$ has compact support in $[-K, K]$, it holds that

$$|\hat{\psi}(\tau)| \leq \|\psi\|_1 \;\text{ for }\; \tau \in [-K, K] \quad\text{and}\quad \hat{\psi}(\tau) = 0 \;\text{ for }\; |\tau| > K . \tag{27}$$

On the one hand, since ψ is even, it holds, by directly bounding (26), that

$$|F_j(t)| \leq \frac{1}{2}\int_{\mathbb{R}} |\psi(u)| du = \frac{1}{2}\|\psi\|_1 \quad \text{for all } t ,$$

and from (27) and the Hilbert transform of Q we deduce on the other hand that

$$|F_j(t)| = \frac{1}{2\pi}\left|\int_{-K}^{K} \frac{\hat{\psi}(\tau)}{t-\tau} d\tau\right| \leq \frac{2K\|\psi\|_1}{(2\pi)(|t|-K)} \quad \text{for } |t| > K \ ,$$

so that it follows that

$$|F_j(t)| \leq \frac{\|\psi\|_1}{2} \min\left(1, \frac{2K}{\pi(|t|-K)_+}\right) \ . \tag{28}$$

Thus, from equations (24), (25) and (28) it follows that

$$\begin{aligned} |F(\boldsymbol{\xi})| &\leq \sum_{S\subseteq[d]} |F_S(\boldsymbol{\xi} - \boldsymbol{\xi}_S)| \\ &\leq \frac{\|\varphi\|_1}{2^d} \sum_{S\subseteq[d]} \prod_{j=1}^{d} \min\left(1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+}\right) \ , \end{aligned}$$

which proves Lemma 25. ■

Proof [Proof of Lemma 26] Let define for any $\boldsymbol{\xi} \in \mathbb{R}^d$ and $\lambda > 0$

$$\mathsf{n}(\boldsymbol{\xi}, \lambda) \doteq |\{j \in [d] : |\xi_j| > \lambda\}| \ .$$

Recall that $\mathbf{v}_S = \mathbf{v} + \mathbf{I}_S\mathbf{w}$ and $\boldsymbol{\xi}_S = r_d\mathbf{v}_S$. Observe that $\boldsymbol{\xi}_S - \boldsymbol{\xi}_{S'} = r_d(\mathbf{I}_S - \mathbf{I}_{S'})\mathbf{w}$, so

$$|\xi_{S,j} - \xi_{S',j}| = \begin{cases} r_d|w_j| & \text{if } j \in (S \cup S') \setminus (S \cap S') \\ 0 & \text{otherwise} \end{cases} \ . \tag{29}$$

If $\mathsf{d}(S, S')$ denotes the Hamming distance between two subsets S, S', then for all S, S', the following holds.

Lemma 28 *It holds that*

$$\mathsf{n}(\boldsymbol{\xi}_S - \boldsymbol{\xi}_{S'}, \gamma d^2) = \mathsf{d}(S \cap \Omega_d, S' \cap \Omega_d) \ . \tag{30}$$

This immediately implies that

$$\mathsf{n}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \frac{\gamma d^2}{2}\right) + \mathsf{n}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_{S'}, \frac{\gamma d^2}{2}\right) \geq \mathsf{d}(S \cap \Omega, S' \cap \Omega) \quad \text{for all } \boldsymbol{\xi} \text{ and } S \neq S' \ . \tag{31}$$

Indeed, if that was not the case, applying the triangle inequality coordinate-wise would contradict equation (30). The first upper bound is obtained by first noticing that, for $d > 2\sqrt{K/\gamma}$, it holds

$$D_S(\boldsymbol{\xi}) \leq \left(\pi(\gamma d^2/2 - K)/(2K)\right)^{-\mathsf{n}(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \gamma d^2/2)} \quad \text{for all } S \text{ and } \xi \ .$$

Now, defining $S^*_{\boldsymbol{\xi}} = \arg\min_{S\subseteq[d]} \mathsf{n}(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \gamma d^2/2)$, from (31) it follows that

$$\mathsf{n}(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \gamma d^2/2) \geq \frac{\mathsf{d}(S \cap \Omega_d, S' \cap \Omega_d)}{2} \quad \text{for all } S \neq S^*_{\boldsymbol{\xi}}$$

and thus, for $d > 2\sqrt{K/\gamma}$, it holds

$$\begin{aligned}
D(\boldsymbol{\xi}) &= D_{S^*_{\boldsymbol{\xi}}}(\boldsymbol{\xi}) + \sum_{S \neq S^*_{\boldsymbol{\xi}}} D_S(\boldsymbol{\xi}) \\
&\leq D_{S^*_{\boldsymbol{\xi}}}(\boldsymbol{\xi}) + \sum_{s=1}^{|\Omega_d|} \sum_{S \,:\, \mathsf{d}(S\cap\Omega_d, S^*_{\boldsymbol{\xi}}\cap\Omega_d)=s} \left(\pi(\gamma d^2/2 - K)/(2K)\right)^{-s/2} \\
&\leq D_{S^*_{\boldsymbol{\xi}}}(\boldsymbol{\xi}) + 2^{d-|\Omega_d|} \sum_{s=1}^{|\Omega_d|} \binom{|\Omega_d|}{s} \left(\pi(\gamma d^2/2 - K)/(2K)\right)^{-s/2} \\
&\leq 1 + 2^{d-|\Omega_d|} \left(1 + \frac{1}{\sqrt{\pi(\gamma d^2/2 - K)/(2K)}}\right)^{|\Omega_d|} \\
&\leq C_{K,\gamma} 2^{d(1-\eta)}
\end{aligned} \tag{32}$$

since $|\{S : \mathsf{d}(S \cap \Omega_d, S^*_{\boldsymbol{\xi}} \cap \Omega_d) = s\}| \leq 2^{d-|\Omega_d|}\binom{|\Omega_d|}{s}$. The term $C_{K,\gamma}$ is a constant that depends only on K and γ; in particular, we can choose $C_{K,\gamma} = 2\exp\left(\sqrt{\frac{8K}{\pi\gamma}}\right)$. The second upper bound is obtained using the above argument as follows. Let $q_{\boldsymbol{\xi}} = \arg\max_j |\xi_j|$. Since $\|\boldsymbol{\xi}_S\|_\infty \leq r_d\tau_d$ for any $S \subseteq [d]$, it holds that

$$\begin{aligned}
D(\boldsymbol{\xi}) &\leq \sum_{S\subseteq[d]} \frac{2K}{\pi(|\xi_{q_\xi} - \xi_{S,q_\xi}| - K)_+} \cdot \prod_{j\neq q_\xi} \min\left(1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+}\right) \\
&\leq 2K(\pi(\|\boldsymbol{\xi}\|_\infty - \tau_d r_d - K)_+)^{-1} \sum_{S\subseteq[d]} \prod_{j\neq q_\xi} \min\left(1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+}\right) \\
&\leq C_{K,\gamma} 2K(\pi(\|\boldsymbol{\xi}\|_\infty - \tau_d r_d - K)_+)^{-1} \cdot 2^{d(1-\eta)}
\end{aligned} \tag{33}$$

by noticing that the argument leading to (32) can now be repeated for the $(d-1)$-dimensional vector $\check{\boldsymbol{\xi}} = (\xi_1, \ldots, \xi_{q_{\boldsymbol{\xi}}-1}, \xi_{q_{\boldsymbol{\xi}}+1}, \ldots \xi_d)$, so that

$$\mathsf{n}(\check{\boldsymbol{\xi}} - \check{\boldsymbol{\xi}}_S, \gamma d^2/2) \geq \frac{\mathsf{d}((S\cap\Omega_d)\setminus\{q_{\boldsymbol{\xi}}\}, (S'\cap\Omega_d)\setminus\{q_{\boldsymbol{\xi}}\})}{2} \quad \text{for all } \; S \neq S^*_{\boldsymbol{\xi}}$$

which proves (33) and concludes the proof of Lemma 26. ■

Proof [Proof of Lemma 28] In fact, it holds that the two sets $A_1 := \{j \in [d] : |\xi_{S,j} - \xi_{S',j}| \geq \gamma d^2\}$ and $A_2 := \{j \in [d] : j \in (S\cap\Omega_d)\setminus(S'\cap\Omega_d)\}$ are equal. Let $j \in A_1$. Then $|\xi_{S,j} - \xi_{S',j}| > \gamma d^2$. Since this quantity is nonzero, equation (29) indicates that therefore $j \in S\setminus S'$ without loss of generality. Moreover, $|\xi_{S,j} - \xi_{S',j}| = r_d|w_j|$ which implies that $r_d|w_j| > \gamma d^2$ and $j \in \Omega_d$. We conclude that $j \in (S\cap\Omega_d)\setminus(S'\cap\Omega_d)$ which implies that $j \in A_2$. Now, let $j \in A_2$. Then, without loss of generality, $j \in (S\cap\Omega_d)\setminus(S'\cap\Omega_d)$. Then, it holds $r|w_j| > \gamma d^2$ since $j \in S\setminus S'$ according to (29) and $|\xi_{S,j} - \xi_{S',j}| = r_d|w_j|$. Combining these two facts, it follows that $|\xi_{S,j} - \xi_{S',j}| > \gamma d^2$ which means that $j \in A_2$. ■

Proof [Proof of Lemma 27] Let

$$\begin{aligned} T_R(\boldsymbol{\nu}) &= T(\boldsymbol{\nu}) \cap [-R,R]^d \\ &= \{\boldsymbol{\xi} : \inf_{\alpha\in\mathbb{R}} \sup_{j\in[d]} |\xi_j - \alpha\nu_j| \leq K \text{ and } \|\boldsymbol{\xi}\|_\infty \leq R\} \ . \end{aligned}$$

The aim is to upper bound the volume of $T_R(\boldsymbol{\nu})$ for any $\boldsymbol{\nu}$. Assume, without loss of generality, that $\|\boldsymbol{\nu}\|_\infty = 1$. The cut-off tube $T_R(\boldsymbol{\nu})$ can be covered with ℓ_∞-balls of radius $K' = \vartheta K$ centered along the ray defined by $\boldsymbol{\nu}$, that is

$$T_R(\boldsymbol{\nu}) \subseteq \bigcup_{j=-\lfloor (K+R)/s \rfloor}^{\lfloor (K+R)/s \rfloor} \left(js\boldsymbol{\nu} + [-\vartheta K, \vartheta K]^d \right) \ . \tag{34}$$

Now, we optimize both the sampling rate $s \in (0, K)$ and the radius ratio $\vartheta \geq 1$ while satisfying (34). Given s, let us first compute the smallest admissible ϑ. Any $\mathbf{x} \in T_R(\boldsymbol{\nu})$ satisfies

$$\|\mathbf{x} - (j+y)s\boldsymbol{\nu}\|_\infty \leq K$$

for some $j \in \mathbb{N}$ and $|y| < 1$. This implies that $\|\mathbf{x} - js\boldsymbol{\nu}\|_\infty \leq K + ys \leq K + s$. Therefore an admissible ϑ is given by the solution of $K + s = \vartheta K$, that is $\vartheta = 1 + sK^{-1}$. Now, the volume of

$$S_R \doteq \bigcup_{j=-\lfloor (K+R)/s \rfloor}^{\lfloor (K+R)/s \rfloor} \left(js\boldsymbol{\nu} + \left[-\left(1 + \frac{s}{K}\right) K, \left(1 + \frac{s}{K}\right) K \right]^d \right)$$

is upper bounded by

$$l(s) \doteq 4 \frac{K+R}{s} \left(2(K+s)\right)^d \ .$$

Minimizing over s gives $s = \frac{K}{d-1}$. Therefore, for all $\boldsymbol{\nu} \in \mathbb{R}^d$, it holds

$$T_R(\boldsymbol{\nu}) \leq (K+R)K^{d-1}(d-1)\left(1 + \frac{1}{d-1}\right)^d \leq (K+R)(d-1)K^{d-1}e^2 \ ,$$

which proves (20). Equation (21) is established analogously. Let $T_{>R}(\boldsymbol{\nu}) = T(\nu) \cap \{\boldsymbol{\xi} : \|\boldsymbol{\xi}\|_\infty > R\}$. Then we have that

$$T_{>R}(\boldsymbol{\nu}) \subseteq \bigcup_{j \geq \lfloor \frac{R-K}{s} \rfloor} \left(js\boldsymbol{\nu} + [-(K+s), (K+s)]^d \right) \ ,$$

where we set $s = K/(d-1)$. Since g is non-increasing, it follows that

$$\begin{aligned}
\int_{T_{>R}(\boldsymbol{\nu})} g(\|\boldsymbol{\xi}\|_\infty)\, d\boldsymbol{\xi} &\leq \sum_{|j|\geq \lfloor \frac{R-K}{s} \rfloor} \int_{\|\boldsymbol{\xi}-js\boldsymbol{\nu}\|_\infty \leq K+s} g(\|\boldsymbol{\xi}\|_\infty)\, d\boldsymbol{\xi} \\
&\leq 2(2(K+s))^d \sum_{j\geq \lfloor \frac{R-K}{s} \rfloor} g(js-(K+s)) \\
&\leq 2(2(K+s))^d \sum_{j\geq \lfloor \frac{R-K}{s} \rfloor} \frac{1}{s} \int_{(j-1)s-(K+s)}^{js-(K+s)} g(u)\, du \\
&\leq \frac{2(2(K+s))^d}{s} \int_{R-K-2s-(K+s)}^{\infty} g(u)\, du \\
&\leq \frac{2e^2(d-1)(2K)^d}{K} \int_{R-K(2+3/(d-1))}^{\infty} g(u)\, du\ .
\end{aligned}$$

This establishes (21) and concludes the proof. ∎

A.2 Proof of Theorem 10

The proof consists in approximating the activation σ_r using Assumption 1.2 on σ. Since σ_r is $(2\pi r)$-Lipschitz, we obtain that there exists, for any $r, Q > 0$, $\alpha_k, \beta_k \in \mathbb{R}$ such that over the interval $[-Q, Q]$ it holds

$$\sup_{|t|\leq Q} \left| \sigma_r(t) - \sum_{k=1}^{N} \alpha_k \sigma(t-\beta_k) \right| \leq \frac{2Qr}{N}$$

as well as

$$\left| \sum_{k=1}^{N} \alpha_k \sigma(t-\beta_k) \right| \leq 1 + 2Qr/N \quad \text{for } t \in \mathbb{R}\ .$$

Let $f_N \in \mathcal{F}_N^\sigma$ be defined as

$$f_N(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \sigma\big(r_d\big(\mathbf{v}_d^T\mathbf{x} + \mathbf{w}_d^T\mathbf{x}_+\big) - \beta_k\big)$$

Now, let $\gamma_d = \|\mathbf{v}_d\|_1 + \|\mathbf{w_d}\|_1$ and $\tilde{Q}_d = \frac{Q_d}{\gamma_d}$, so that by definition when $\|\mathbf{x}\|_\infty \leq \tilde{Q}_d$ it holds that

$$|\mathbf{v}_d^T\mathbf{x} + \mathbf{w}_d^T\mathbf{x}_+| \leq Q_d\ .$$

The approximation error can be decomposed as follows:

$$\begin{aligned}
&\int_{\mathbb{R}^d} (f_{r_d,\mathbf{w}_d,\mathbf{v}_d}(\mathbf{x}) - f_N(\mathbf{x}))^2 \varphi(\mathbf{x})^2 \, d\mathbf{x} = \\
&= \int_{\|\mathbf{x}\|_\infty \le \tilde{Q}_d} (f_{r_d,\mathbf{w}_d,\mathbf{v}_d}(\mathbf{x}) - f_N(\mathbf{x}))^2 \varphi(\mathbf{x})^2 \, d\mathbf{x} + \int_{\|\mathbf{x}\|_\infty > \tilde{Q}_d} (f_{r_d,\mathbf{w}_d,\mathbf{v}_d}(\mathbf{x}) - f_N(\mathbf{x}))^2 \varphi(\mathbf{x})^2 \, d\mathbf{x} \\
&\qquad \le \frac{4Q_d^2 r_d^2}{N^2} \|\varphi \cdot \mathbb{1}_{B^d_{\tilde{Q}_d,\infty}}\|_2^2 + 4\left(1 + \frac{Q_d r_d}{N}\right)^2 (\|\varphi\|_2^2 - \|\varphi \cdot \mathbb{1}_{B^d_{\tilde{Q}_d,\infty}}\|_2^2) \\
&\qquad\qquad \le \frac{4\tilde{Q}_d^2 \gamma_d^2 r_d^2}{N^2} \|\varphi\|_2^2 + 4\left(1 + \frac{Q_d r_d}{N}\right)^2 \left(1 - (1 - \alpha\, \tilde{Q}_d^{-1})^d\right) \\
&\qquad\qquad\qquad\qquad \le \|\varphi\|_2^2 \left(\frac{4\tilde{Q}_d^2 \gamma_d^2 r_d^2}{N^2} + 16\alpha d \tilde{Q}_d^{-1}\right) \, ,
\end{aligned}$$

since $|\psi(x)|^2 \le \alpha |x|^{-2}/2$ for some $\alpha > 0$, as long as $\tilde{Q}_d > \alpha$ and $N > r_d \tilde{Q}_d$. Optimizing this upper bound with respect to $\tilde{Q}_d$ gives

$$\tilde{Q}_d = \left(2\alpha d \frac{N^2}{r_d^2 \gamma_d^2}\right)^{1/3} ,$$

which results in

$$\|f_{r,\mathbf{w},\mathbf{v}} - f\|_\varphi^2 \lesssim \left(\frac{d\gamma_d r_d}{N}\right)^{2/3} ,$$

as long as $N > \alpha r_d \gamma_d$. This concludes the proof.

Appendix B. Proofs of $\mathrm{poly}(d)$ upper bounds

B.1 Proof of Lemma 3

We show this for the case $L(f^{(d)}) = 2$, but the proof it is analogous for the other cases. The function $f^{(d)}$ has the form

$$f^{(d)}(\mathbf{x}) = \boldsymbol{\gamma}_d^T \boldsymbol{\rho}_2(\mathbf{W}_d \boldsymbol{\rho}_1(\mathbf{U}_d \mathbf{x}))$$

where $\boldsymbol{\rho}_1^{(d)}, \boldsymbol{\rho}_2^{(d)}$ are component-wise activations satisfying Assumption 1, and $\boldsymbol{\gamma}_d \in \mathbb{R}^{q_d}$, $\mathbf{W} \in \mathbb{R}^{q_d \times p_d}$, $\mathbf{U} \in \mathbb{R}^{p_d \times d}$, with

$$p_d, q_d, \|\boldsymbol{\gamma}\|_\infty, \|\mathbf{W}\|_{F,\infty}, \|\mathbf{U}\|_{F,\infty} \le \mathrm{poly}(d) \, .$$

Thanks to Assumption 1.2, there exists $\mathbf{A} \in \mathbb{R}^{N p_d \times d}$, $\mathbf{B} \in \mathbb{R}^{p_d \times N p_d}$, $\mathbf{c} \in \mathbb{R}^{N p_d}$ such that

$$\sup_{\mathbf{x} \in K} \left|\boldsymbol{\gamma}^T \boldsymbol{\rho}_2(\mathbf{W}\boldsymbol{\rho}_1(\mathbf{U}\mathbf{x})) - \boldsymbol{\gamma}^T \boldsymbol{\rho}_2(\mathbf{W}\mathbf{B}\, \boldsymbol{\sigma}(\mathbf{A}\mathbf{x} + \mathbf{c}))\right| \le \frac{\epsilon}{2}$$

and

$$N, \|\mathbf{c}\|_\infty, \|\mathbf{B}\|_{F,\infty}, \|\mathbf{A}\|_{F,\infty} \le \epsilon^{-1} \cdot \mathrm{poly}(d) \, .$$

Let $K_1 = \{\mathbf{B}\boldsymbol{\sigma}(\mathbf{A}\mathbf{x}+\mathbf{c}) : \mathbf{x} \in K\}$; it holds $\mathrm{diam}(K_1) \leq \mathrm{poly}(d)$. Similarly as before, we get that there exists $\mathbf{D} \in \mathbb{R}^{Mq_d \times p_d}$, $\mathbf{E} \in \mathbb{R}^{q_d \times Mq_d}$, $\mathbf{f} \in \mathbb{R}^{Mq_d}$ such that

$$\sup_{\mathbf{y}\in K_1} \left|\boldsymbol{\gamma}^T \boldsymbol{\rho}_2(\mathbf{W}\mathbf{y}) - \boldsymbol{\gamma}^T \mathbf{E}\, \boldsymbol{\sigma}(\mathbf{D}\mathbf{y}+\mathbf{f})\right| \leq \frac{\epsilon}{2}$$

and

$$M, \|\mathbf{f}\|_\infty, \|\mathbf{E}\|_{F,\infty}, \|\mathbf{D}\|_{F,\infty} \leq \epsilon^{-1} \cdot \mathrm{poly}(d) \ .$$

By calling $\tilde{\boldsymbol{\gamma}} = \mathbf{E}^T\boldsymbol{\gamma}$, $\tilde{\mathbf{W}} = \mathbf{DWB}$ and $\tilde{\mathbf{U}} = \mathbf{UA}$, we get that

$$g^\sigma(\mathbf{x}) \doteq \tilde{\boldsymbol{\gamma}}^T \boldsymbol{\sigma}(\tilde{\mathbf{W}}\boldsymbol{\sigma}(\tilde{\mathbf{U}}\mathbf{x}+\mathbf{c})+\mathbf{f})$$

satisfies the statement of the theorem.

B.2 Preliminary lemmas

The first lemma is a known results in approximation theory.

Lemma 29 *(Jackson's Theorem, Theorem 1.4 in (Rivlin, 1981)) Let $f : [a,b] \to \mathbb{R}$ with modulus of continuity ω. Then there exists a polynomial $p_n(t) = \sum_{k=0}^n p_k t^k$, $p_k \in \mathbb{R}$, such that*

$$\sup_{t\in[-r,r]} |f(t) - p_n(t)| \leq 6\,\omega\left(\frac{b-a}{2n}\right) \ .$$

The next lemma yields a worst approximation rate but allows us to control the coefficients of the polynomial. It is a small modification of Lemma 4 in (Safran et al., 2019).

Lemma 30 *Let $f : [-r,r] \to \mathbb{R}$ $(1,\alpha)$-Holder. Then for any $\epsilon > 0$ there exists a polynomial $p_n(t) = \sum_{k=0}^n r_k t^k$, $r_k \in \mathbb{R}$, of degree $n = \left\lceil \frac{4^{\frac{1}{\alpha}} r^\alpha}{\epsilon^{1+\frac{2}{\alpha}}} \right\rceil$ such that*

$$\sup_{t\in[-r,r]} |f(t) - p_n(t)| \leq \epsilon \ .$$

Moreover, p_n can be chosen such that $|r_k| \leq 2^n r^{\alpha-k}$, $k \in [n]$, and $|r_0| \leq r^\alpha + |f(0)|$.

Proof [Proof] Notice that we can assume $f(0) = 0$ without loss of generality. Define $g(t) = f(r(2t-1))$ for $t \in [0,1]$ and notice that g is $((2r)^\alpha, \alpha)$-Holder. Also, define the n Bernstein polynomial $b_{n,i}$, $i \in [0,n]$, as

$$b_{n,i}(t) = \binom{n}{i} t^i (1-t)^{n-i}$$

for $t \in [0,1]$. Notice that they form a partition of unity. We define

$$g_n(t) = \sum_{i=0}^n g\left(\frac{i}{n}\right) b_{n,i}(t) \ .$$

We have that

$$\begin{aligned}|g_n(t) - g(t)| &\leq \sum_{i=0}^{n} b_{n,i}(t)\left|g(t) - g\left(\frac{i}{n}\right)\right| \\ &= \sum_{i\,:\,\left|\frac{i}{n}-t\right|<\epsilon} b_{n,i}(t)\left|g(t) - g\left(\frac{i}{n}\right)\right| + \sum_{i\,:\,\left|\frac{i}{n}-t\right|\geq\epsilon} b_{n,i}(t)\left|g(t) - g\left(\frac{i}{n}\right)\right| \\ &\leq \epsilon^\alpha + 2r^\alpha \sum_{i\,:\,\left|\frac{i}{n}-t\right|\geq\epsilon} b_{n,i}(t) \leq \epsilon^\alpha + \frac{r^\alpha}{2n\epsilon^2}\ .\end{aligned}$$

In particular $\frac{r^\alpha}{2n\epsilon^2} \leq \epsilon^\alpha$ if

$$n \geq \frac{r^\alpha}{2\epsilon^{2+\alpha}}\ .$$

If we define $p_n(t) = g_n\left(\frac{t}{2r} + \frac{1}{2}\right)$, then we have that

$$\sup_{x\in[-r,r]} |f(t) - p_n(t)| \leq \epsilon$$

if

$$n \geq \frac{4^{\frac{1}{\alpha}} r^\alpha}{\epsilon^{1+\frac{2}{\alpha}}}\ .$$

Finally, we want to upper bound the coefficients of p_n. Notice that we have

$$p_n(t) = (2r)^{-n} \sum_{i=0}^{n} \binom{n}{i} g\left(\frac{i}{n}\right)(t+r)^i(t-r)^{n-i}\ .$$

It follows that the coefficients of p_n can be bounded by those of

$$(2r)^{-n} \sum_{i=0}^{n} \binom{n}{i}\left|g\left(\frac{i}{n}\right)\right|(t+r)^n \leq r^{\alpha-n}(t+r)^n\ .$$

Let r_k the k-th coefficients of $r^{\alpha-n}(t+r)^n$. Then

$$r_k = r^{\alpha-n}\binom{n}{k} r^{n-k} \leq 2^n r^{\alpha-k}\ .$$

This concludes the proof. ∎

B.3 Approximation by shallow Fourier neural networks

We start by reporting a known result ((Burkill, 1959), Theorem 18).

Lemma 31 *Let $g : [-\pi, \pi] \to \mathbb{R}$ 2π-periodic with modulus of continuity ω. Then there exists a trigonometric polynomial $q_n(t) = \sum_{k=-n}^{n} b_k e^{ikt}$, $b_k \in \mathbb{C}$, with real values (i.e. $q_n(t) \in \mathbb{R}$ for all $t \in [-\pi, \pi]$), such that*

$$\sup_{t\in[-\pi,\pi]} |g(t) - q_n(t)| \leq \frac{2}{\pi}\omega\left(\frac{2}{n}\right)\left[2 + \omega(\pi) - \log\omega\left(\frac{2}{n}\right)\right]\ .$$

Moreover, it holds that

$$|b_k| \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |g(t)|\, dt \ .$$

Proof [Proof] The polinomyal q_n is given by the Fejer sum of the Fourier series of g, that is

$$q_n(t) = \frac{1}{n}\sum_{j=0}^{n-1}\sum_{k=-j}^{j} \hat{g}_k e^{ikt} = \sum_{k=-(n-1)}^{n-1} \frac{n-|k|}{n}\hat{g}_k e^{ikt}$$

where

$$\hat{g}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(t)e^{-ikt}\, dt \ .$$

The proof of the upper bound can be found in (Burkill, 1959), Theorem 18. Finally, notice that q_n is real-valued since

$$\hat{g}_k e^{ikt} + \hat{g}_{-k}e^{-ikt} = 2\mathrm{Re}\Big(\hat{g}_k e^{ikt}\Big)$$

because $\hat{g}_{-k} = \overline{\hat{g}_k}$ since g takes values in $\mathbb{R}$. ■

The above result immediately implies a convergence rate for univariate approximation by shallow Fourier networks (that is, with activation $\sigma_1(t) = e^{2\pi i t}$).

Lemma 32 *Let $f : [-r, r] \to \mathbb{R}$ be L-Lipschitz. Then there exists a real-valued Fourier shallow network $q_n(t) = \sum_{k=-n}^{n} b_k e^{iw_k t}$, $b_k \in \mathbb{C}$, $w_k \in \mathbb{R}$, such that*

$$\sup_{x\in[-r,r]} |f(x) - q_n(x)| \leq 3\big(1 + 2L^2r^2\big)\frac{\log n}{n}$$

for any $n \geq 2$. Moreover q_n can be chosen such that $|w_k| \leq \frac{\pi|k|}{r}$ and $|b_k| \leq \|f\|_\infty$ for any $k \in [-n, n]$.

Proof [Proof] Assume, w.l.o.g., that $f(r) \leq f(-r)$ (otherwise we can consider $f(-x)$ in place of $f(x)$). First, we want to transform f into a 2-pi periodic function on $[-\pi, \pi]$. To do this we consider $\tilde{g}$ defined as

$$\tilde{g}(x) = \begin{cases} L(x+r) + f(-r) & \text{if } x \in \left[-r - \frac{c}{2L}, -r\right] \\ f(x) & \text{if } x \in [-r, r] \\ L(x-r) + f(r) & \text{if } x \in \left[r, r + \frac{c}{2L}\right] \end{cases}$$

where $c = f(-r) - f(r)$. Notice that $\tilde{g}$ is L-Lipschitz and $2\big(r + \frac{c}{2L}\big)$-periodic. Finally, let $g : [-\pi, \pi] \to \mathbb{R}$ defined as

$$g(x) = \tilde{g}\left(\frac{2Lr + c}{2L\pi}x\right) \ .$$

We have that g is 2π-periodic and ℓ-Lipschitz for

$$\ell = \frac{2Lr + c}{2\pi} \leq \frac{2Lr}{\pi} \ .$$

Therefore, we can apply Lemma 31 to g. This gives us a (real-valued) trigonometric polynomial $r_n(t) = \sum_{=-n}^{n} b_k e^{ikt}$ such that

$$\begin{aligned}\sup_{x\in[-\pi,\pi]} |g(x) - r_n(x)| &\leq \frac{4\ell}{\pi n}\left[2 + \ell\pi - \log\frac{2\ell}{n}\right] \\ &\leq 3\big(1 + 2L^2r^2\big)\frac{\log n}{n}\end{aligned}$$

for $n \geq 2$. Since

$$\sup_{x\in[-r,r]}\left|f(x) - r_n\left(\frac{L}{\ell}x\right)\right| \leq \sup_{x\in\left[-r-\frac{c}{2L},r+\frac{c}{2L}\right]}\left|\tilde{g}(x) - r_n\left(\frac{L}{\ell}x\right)\right| = \sup_{x\in[-\pi,\pi]}|g(x) - r_n(x)|$$

the thesis follows. ■

To conclude we make some remarks about shallow Fourier networks. Note that a generic shallow Fourier network f_N with N units can be represented as

$$f(\mathbf{x}) = \sum_{k=1}^{N} u_k e^{i\mathbf{w}_k^T\mathbf{x}}\,. \tag{35}$$

Indeed we have that

$$\sum_{k=1}^{N} u_k e^{i(\mathbf{w}_k^T\mathbf{x}+b_k)} + b = \sum_{k=1}^{N}\Big(u_k e^{ib_k}\Big)e^{i\mathbf{w}_k^T\mathbf{x}} + b\cdot e^{i\mathbf{0}^T\mathbf{x}}$$

for any b, $b_k \in \mathbb{C}$. Let $\mathcal{F}_N^f$ be the space of networks as in equation (35). Notice that a universal approximation theorem holds for shallow Fourier networks as well. This is because the universal approximation theorem holds for shallow networks with activation $\sigma(t) = \cos(t)$ and since $\cos(t) = \big(e^{it} + e^{-it}\big)/2$, the thesis follows. Finally, the following lemma will be used in the proof of Theorem 11.

Lemma 33 *If f is a (real-valued) shallow Fourier neural network, then so is f^k, for k non-negative integer. Moreover, if f has n units, then the number of units of f^k is upper bounded by*

$$\binom{n+k-1}{k}\,.$$

Proof [Proof] Let $f(\mathbf{x}) = \sum_{j=1}^{n} u_j e^{i\mathbf{w}_j^T\mathbf{x}}$ be a shallow Fourier neural network. Then, by the multinomial formula, we have that

$$\begin{aligned}f^k(\mathbf{x}) = \left(\sum_{j=1}^{n} u_j e^{i\mathbf{w}_j^T\mathbf{x}}\right)^k &= \sum_{p_1+\cdots+p_n=k}\binom{k}{p_1,\ldots,p_n}\prod_{j=1}^{n}\Big(u_j^{p_j}\Big(e^{i\mathbf{w}_j^T\mathbf{x}}\Big)^{p_j}\Big) \\ &= \sum_{p_1+\cdots+p_n=k}\binom{k}{p_1,\ldots,p_n}\left(\prod_{j=1}^{n}u_j^{p_j}\right)e^{i\left(\sum_{j=1}^{n}p_j\mathbf{w}_j\right)^T\mathbf{x}}\,.\end{aligned}$$

Clearly, if f is real-valued, so is f^k. Finally notice that by the formula above, the number of units of f^k is upper bounded by $|\{(p_1,\ldots,p_n) : p_1+\cdots+p_n = k\}|$. ■

B.4 poly(d) upper bounds for two-hidden-layers networks

Consider a two-hidden-layers neural network f defined as

$$f : \mathbf{x} \in \mathbb{R}^d \mapsto \boldsymbol{\gamma}^T \mathbf{g}\big(\mathbf{W}^T \mathbf{h}\big(\mathbf{U}^T \mathbf{x}\big)\big) \in \mathbb{C} \ ,$$

where $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^p$ and $\mathbf{g} : \mathbb{R}^o \to \mathbb{R}^o$ are, respectively, component-wise 1-Lipschitz and $(1, \alpha)$-Holder activation functions, and $\mathbf{U} \in \mathbb{R}^{d \times p}$, $\mathbf{W} \in \mathbb{R}^{p \times o}$, $\boldsymbol{\gamma} \in \mathbb{C}^o$. We wish to approximate f with a one-hidden-layer neural network with a given activation σ satisfying Assumption 1.2, for some constant $\nu_\sigma > 0$. We start by proving a result for approximation by shallow Fourier networks at a poly(d) rate.

Proposition 34 *Let $K \subset \mathbb{R}^d$ be a compact set. There exist $f_N \in \mathcal{F}_N^f$ such that*

$$\left\| f - f_N^f \right\|_{K,\infty} \leq \epsilon$$

with

$$f_N^f(\mathbf{x}) = \sum_{\nu=1}^{N} b_\nu e^{i \mathbf{v}_\nu^T \mathbf{x}} \ ,$$

for

$$N = (2np + 1)^m$$

with

$$n = \left\lceil \frac{9 \cdot 4^{\frac{1}{\alpha}} \|\boldsymbol{\gamma}\|_1^2 \|\mathbf{W}\|_\infty^2 (1 + 2C^2)^2}{\epsilon^{\frac{2}{\alpha}}} \right\rceil \quad \textit{and} \quad m = \left\lceil \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1 + \frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} \left(\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \right)^\alpha \right\rceil \ ,$$

where we denoted

$$C = \sup_{x \in K} \|\mathbf{U}^T \mathbf{x}\|_\infty \quad \textit{and} \quad M = \sup_{x \in K} \left\| \mathbf{W}^T h\big(\mathbf{U}^T \mathbf{x}\big) \right\|_\infty \ .$$

Moreover f_N^f can be chosen such that it holds

$$\sup_{\mathbf{x} \in K} \left| \mathbf{v}_\nu^T \mathbf{x} \right| \leq \pi m n \quad \textit{and} \quad |b_\nu| \leq 2\|\boldsymbol{\gamma}\|_1 \left[1 + \left(\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \right)^\alpha \right] (4npH \|\mathbf{W}\|_{F,\infty})^m \quad (36)$$

where $H = \sup_{\mathbf{x} \in [-C,C]^d} \|h(\mathbf{x})\|_\infty$.

Proof [Proof] Let q_n^j given by Lemma 32 to approximate h_j over $[-C, C]$ and

$$q_k^{(n)}(\mathbf{x}) = \sum_{j=1}^{p} w_{k,j} q_n^j(\mathbf{u}_j^T \mathbf{x})$$

for $k \in [o]$. We have that

$$\begin{aligned} \left| q_k^{(n)}(\mathbf{x}) - \mathbf{w}_k^T h\big(\mathbf{U}^T \mathbf{x}\big) \right| &\leq \sum_{j=1}^{p} |w_{k,j}| \left| q_n^j(\mathbf{u}_j^T \mathbf{x}) - h_j(\mathbf{u}_j^T \mathbf{x}) \right| \\ &\leq 3\|\mathbf{W}\|_\infty \big(1 + 2C^2\big) \frac{\log n}{n} \doteq \|\mathbf{W}\|_\infty (1 + 2C^2) \epsilon_n \end{aligned}$$

for $\mathbf{x} \in K$. It holds that $q_k^{(n)}$ is a real-valued shallow Fourier network with $(2n-1)p$ terms and first layers weights given by $\frac{\pi k}{C}\mathbf{u}_j$ for $k \in [-(n-1), n-1]$. Moreover, it holds that

$$\left|q_k^{(n)}(\mathbf{x})\right| \leq \left|q_k^{(n)}(\mathbf{x}) - \mathbf{w}_k^T h\big(\mathbf{U}^T\mathbf{x}\big)\right| + \left|\mathbf{w}_k^T h\big(\mathbf{U}^T\mathbf{x}\big)\right| \leq \|\mathbf{W}\|_\infty(1+2C^2)\epsilon_n + M \doteq L \ .$$

Let $p_m^k(t) = \sum_{h=0}^m \beta_h^k t^h$ given by Corollary 3 to approximate g_k over the interval $[-L, L]$ and ϵ_m the relative error. Let then

$$f_{n,m}(\mathbf{x}) = \sum_{k=1}^o \gamma_k p_m^k(q_k^n(\mathbf{x})) \ .$$

It holds that

$$\begin{aligned}
|f(\mathbf{x}) - f_{n,m}(\mathbf{x})| &\leq \sum_{k=1}^o |\gamma_k| \left|g_k(\mathbf{w}_k^T h(\mathbf{U}^T\mathbf{x})) - p_m^k(q_k^{(n)}(\mathbf{x}))\right| \\
&\leq \sum_{k=1}^o |\gamma_k| \left|g_k(\mathbf{w}_k^T h(\mathbf{U}^T\mathbf{x})) - g_k(q_k^n(\mathbf{x}))\right| + \sum_{k=1}^o |\gamma_k| \left|g_k(q_k^{(n)}(\mathbf{x})) - p_m^k(q_k^{(n)}(\mathbf{x}))\right| \\
&\leq \|\boldsymbol{\gamma}\|_1 \sup_{k\in[o]} \left|\mathbf{w}_k^T h(\mathbf{U}^T\mathbf{x}) - q_k^{(n)}(\mathbf{x})\right|^\alpha + \|\boldsymbol{\gamma}\|_1 \epsilon_m \\
&\leq \|\boldsymbol{\gamma}\|_1 \|\mathbf{W}\|_\infty^\alpha (1+2C^2)^\alpha \epsilon_n^\alpha + \|\boldsymbol{\gamma}\|_1 \epsilon_m \ .
\end{aligned}$$

It holds that

$$\|\boldsymbol{\gamma}\|_1 \|\mathbf{W}\|_\infty^\alpha (1+2C^2)^\alpha \epsilon_n^\alpha \leq \frac{\epsilon}{2}$$

as long as

$$n \geq \frac{9 \cdot 4^{\frac{1}{\alpha}} \|\boldsymbol{\gamma}\|_1^2 \|\mathbf{W}\|_\infty^2 \big(1+2C^2\big)^2}{\epsilon^{\frac{2}{\alpha}}} \ . \tag{37}$$

Similarly

$$\|\boldsymbol{\gamma}\|_1 \epsilon_m \leq \frac{\epsilon}{2}$$

as long as

$$m \geq L\left(\frac{12\|\boldsymbol{\gamma}\|_1}{\epsilon}\right)^{\frac{1}{\alpha}} = \left(\frac{12\|\boldsymbol{\gamma}\|_1}{\epsilon}\right)^{\frac{1}{\alpha}} \big[\|\mathbf{W}\|_\infty(1+2C^2)\epsilon_n + M\big] \ .$$

Moreover, by Lemma 30, $p_m^k(t) = \sum_{h=0}^m \beta_h^k t^h$ can be chosen with

$$m \geq \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} L^\alpha = \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} \big[\|\mathbf{W}\|_\infty(1+2C^2)\epsilon_n + M\big]^\alpha$$

such that its coefficients β_h^k, $k \in [m]$, are bounded by

$$\begin{aligned}
|\beta_k| &\leq \max\Big\{2^m L^{\alpha-k}, L^\alpha + |g(0)|\Big\} \leq 2^m(1+L^\alpha) + |g(0)| \\
&= 2^m\big(1 + \big[\|\mathbf{W}\|_\infty(1+2C^2)\epsilon_n + M\big]^\alpha\big) + |g(0)| \ .
\end{aligned}$$

Notice that we can assume $g(0) = 0$ without loss of generality. Therefore

$$\sup_{x\in K}|f(\mathbf{x}) - f_{n,m}(\mathbf{x})| \leq \epsilon$$

as long as (37) holds and

$$m \geq \left(\frac{12\|\boldsymbol{\gamma}\|_1}{\epsilon}\right)^{\frac{1}{\alpha}}\left[\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1}\right)^{\frac{1}{\alpha}} + M\right] = 6^{\frac{1}{\alpha}}\left(1 + M\left(\frac{2\|\boldsymbol{\gamma}\|_1}{\epsilon}\right)^{\frac{1}{\alpha}}\right). \tag{38}$$

If we further assume that

$$m \geq \frac{2\cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}}\|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}}\left[\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1}\right)^{\frac{1}{\alpha}} + M\right]^{\alpha}$$

we can also assume that

$$\left|\beta_h^k\right| \leq 2^{1+\frac{2\cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}}\|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}}\left[\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1}\right)^{\frac{1}{\alpha}}+M\right]^{\alpha}}\left(1 + \left[\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1}\right)^{\frac{1}{\alpha}} + M\right]^{\alpha}\right)$$

for $k \in [m]$. Finally, notice that, by Lemma 33, $f_{n,m}$ is a shallow Fourier neural network with number of units upper bounded by

$$\begin{aligned} N &= \sum_{k=0}^{m}\binom{(2n-1)p+k-1}{k} = \binom{(2n-1)p+m}{m} \\ &= \frac{1}{m!}((2n-1)p+k+m)\cdots((2n-1)p+1) \\ &\leq ((2n-1)p+1)^m\ . \end{aligned}$$

Therefore, it holds that

$$\inf_{f_N\in\mathcal{F}_N^f}\sup_{\mathbf{x}\in K}|f(\mathbf{x}) - f_N(\mathbf{x})| \leq \epsilon$$

as long as

$$N \geq (2np+1)^m$$

with n and m given by (37) and (38) respectively. Finally, notice that the first layer weights of $f_{n,m}$ are given by

$$\sum_{j=1}^{p}\sum_{k=-(n-1)}^{n-1} s_{k,j}\frac{\pi k}{C}u_j$$

over all non-negative integers $s_{k,j}$ such that $\sum_{j=1}^{p}\sum_{k=-(n-1)}^{n-1} s_{k,j} \leq m$. Therefore, if

$$f_{n,m}(\mathbf{x}) = \sum_{\nu=1}^{N} b_\nu e^{i\mathbf{v}_\nu^T\mathbf{x}}\ ,$$

then

$$\left|\mathbf{v}_\nu^T\mathbf{x}\right| \leq m\frac{\pi(n-1)}{C}\max_{j\in[p]}|\mathbf{u}_j^T\mathbf{x}| \leq mn\pi\ .$$

On the other hand, the coefficients b_k have the form

$$b_\nu = \binom{h}{s} \sum_{k=1}^{o} \gamma_k \beta_h^k \big(w_{k,j}(q_n^j)_l\big)^{s_{l,j}}$$

for all non-negative integers $s = (s_{l,j})_{l,j}$ such that $\sum_{j=1}^{p} \sum_{l=-(n-1)}^{n-1} s_{l,j} = h \leq m$, where $(q_n^j)_l$ denotes the l-th coefficients of q_n^j. By Lemma 31, we know that

$$\big|(q_n^j)_l\big| \leq \sup_{t\in[-C,C]} |h_j(t)| \ .$$

Therefore

$$\begin{aligned} |b_\nu| &\leq ((2n-1)p)^h \sup_{t\in[-C,C]} |h_j(t)|^{s_{l,j}} \sum_{k=1}^{o} |\gamma_k||\beta_h^k||w_{k,j}|^{s_{l,j}} \\ &\leq [(2n-1)p\, H\, \|\mathbf{W}\|_{F,\infty}]^m \|\boldsymbol{\gamma}\|_1 \|\boldsymbol{\beta}\|_{F,\infty} \ . \end{aligned}$$

This concludes the proof. ■

We can now conclude with a detailed version of Theorem 11.

Theorem 35 *Let K be a compact set and*

$$C = \sup_{\mathbf{x}\in K} \|\mathbf{U}^T\mathbf{x}\|_\infty \,, \quad M = \sup_{\mathbf{x}\in K} \big\|\mathbf{W}^T\mathbf{h}\big(\mathbf{U}^T\mathbf{x}\big)\big\|_\infty \quad \textit{and} \quad H = \sup_{\mathbf{x}\in[-C,C]^d} \|\mathbf{h}(\mathbf{x})\|_\infty \ .$$

It holds that

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f(\mathbf{x}) - f_N^\sigma(\mathbf{x})\|_{K,\infty} \leq \epsilon$$

for some

$$N \leq \frac{16\pi\nu_\sigma}{\epsilon} \|\boldsymbol{\gamma}\|_1 mn(4np+1)^{2m} (H\|\mathbf{W}\|_{F,\infty})^m \left[1 + \left(\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1}\right)^{\frac{1}{\alpha}} + M\right)^\alpha\right] ,$$

where

$$n = \frac{9\cdot 4^{\frac{1}{\alpha}} \|\boldsymbol{\gamma}\|_1^2 \|\mathbf{W}\|_\infty^2 (1+2C^2)^2}{\epsilon^{\frac{2}{\alpha}}} \quad \textit{and} \quad m = \frac{2\cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} \left(\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1}\right)^{\frac{1}{\alpha}} + M\right)^\alpha .$$

Moreover, it is possible to choose f_N^σ such that $m_\infty(f_N^\sigma) \leq (1+N^2)$.

Proof [Proof] Let f_N given by Proposition 34 such that

$$\sup_{\mathbf{x}\in K} |f(\mathbf{x}) - f_N(\mathbf{x})| \leq \frac{\epsilon}{2} \ .$$

We know that

$$f_N(\mathbf{x}) = \sum_{k=1}^{N} b_k e^{i\mathbf{v}_k^T\mathbf{x}} = f_N^c(\mathbf{x}) + i f_N^s(\mathbf{x})$$

where

$$f_N^c(\mathbf{x}) = \sum_{k=1}^{N} b_k \cos(\mathbf{v}_k^T \mathbf{x}) \quad \text{and} \quad f_N^s(\mathbf{x}) = \sum_{k=1}^{N} b_k \sin(\mathbf{v}_k^T \mathbf{x})$$

and $|b_k| \leq B$ and $\left|\mathbf{v}_k^T \mathbf{x}\right| \leq V$ for $\mathbf{x} \in K$, where B and V are given by (36). Using the assumption on σ, we know that, for each $k \in [N]$, there exist shallow networks f_k^c and f_k^s with activation σ and number of units

$$n \leq c_\sigma \frac{4VBN}{\epsilon}$$

such that

$$\sup_{\mathbf{x}\in K} \left| f_k^c(\mathbf{x}) - \cos(\mathbf{v}_k^T \mathbf{x}) \right| \leq \frac{\epsilon}{4NB} \quad \text{and} \quad \sup_{\mathbf{x}\in K} \left| f_k^s(\mathbf{x}) - \sin(\mathbf{v}_k^T \mathbf{x}) \right| \leq \frac{\epsilon}{4NB} \ .$$

Letting $f_{\mathcal{N}}(\mathbf{x}) = \sum_{k=1}^{N} b_k f_k^c(\mathbf{x}) + i \sum_{k=1}^{N} b_k f_k^s(\mathbf{x})$ it holds that

$$\begin{aligned}
\sup_{\mathbf{x}\in K} |f_{\mathcal{N}}(\mathbf{x}) - f_N(\mathbf{x})| &\leq \sup_{\mathbf{x}\in K} \left| \sum_{k=1}^{N} b_k \left(f_k^c(\mathbf{x}) - \cos(\mathbf{w}_k^T \mathbf{x}) \right) \right| + \sup_{\mathbf{x}\in K} \left| \sum_{k=1}^{N} b_k \left(f_k^s(\mathbf{x}) - \sin(\mathbf{w}_k^T \mathbf{x}) \right) \right| \\
&\leq \sum_{k=1}^{N} |b_k| \sup_{\mathbf{x}\in K} \left| f_k^c(\mathbf{x}) - \cos(\mathbf{w}_k^T \mathbf{x}) \right| + \sum_{k=1}^{N} |b_k| \sup_{\mathbf{x}\in K} \left| f_k^s(\mathbf{x}) - \sin(\mathbf{w}_k^T \mathbf{x}) \right| \\
&\leq NB \frac{\epsilon}{4NB} + NB \frac{\epsilon}{4NB} = \frac{\epsilon}{2}
\end{aligned}$$

which implies that

$$\sup_{\mathbf{x}\in K} |f_{\mathcal{N}}(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon \ .$$

Moreover notice that we can assume that all second layer weights of $f_{\mathcal{N}}$ are real; indeed, if this is not the case, one can replace them by the real part, and upper bound above can only decrease. Finally, we have that the number of units of $f_{\mathcal{N}}$ is given by

$$\mathcal{N} \leq \frac{8c_\sigma}{\epsilon} \cdot V \cdot B \cdot N \ .$$

Applying Proposition 34 concludes the proof. ■

B.5 Proofs of special cases

B.5.1 Radial functions

Let $f(\mathbf{x}) = \varphi(\|\mathbf{x}\|)$ with φ 1-Lipschitz. Then it holds that $f(\mathbf{x}) = g(\mathbf{1}^T \mathbf{h}(\mathbf{x}))$ where $g(t) = \varphi(\sqrt{t})$ and $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ is defined as $h_i(\mathbf{x}) = x_i^2$. Clearly, $\sup_{\mathbf{x}\in B_{1,2}^d} \|\mathbf{x}\|_\infty = 1$, $\sup_{\mathbf{x}\in B_{1,2}^d} \left|\mathbf{1}^T \mathbf{h}(\mathbf{x})\right| = \sup_{\mathbf{x}\in B_{1,2}^d} \|\mathbf{x}\|^2 = 1$ and $\sup_{\mathbf{x}\in[-1,1]^d} \|\mathbf{h}(\mathbf{x})\|_\infty = \sup_{x\in[-1,1]} |x|^2 = 1$. Moreover, $\|\mathbf{1}\|_1 = d$ and g is $(1, 1/2)$-Holder. Then, by applying Theorem 35, we get the following.

Corollary 36 (Radial functions) *It holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f_N^\sigma - f\|_{B^d_{1,2},\infty} \leq \epsilon$$

for some

$$N \leq \nu_\sigma \alpha \cdot d^2 \cdot \frac{(4+\epsilon)^2}{\epsilon^{10}} \left(\alpha \frac{d^3}{\epsilon^4} + 1 \right)^{\frac{\alpha}{\epsilon^5}(2+\epsilon)}$$

where $\alpha > 0$ is a numerical constant.

B.5.2 Shallow approximation of piece-wise oscillatory functions

Consider $f_{\mathbf{w},\mathbf{U}} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{i\mathbf{w}^T(\mathbf{U}\mathbf{x})_+}$ for some $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{U} \in \mathbb{R}^{p\times d}$. Then Theorem 35 implies the following.

Corollary 37 (Approximation of (2) by shallow networks) *It holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f_{\mathbf{w},\mathbf{U}} - f_N^\sigma\|_{B^d_{r,p},\infty} \leq \epsilon$$

for some

$$N \leq \frac{\nu_\sigma \beta}{\epsilon^6} \cdot (2 + \epsilon + 2r\|\mathbf{w}\|_1 \|\mathbf{U}\|_{p,\infty})^2 \cdot \left[r\|\mathbf{w}\|_\infty \|\mathbf{U}\|_{p,\infty} \left(\frac{4p\beta}{\epsilon^2} + 1 \right)^2 \right]^{\frac{\alpha}{\epsilon^2}(\epsilon + 2r\|\mathbf{w}\|_1\|\mathbf{U}\|_{p,\infty})}$$

where $\beta = \alpha\|\mathbf{w}\|_1^2 \cdot \left(1 + 2r^2\|\mathbf{U}\|_{p,\infty}^2\right)^2$ and α is a numerical constant.

B.5.3 Approximation bounds under the Gaussian metric

For sake of simplicity in this section we consider approximation bounds for the function of interest

$$f_{\mathbf{w},\mathbf{U}} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{i\mathbf{w}^T(\mathbf{U}\mathbf{x})_+}$$

for some $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_p]^T \in \mathbb{R}^{p\times d}$. Notice that the following results can be naturally extended to any three-layer network target. We are interested in upper bounding the error

$$\inf_{f_N \in \mathcal{F}_N^f} \left(\mathbb{E}|f_{\mathbf{w},\mathbf{U}}(\mathbf{X}) - f_N(\mathbf{X})|^2 \right)^{\frac{1}{2}}$$

where the expectation is taken over $\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. For sake of simplicity of notation, we denote

$$\|f - g\|_{\sigma,2} \doteq \left(\mathbb{E}|f(\mathbf{X}) - g(\mathbf{X})|^2 \right)^{\frac{1}{2}} .$$

It is a well known fact that Gaussian vectors concentrates in a ball of radius $\sqrt{d}$. We recall a quantitative version of this fact in the following.

Lemma 38 *Let $\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ a d-dimensional Gaussian vector. Then it holds that*

$$P\Big\{ \|\mathbf{X}\|_2 \geq \sigma\sqrt{d} + t \Big\} \leq e^{-\frac{t^2}{2\sigma^2}} .$$

Thanks to Proposition 34, the following holds.

Lemma 39 *Let $r > 0$. Then it holds that*

$$\inf_{f_N \in \mathcal{F}_N^f} \|f_N - f_{\mathbf{w},\mathbf{U}}\|_{B^d_{r,2},\infty} \leq \delta \tag{39}$$

as long as

$$N \geq (2np+1)^m$$

where

$$n = \frac{36}{\delta^2}\|\mathbf{w}\|_1^2\big(1 + r^2\|\mathbf{U}\|_{2,\infty}^2\big)^2 \quad \textit{and} \quad m \geq \frac{16}{\delta^3}(\delta + 2r\|\mathbf{w}\|_1\|\mathbf{U}\|_{2,\infty}) \ .$$

Moreover, under the same assumption, we can also assume that the function f_N that satisfies (39) *also satisfies*

$$\|f_N\|_\infty \leq N(2 + \delta + 2r\|\mathbf{w}\|_1\|\mathbf{U}\|_{2,\infty})(4npr\|\mathbf{w}\|_\infty\|\mathbf{U}\|_{2,\infty})^m \ .$$

Thanks to these two lemmas, the following proposition follows.

Proposition 40 *Let $\sigma = d^{-1/2}$ and assume that $\|\mathbf{U}\|_{2,\infty} \leq 1$. Then it holds*

$$\inf_{f_N \in \mathcal{F}_N^f} \|f_N - f_{\mathbf{w},\mathbf{U}}\|_{\sigma,2} \leq \epsilon \tag{40}$$

as long as

$$N \geq \left[Kp\left(1 + \frac{1}{\epsilon^s}\right)(1 + \|\mathbf{w}\|_1^s)\right]^{K\left(1+\left(\frac{\log p}{d}\right)^s\right)\left(1+\frac{1}{\epsilon^s}\right)\left(1+\|\mathbf{w}\|_1^s\right)}$$

where $K > 0$ and $s \geq 1$ are some numerical constant.

Proof [Proof] Let $c = \|\mathbf{w}\|_1$. First, notice that $\|f_{\mathbf{w},\mathbf{U}}\|_\infty = 1$. Let $\chi_r(\mathbf{x}) = \mathbb{1}\{\|\mathbf{x}\|_2 \leq r\}$ and f_N given by Lemma 39 for a certain $\delta > 0$. Then it holds that

$$\begin{aligned}\|f_N - f_{\mathbf{w},\mathbf{U}}\|_{\sigma,2} &\leq \|(f_N - f_{\mathbf{w},\mathbf{U}})(1-\chi_r)\|_{\sigma,2} + \|(f_N - f_{\mathbf{w},\mathbf{U}})\chi_r\|_{\sigma,2} \\ &\leq \|f_N - f_{\mathbf{w},\mathbf{U}}\|_{B^d_{r,2},\infty} + P(\|\mathbf{x}\|_2 > r)(\|f_N\|_\infty + \|f_{\mathbf{w},\mathbf{U}}\|_\infty) \ .\end{aligned}$$

If $r = 1 + t$ for $t > 0$, it follows

$$\|f_N - f_{\mathbf{w},\mathbf{U}}\|_{2,\sigma} \leq \delta + e^{-\frac{dt^2}{2}}(1 + \|f_N\|_\infty)$$

as long as

$$N \geq \left(\frac{72p}{\delta^2}c^2\big(1+r^2\big)^2 + 1\right)^{\frac{1}{\delta^3}(\delta+2rc)} \ .$$

Moreover, one can assume

$$\begin{aligned}\|f_N\|_\infty &\leq (2 + \delta + 2rc)\left(\frac{72p}{\delta^2}c^2\big(1+r^2\big)^2 + 1\right)^{\frac{16}{\delta^3}(\delta+2rc)}\left(144\frac{pr}{\delta^2}c^3\big(1+r^2\big)^2\right)^{\frac{16}{\delta^3}(\delta+2r)} \\ &\leq (2 + \delta + 2r\omega)\Big(144\frac{p}{\delta^2}\omega^3 r(1+r^2)^2 + 1\Big)^{\frac{32}{\delta^3}(\delta+2r\omega)}\end{aligned}$$

where $\omega = \max(1, c)$. Let $\delta = \frac{\epsilon}{2}$. If $t \geq 1$, it holds that

$$\begin{aligned}\|f_N\|_\infty &\leq (4\omega + \epsilon + 2\omega t)\Big(576\frac{p}{\epsilon^2}\omega^3(1+t)\big(1+(1+t)^2\big)^2 + 1\Big)^{\frac{256}{\epsilon^3}(\epsilon+2\omega+2\omega t)} \\ &\leq K(\epsilon + \omega + \omega t)\Big(K\frac{p}{\epsilon^2}\omega^2 t^5 + 1\Big)^{\frac{K}{\epsilon^3}(\epsilon+\omega+\omega t)} \ .\end{aligned}$$

In the equation above above and in the following, K denotes a (large enough) numerical constant. Therefore

$$e^{-\frac{dt^2}{2}}(1 + \|f_N\|_\infty) \leq \frac{\epsilon}{2} \tag{41}$$

as long as

$$\frac{dt^2}{2} - \log\Big(1 + K(\epsilon + \omega + \omega t)\Big(K\frac{p}{\epsilon^2}\omega^2 t^5 + 1\Big)^{\frac{K}{\epsilon^3}(\epsilon+\omega+\omega t)}\Big) + \log\frac{\epsilon}{2} \geq 0 \ .$$

Since $\log(1 + Cs^\alpha) \leq \log(1 + C) + \alpha\log(s)$ if $s \geq 1$, $C > 0$ and $\alpha > 0$, the above is implied by

$$\frac{dt^2}{2} - \log(1 + K(\epsilon + \omega + \omega t)) - \frac{K}{\epsilon^3}(\epsilon + \omega + \omega t)\log\Big(K\frac{p}{\epsilon^2}\omega^2 t^5 + 1\Big) + \log\frac{\epsilon}{2} \geq 0 \ .$$

Since

$$\log(1 + K(\epsilon + \omega + \omega t)) \leq K(\epsilon + \omega + \omega t)$$

and

$$\log\Big(K\frac{p}{\epsilon^2}\omega^2 t^5 + 1\Big) \leq \log\Big(1 + K\frac{p\omega^2}{\epsilon^2}\Big) + 5\log t \leq \log\Big(1 + K\frac{p\omega^2}{\epsilon^2}\Big) + 5\sqrt{t}$$

equation (41) holds if

$$\frac{dt^2}{2} - \alpha - \beta t^{1/2} - \gamma t - \eta t^{3/2} \geq 0$$

where

$$\begin{aligned}\alpha &= K(\epsilon + \omega) + \frac{K}{\epsilon^3}(\epsilon + \omega)\log\Big(1 + K\frac{p\omega^2}{\epsilon^2}\Big) - \log\frac{\epsilon}{2} > 0 \ , \\ \beta &= \frac{K}{\epsilon^3}(\epsilon + \omega) > 0 \ , \\ \gamma &= K\omega t + \frac{K}{\epsilon^3}\omega\log\Big(1 + K\frac{p\omega^2}{\epsilon^2}\Big) > 0 \ , \\ \eta &= \frac{K}{\epsilon^3}\omega t > 0 \ .\end{aligned}$$

It follows that eq. (41) holds if

$$t \geq 1 + 4\Big(\frac{\alpha + \beta + \gamma + \eta}{d}\Big)^2 \ .$$

It follows that the error bound (40) holds as long as

$$N \geq \left(\frac{Kp}{\epsilon^2}(1+c)^2 \left(1 + 4\left(\frac{\alpha+\beta+\gamma+\eta}{d} \right)^2 \right)^4 + 1 \right)^{\frac{K}{\epsilon^3}\left(\epsilon + c\left(1+\left(\frac{\alpha+\beta+\gamma+\eta}{d}\right)^2\right)\right)} .$$

The thesis follows. ■

B.6 Extension to generic L-layers networks

The results presented in the previous section can be generalized to hold for approximating generic multi-layer neural networks. In this section we present an analogous result to Theorem 11 for this more general case. Consider a multi-layer neural network f defined as

$$f : \mathbf{x} \in \mathbb{R}^d \to x^{(L)}(\mathbf{x}) \in \mathbb{C}$$

where $x^{(L)}$ is defined by recursion by $\mathbf{x}^{(0)}(\mathbf{x}) = \mathbf{x}$,

$$\mathbf{x}^{(k)}(\mathbf{x}) = \boldsymbol{\sigma}^{(k)}(\mathbf{A}^{(k)}\mathbf{x}^{(k-1)}(\mathbf{x})) \text{ for } k \in [L] \quad \text{and} \quad x^{(L+1)}(\mathbf{x}) = \left[\mathbf{a}^{(L+1)}\right]^T \mathbf{x}^{(L)}(\mathbf{x}) ,$$

where $\mathbf{A}^{(k)} = [\mathbf{a}_1^{(k)}| \cdots |\mathbf{a}_{d_k}^{(k)}]^T \in \mathbb{R}^{d_k \times d_{k-1}}$ for $k \in [L]$ (with $d_0 = d$), $\mathbf{a}^{(L+1)} \in \mathbb{C}^{d_L}$ and $\boldsymbol{\sigma}^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ are $\frac{1}{6}$-Lipschitz component-wise activation functions and verify $\boldsymbol{\sigma}^k(\mathbf{0}) = \mathbf{0}$ for $k \in [L]$. In the following we also assume that $\|\mathbf{A}^{(k)}\|_\infty \leq 1$ for $k \in [L]$ and $\|\mathbf{a}_{L+1}\|_1 \leq 1$. Note that these assumption can easily be relaxed, but we adopt them here for sake of simplicity.

Proposition 41 *Let f as above. It holds that*

$$\inf_{f_N \in \mathcal{F}_N^f} \|f - f_N\|_{B^d_{1,\infty},\infty} \leq \epsilon$$

as long as

$$N \geq \left(2^L C \left(1 + \frac{1}{\epsilon^2} \right) d_1 \right)^{CL\left(1+\frac{1}{\epsilon}\right)^{L-1}}$$

where C is a numerical constant.

Before proving the above proposition, we prove two preliminary lemmas.

Lemma 42 *Let $\mathcal{W} = \{\mathbf{w}_\ell\}_{\ell \in [K]} \subset \mathbb{R}^d$ and $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^p$ such that h_j is a shallow Fourier neural networks with first layer weights given by $\mathcal{W}$, for all $j \in [p]$. Consider $\mathbf{q} : \mathbb{R}^p \to \mathbb{R}^m$ of the form*

$$\mathbf{q}(\mathbf{x}) = \mathbf{B}\boldsymbol{\sigma}(\mathbf{x})$$

where $\boldsymbol{\sigma} : \mathbb{R}^p \to \mathbb{R}^p$ is a component-wise polynomial activation function of degree at most D and $\mathbf{B} \in \mathbb{C}^{m \times p}$. Then there exists $\mathcal{V} \subset \mathbb{R}^d$ finite such that $\mathbf{f} \doteq \mathbf{q} \circ \mathbf{h}$ is such that f_j is a Fourier neural nets with first layer weights given by $\mathcal{V}$ for each $j \in [p]$ and such that

$$|\mathcal{V}| \leq (2K)^D .$$

Proof [Proof] The functions f_j have the form

$$f_j(\mathbf{x}) = \sum_{k=1}^{p} b_{jk} \sum_{l=0}^{D} \alpha_{k,l} (h_k(\mathbf{x}))^l = \sum_{k=1}^{p} b_{jk} \sum_{l=0}^{D} \alpha_{k,l} \left(\sum_{\nu=1}^{K} \beta_{k,\nu} e^{i \mathbf{w}_\nu^T \mathbf{x}} \right)^l .$$

By Lemma 33, we see that each f_j is a Fourier neural network with the same set of first layer weights of size at most

$$\sum_{l=0}^{D} \binom{K+l-1}{l} = \binom{K+D}{D} \leq (K+1)^D \leq (2K)^D .$$

This concludes the proof. ■

Lemma 43 *Consider the same assumption as Proposition 41. Then, there exists a polynomial*

$$f_{N_1,\ldots,N_L} : \mathbf{x} \in \mathbb{R}^d \to y^{(L+1)}(\mathbf{x}) \in \mathbb{C}$$

given by the recursion $\mathbf{y}^{(0)}(\mathbf{x}) = \mathbf{x}$,

$$\begin{aligned} \mathbf{y}^{(k)}(\mathbf{x}) &= \mathbf{p}_{N_k}^k (\mathbf{A}^{(k)} \mathbf{y}^{(k-1)}(\mathbf{x})) \quad \textit{for } k \in [L] \\ y^{(L+1)}(\mathbf{x}) &= \left[\mathbf{a}^{(L+1)} \right]^T \mathbf{y}^{(L)}(\mathbf{x}) \end{aligned}$$

where $\mathbf{p}_{N_k}^k$ *are component-wise polynomial activation functions of degree* N_k*, such that*

$$\| f - f_{N_1,\ldots,N_L} \|_{B^d_{1,\infty},\infty} \leq \epsilon \tag{42}$$

as long as $N_k \geq \frac{L}{\epsilon} + (L-1)$ *for* $k \in [L]$*. In particular,* f *is a polynomial of degree* $\prod_{k=1}^{L} N_k$*.*

Proof [Proof] We can show this by induction over L. First, consider the case $L = 1$. By Lemma 29, for each $j \in [d_1]$, there exist polynomials $p_{N,j} : \mathbb{R} \to \mathbb{R}$ of degree N which verify

$$\left| p_{N,j}((\mathbf{a}_i^{(1)})^T \mathbf{x}) - \sigma_j^{(1)}((\mathbf{a}_j^{(1)})^T \mathbf{x}) \right| \leq \frac{1}{N}$$

since $\left| (\mathbf{a}_i^{(1)})^T \mathbf{x} \right| \leq 1$ by assumption. Since $\| \mathbf{a}^{(2)} \|_1 \leq 1$, it follows that

$$\left| (\mathbf{a}^{(2)})^T \mathbf{p}_N(\mathbf{A}^{(1)} \mathbf{x}) - (\mathbf{a}^{(2)})^T \boldsymbol{\sigma}^{(1)}(\mathbf{A}^{(1)} \mathbf{x}) \right| \leq \frac{1}{N} .$$

This implies the thesis for the case $L = 1$. Now consider the induction step, that is, assume that, for every $\delta > 0$ and j, there exists a certain $f^j_{N_1,\ldots,N_{L-1}}$ such that

$$\left| x_j^{(L-1)}(\mathbf{x}) - f^j_{N_1,\ldots,N_{L-1}}(\mathbf{x}) \right| \leq \delta$$

as long as $N_k \geq \frac{L-1}{\delta} + (L-2)$ for $k \in [L-1]$. Notice that this implies that

$$\left|(\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x})\right| \leq 1+\delta \ ,$$

where $\mathbf{f}_{N_1,\dots,N_{L-1}} = (f^1_{N_1,\dots,N_{L-1}}, \dots, f^{d_{L-1}}_{N_1,\dots,N_{L-1}})$. Therefore for each $j \in [d_L]$, by Lemma 29, there exist polynomials $p_{N,j}$ of degree N such that

$$\left|p_{N,j}((\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x})) - \sigma_j^{(L)}((\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x}))\right| \leq \frac{1+\delta}{N} \ .$$

Let then $f_{N_1,\dots,N_{L-1},N}$ be defined as

$$f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) = \sum_{j=1}^{N} a_j^{(L+1)} p_{N,j}((\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x})) \ .$$

Since $\|\mathbf{a}^{(L+1)}\|_1 \leq 1$, it holds that

$$\begin{aligned}\left|f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) - f(\mathbf{x})\right| \leq& \left|f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) - \mathbf{a}_{L+1}^T \boldsymbol{\sigma}^{L+1}\big(f_{N_1,\dots,N_{L-1},N}(\mathbf{x})\big)\right| \\ &+ \left|\mathbf{a}_{L+1}^T \boldsymbol{\sigma}^{L+1}\big(f_{N_1,\dots,N_{L-1},N}(\mathbf{x})\big) - f(\mathbf{x})\right| \\ \leq& \frac{1+\delta}{N} + \delta \ .\end{aligned}$$

If $\delta = \frac{L-1}{L}\epsilon$ then equation (42) holds as long as

$$N \geq \frac{1+\frac{L-1}{L}\epsilon}{\frac{\epsilon}{L}} = \frac{L}{\epsilon} + (L-1) \ .$$

This concludes the proof of the lemma. ■

Proof [Proof of Proposition 41] It holds that

$$f(\mathbf{x}) = g(\boldsymbol{\sigma}^{(1)}(\mathbf{A}^{(1)}\mathbf{x}))$$

where g is a $(L-1)$-hidden-layers neural network with input dimension d_1. By Lemma 31, for every $\delta > 0$ and $j \in [d_1]$, there exists Fourier networks $q_{N_1,j}(\mathbf{x})$ with $2N_1 - 1$ units such that

$$\left|\sigma_j^{(1)}((\mathbf{a}_j^{(1)})^T\mathbf{x}) - q_{N_1,j}((\mathbf{a}_j^{(1)})^T\mathbf{x})\right| \leq \frac{C}{\sqrt{N_1}}$$

where $C > 0$ is a numerical constant. Notice that this implies that, for $N_1 \geq 4C^2$, it holds

$$\left\|\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x})\right\|_\infty \leq 1 \ .$$

Now, we can approximate g with a polynomial neural network $g_{N_L,\dots,N_2}$ as given by Lemma 43. In particular, for any $\delta > 0$, there exist $g_{N_L,\dots,N_2}$ such that

$$\sup_{\mathbf{x}\in[-1,1]^d} |g_{N_L,\dots,N_2}(\mathbf{x}) - g(\mathbf{x})| \leq \delta$$

as long as $N_k \geq \frac{L-1}{\delta} + (L-2)$ for $k \in [2, L]$. It follows that

$$\left|g_{N_L,\dots,N_2}(\mathbf{q}_{N_1}(\mathbf{A}^1\mathbf{x})) - f(\mathbf{x})\right| \leq \delta + \frac{C}{\sqrt{N_1}} \ .$$

Let $f_N(\mathbf{x}) = g_{N_L,\dots,N_2}(\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x}))$. By choosing $\delta = \epsilon/2$, it holds that

$$\sup_{\mathbf{x}\in[-1,1]^d} |f_N(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon$$

as long as $N_k \geq 2\frac{L-1}{\epsilon} + (L-2)$ for $k \in [2, L]$ and $N_1 \geq C^2\left(1 + \frac{4}{\epsilon^2}\right)$. We claim that f_N is a Fourier network with at most

$$N = \left(2^L N_1 d_1\right)^{\prod_{k=2}^{L} N_k} \tag{43}$$

units. We can prove this by induction over $L \geq 2$. Remember that $g_{N_L,\dots,N_2}$ is is the form

$$g_{N_L,\dots,N_2}(\mathbf{x}) = \left[\mathbf{a}^{(L+1)}\right]^T \mathbf{g}^L_{N_L}\left(\mathbf{A}^{(L)}\mathbf{g}^{L-1}_{N_{L-1}}\left(\mathbf{A}_{(L-1)}\cdots\mathbf{g}^2_{N_2}(\mathbf{A}^{(2)}\mathbf{x})\right)\right)$$

where $\mathbf{g}^k_{N_k}$ is a component-wise polynomial of degree at most N_k, for $k \in [2, L]$. We start by the case $L = 2$. Notice that each component of $\mathbf{A}^{(2)}\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x})$ is a Fourier network with the same set of first layer weights, of size at most $(2N-1)d_1$. Then, by Lemma 42, we have that each component of

$$\mathbf{f}^2_{N_2,N_1}(\mathbf{x}) \doteq \mathbf{A}^{(3)}\mathbf{g}^2_{N_2}(\mathbf{A}^{(2)}\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x}))$$

is a Fourier network with the same set of first layer weights of size at most

$$(2(2N_1-1)d_1)^{N_2} \ .$$

Finally, consider the induction step. By the assumption hypothesis, the function

$$\mathbf{f}^{L-1}_{N_{L-1},\dots,N_1}(\mathbf{x}) \doteq \mathbf{A}^{(L)}\mathbf{g}^{L-1}_{N_{L-1}}(\mathbf{A}^{(L-1)}\cdots\mathbf{g}^2_{N_2}(\mathbf{A}^{(2)}\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x})))$$

is such that each component is a Fourier network with the same set of first layer weights of size at most

$$\left(2^{L-2}(2N_1-1)d_1\right)^{\prod_{k=2}^{L-1} N_k} \ .$$

Then, by Lemma 42, the function

$$f_N(\mathbf{x}) = \left[\mathbf{a}^{(L+1)}\right]^T \mathbf{g}^L_{N_L}(\mathbf{f}^{L-1}_{N_{L-1},\dots,N_1}(\mathbf{x}))$$

is a Fourier network with at most

$$\left(2\cdot\left(2^{L-2}(2N_1-1)d_1\right)^{\prod_{k=2}^{L-1} N_k}\right)^{N_L} = 2^{N_L}2^{(L-2)\prod_{k=2}^{L-1} N_k}((2N_1-1)d_1)^{\prod_{k=2}^{L} N_k}$$

which implies equation (43). Plugging in the lower bounds on N_k in terms of ϵ, the thesis follows. ■

B.7 Fixed-dimension approximation

The results of Section 4 on fixed-threshold approximation can be complemented by the following result on fixed-dimension approximation. The proposition below is a straightforward generalization of Theorem 3 in (Safran et al., 2019).

Proposition 44 *Let σ be an activation satisfying Assumption 1. Then there exists a constant $\beta > 0$ such that for any $f : B^d_{1,2} \to \mathbb{C}$ 1-Lipschitz function and $\epsilon > 0$ there exists a network $f_N \in \mathcal{F}^\sigma_N$ such that*

$$\|f - f_N\|_{B^d_{1,\infty},\infty} \leq \epsilon$$

for some $N \leq 2 + \beta d^7 \left(\beta \epsilon^{-1}\right)^d \epsilon^{-6}$.

Proof [Proof] The result is proved by noticing that the proof of Theorem 3 in (Safran et al., 2019) actually holds for any function f as in the statement. Moreover, using Assumption 1, f_N can also be chosen so that an equivalent bound holds for $m_\infty(f_N)$. ■

Appendix C. Proofs related to spherical harmonics analysis of shallow networks

C.1 Proof of Proposition 16

Let $f_N : \mathbb{R}^d \to \mathbb{R}$ a one-hidden-layer network defined by

$$f_N(\mathbf{x}) = \sum_{i=1}^N u_i f^{\sigma_i, \mathbf{w}_i}(\mathbf{x}) \doteq \sum_{i=1}^N u_i \sigma_i\left(\mathbf{w}_i^T \mathbf{x}\right)$$

where $\mathbf{u} \in \mathbb{R}^N$, $\mathbf{w}_i \in \mathbb{S}^{d-1}$, and σ_i are linearly bounded activations. Thanks to Parseval's formula, it holds that

$$\begin{aligned}
\|f_N - f^{(d)}\|_2^2 &\geq \|\mathcal{P}_{I_d} f_N - \mathcal{P}_{I_d} f^{(d)}\|_2^2 \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \sum_{j \in I_d} \sum_{i=1}^N u_i \langle f^{\sigma_i, \mathbf{w}_i}, f_j^{(d)} \rangle \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \sum_{j \in I_d} \sum_{i=1}^N \frac{1}{\sqrt{N_j^d}} |u_i| \|f_j^{(d)}\|_\infty \|f_j^{\sigma_i, \mathbf{w}_i}\|_2 \qquad (44) \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \|f^{(d)}\|_2 \sum_{i=1}^N |u_i| \|f^{\sigma_i, \mathbf{w}_i}\|_2 \left[\sum_{j \in I_d} c_{d,j}^2\right]^{1/2} \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \cdot O(d^M) \cdot \epsilon^{d^\alpha} \|f^{(d)}\|_2 \sum_{i=1}^N |u_i| \|f^{\sigma_i, \mathbf{w}_i}\|_2 \ .
\end{aligned}$$

Finally, notice that it holds that

$$\|f^{\sigma_i, \mathbf{w}_i}\|_2 \leq 2\, m_\infty(f_N)$$

and therefore

$$\|f_N - f^{(d)}\|_2^2 \geq \Omega(d^{-2M}) - 4 \cdot O(d^M) \cdot \epsilon^{d^\alpha} \cdot m_\infty^2(f_N) \cdot N \ .$$

This concludes the proof.

C.2 Low-coherence zonal harmonics frames

In this section, we wish to quantify how much incoherent can a frame composed of zonal harmonics be. More specifically, we wish to find a lower bound for

$$N(d,k,\epsilon) = \sup\left\{ N \geq 1 : \exists\, \mathbf{w}_1, \dots, \mathbf{w}_N \in \mathbb{S}^{d-1} : \sup_{i \neq j} \left| P_k^d\big(\mathbf{w}_i^T \mathbf{w}_j\big) \right| \leq \epsilon \right\}$$

for $\epsilon \in (0,1)$.

Lemma 45 *It holds that*

$$N(d,k,\epsilon) \geq \sup\left\{ N \geq 1 : \exists\, \mathbf{w}_1, \dots, \mathbf{w}_N \in \mathbb{S}^{d-1} : \sup_{i \neq j} \left| \mathbf{w}_i^T \mathbf{w}_j \right| \leq \sqrt{1 - \frac{d}{k\epsilon^{4/d}}} \right\}$$

for $k > d \geq 5$ *and* $\left(\frac{d}{k}\right)^{d/4} \leq \epsilon < 1$.

Proof [Proof] We recall that it holds

$$\left| P_k^d(t) \right| \leq \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{d-1}{2}\right) \left(\frac{4}{k(1-t^2)}\right)^{(d-2)/2}$$

for $d \geq 2$ and $t \in (-1,1)$ (cfr. eq. (2.117) in (Atkinson and Han, 2012)) and that

$$\Gamma(x) \leq \left(\frac{x}{2}\right)^{x-1}$$

for $x \geq 2$. Therefore it holds that

$$\begin{aligned} \left| P_k^d(t) \right| &\leq \frac{1}{\sqrt{\pi}} \left(\frac{d-1}{4}\right)^{(d-3)/2} \left(\frac{4}{k(1-t^2)}\right)^{(d-2)/2} \\ &\leq \frac{1}{\sqrt{\pi}} \left(\frac{d}{4}\right)^{-1/2} \left(\frac{d}{k(1-t^2)}\right)^{(d-2)/2} \leq \left(\frac{d}{k(1-t^2)}\right)^{(d-2)/2} \end{aligned}$$

for $d \geq 5$ and $|t| < 1$. In particular, for $\epsilon \in (0,1)$, it holds that $\left| P_k^d(t) \right| \leq \epsilon$ if

$$\frac{d}{k(1-t^2)} \leq \epsilon^{4/d}$$

that is if

$$|t| \leq \sqrt{1 - \frac{d}{k\epsilon^{4/d}}} \ .$$

The thesis follows. ■

Define

$$N(d,\delta) = \sup\left\{ N \geq 1 : \exists\, \mathbf{w}_1, \dots, \mathbf{w}_N \in \mathbb{S}^{d-1} : \sup_{i\neq j} \left|\mathbf{w}_i^T \mathbf{w}_j\right| \leq \delta \right\}$$

for $\delta \in (0,1)$. The previous lemma says that

$$N(d,k,\epsilon) \geq N\left(d, \sqrt{1 - \frac{d}{k\epsilon^{4/d}}}\right) .$$

Example 7 *Taking*

$$\{\mathbf{w}_i\}_{i=1}^N = \left\{ \epsilon \in \left\{ \pm \frac{1}{\sqrt{d}} \right\}^d : \epsilon_1 > 0 \right\} \tag{45}$$

it holds that $N = 2^{d-1}$ *and*

$$\max_{i\neq j} \left|\mathbf{w}_i^T \mathbf{w}_j\right| = 1 - \frac{2}{d} .$$

Therefore

$$N\left(d, 1 - \frac{2}{d}\right) \geq 2^{d-1} .$$

Taking $\epsilon = 2^{-d}$*, it holds that, if* $k \geq 8d^2$*, then*

$$N\left(d, k, 2^{-d}\right) \geq 2^{d-1} .$$

Using this fact it is possible to explicitly construct a high energy sparse function.

Lemma 46 *Take* $k \geq 16d^2$ *even and let*

$$\hat{P}(\mathbf{x}) = \beta_d \sum_{i=1}^{2^{d-1}} (N_k^d)^{1/2} P_k^d(\mathbf{w}_i^T \mathbf{x})$$

with $\beta_d = 2(2^d+2)^{-1/2}$ *and* $\mathbf{w}_i$ *as in equation* (45)*. Then* $\|\hat{P}\|_2 = \Theta_d(1)$ *and it is exponentially spread, that is* $\ell_{\infty,2}(\hat{P}) \leq O_d(2^{-d/2})\sqrt{N_k^d}$*.*

Proof [Proof] It holds that

$$\begin{aligned}
\|\hat{P}\|_2^2 &= \beta_d^2 \left[2^{d-1} + \sum_{i\neq j} P_k^d\left(\mathbf{w}_i^T \mathbf{w}_j\right) \right] \\
&\leq \frac{2}{2^{d-1}+1} \left[2^{d-1} + \left(2^{2d-2} - 2^{d-1}\right) 2^{-d} \right] \\
&= \frac{2}{2^{d-1}+1} \left[2^{d-1} + 2^{d-2} - 2^{-1} \right] \leq 3
\end{aligned}$$

and that

$$\begin{aligned}\|\hat{P}\|_2^2 &\geq \frac{2}{2^{d-1}+1}\Big[2^{d-1} - \Big(2^{2d-2} - 2^{d-1}\Big)2^{-d}\Big] \\ &= \frac{2}{2^{d-1}+1}\Big[2^{d-1} - 2^{d-2} + 2^{-1}\Big] \geq 1 \ .\end{aligned}$$

On the other hand, it holds that

$$\|\hat{P}\|_\infty \leq \beta_d (N_k^d)^{1/2} \sup_{x\in\mathbb{S}^{d-1}} \sum_{i=1}^{2^{d-1}} \Big|P_k^d(\mathbf{w}_i^T\mathbf{x})\Big| \ .$$

By definition of the vectors $\{\mathbf{w}_i\}_{i=1}^{2^{d-1}}$, it holds

$$\begin{aligned}\sup_{x\in\mathbb{S}^{d-1}} \sum_{i=1}^{2^{d-1}} \Big|P_k^d(\mathbf{w}_i^T\mathbf{x})\Big| &= \frac{1}{2} \sup_{x\in\mathbb{S}^{d-1},\, x>0} \sum_{\boldsymbol{\epsilon}\in\{\pm d^{-1/2}\}^d} \Big|P_k^d(\mathbf{x}^T\boldsymbol{\epsilon})\Big| \\ &\leq 1 + \frac{1}{2} \sup_{\mathbf{x}\in\mathbb{S}^{d-1},\, \mathbf{x}\succ 0} \sum_{\boldsymbol{\epsilon}\in\{\pm d^{-1/2}\}^d \,:\, |\mathbf{1}^T\boldsymbol{\epsilon}|<\sqrt{d}} \left(\frac{1}{16d\Big(1-|\mathbf{x}^T\boldsymbol{\epsilon}|^2\Big)}\right)^{(d-2)/2} \\ &\leq 1 + \frac{1}{2}(2^d-2)\left(\frac{1}{16d\big(1-\frac{d-1}{d}\big)}\right)^{(d-2)/2} \leq 1 + \frac{2^{d-1}-1}{4^{d-2}} \leq 2 \ .\end{aligned}$$

This proves the claim. ∎

C.3 Proof of Proposition 21

Assume first that $f \in \mathcal{H}^1$. Then $f = h_\pi$ for some π even signed Radon measure. Thus

$$\begin{aligned}\gamma_1(f) = \|\pi\|_1 &= \sup_{\varphi\in C(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty\leq 1} \int_{\mathbb{S}^{d-1}} \varphi(\mathbf{w})\, d\pi(\mathbf{w}) \\ &= \sup_{\varphi\in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty\leq 1} \int_{\mathbb{S}^{d-1}} \varphi(\mathbf{w})\, d\pi(\mathbf{w}) \\ &= \sup_{\varphi\in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty\leq 1} \int_{\mathbb{S}^{d-1}} T(T^{-1}\varphi)(\mathbf{w})\, d\pi(\mathbf{w}) \\ &= \sup_{\varphi\in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty\leq 1} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} \big|\mathbf{w}^T\mathbf{x}\big|(T^{-1}\varphi)(\mathbf{x})\, dS(\mathbf{x})\, d\pi(\mathbf{w}) \\ &= \sup_{\varphi\in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty\leq 1} \langle T^{-1}\varphi, f\rangle \ .\end{aligned}$$

This shows one side of the statement. On the other hand, assume that

$$\sup_{\varphi\in C^\infty_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_\infty\leq 1} \langle T^{-1}\varphi, f\rangle < \infty \ .$$

Then, the transformation

$$S_f(\varphi) \doteq \langle T^{-1}\varphi, f\rangle$$

defines a bounded linear operator $S_f : C^\infty_{even} \to \mathbb{R}$. Since $C^\infty_{even}(\mathbb{S}^{d-1})$ is dense in $C_{even}(\mathbb{S}^{d-1})$ (the set of even function in $C(\mathbb{S}^{d-1})$), S_f can be extended to a bounded linear operator on $C_{even}(\mathbb{S}^{d-1})$. By setting

$$S_f(\varphi) = S_f(\varphi_{even})$$

we can extend it on $C(\mathbb{S}^{d-1})$. By the Riesz representation theorem, there exists a signed Radon measure π on $\mathbb{S}^{d-1}$ such that

$$S_f(\varphi) = \int_{\mathbb{S}^{d-1}} \varphi(\mathbf{w})\, d\pi(\mathbf{w})$$

for every $\varphi \in C(\mathbb{S}^{d-1})$. Moreover, since $S_f(\varphi) = 0$ for every odd φ, we can assume that π is even. Let h_π be the function in $\mathcal{H}^1$ defined by π. Then it holds that

$$\langle T^{-1}\varphi, f\rangle = \|\pi\|_1 = \langle T^{-1}\varphi, h_\pi\rangle$$

for every $\varphi \in C^\infty_{even}(\mathbb{S}^{d-1})$. Since T is an automorphism over $C^\infty_{even}(\mathbb{S}^{d-1})$, then it holds

$$\langle \varphi, f\rangle = \langle \varphi, h_\pi\rangle$$

for every $\varphi \in C^\infty_{even}(\mathbb{S}^{d-1})$. Since f and h_π are even, this implies that $f = h_\pi$. This concludes the proof.