Function approximation with ReLU-like zonal function networks

H. N. Mhaskar*

Abstract

A zonal function (ZF) network on the $q$ dimensional sphere $S^q$ is a network of the form $\mathbf{x} \mapsto \sum_{k=1}^{n} a_k \phi(\mathbf{x} \cdot \mathbf{x}_k)$ where $\phi : [-1, 1] \to \mathbb{R}$ is the activation function, $\mathbf{x}_k \in S^q$ are the centers, and $a_k \in \mathbb{R}$. While the approximation properties of such networks are well studied in the context of positive definite activation functions, recent interest in deep and shallow networks motivate the study of activation functions of the form $\phi(t) = |t|$, which are not positive definite. In this paper, we define an appropriate smoothness class and establish approximation properties of such networks for functions in this class. The centers can be chosen independently of the target function, and the coefficients are linear combinations of the training data. The constructions preserve rotational symmetries.

1 Introduction

A neural network is a function of the form $\mathbf{x} \mapsto \sum_{k=1}^{n} a_k \phi(\mathbf{x} \cdot \mathbf{x}_k)$ where $\phi : [-1, 1] \to \mathbb{R}$ is the activation function, $\mathbf{x}_k \in \mathbb{R}^q$, $a_k \in \mathbb{R}$, and $\phi : \mathbb{R}^q \to \mathbb{R}$ is an activation function. An attractive property of these networks is that unlike their predecessors, perceptrons [25], they have the universal approximation property: any continuous function on any compact subset of any Euclidean space $\mathbb{R}^q$ can be approximated arbitrarily well by neural networks with very minimal conditions on the activation function. Clearly, neural networks are also important from the point of view of computation in that they can be evaluated in parallel. There are many journals devoted to the topic of neural networks and their properties.

Approximation of multivariate functions by neural networks is a very old topic with some of the first papers dating back to the late 1980s e.g., [9, 6, 8, 30]. Some very general conditions on the activation function that permits the universal approximation property are given in [21, 13]. An important question in this theory is to estimate the number of nonlinear units in the network required to approximate a given function up to a given accuracy. Clearly, as with all similar questions in classical approximation theory, the answer to this question depends upon how one measures the “smoothness” of the target function. Thus, various dimension independent bounds are given in [21, 22, 16]. When the smoothness is measured by the number of derivatives, the optimal estimate consistent with that in the classical theory of polynomial approximation [32] is given in [14]. A survey of these ideas can be found in [23, 15].

Many problems in geophysics lead to the question of approximation of functions on the sphere. Therefore, we studied in [23, 19] approximation on the sphere by neural networks. Unlike the setting of the Euclidean space, translation is not possible anymore, and the neural networks take the form of what we have called Zonal Function (ZF) networks. A zonal function (ZF) network on the $q$ dimensional sphere $S^q$ is a network of the form $\mathbf{x} \mapsto \sum_{k=1}^{n} a_k \phi(\mathbf{x} \cdot \mathbf{x}_k)$ where $\phi : [-1, 1] \to \mathbb{R}$ is the activation function, $\mathbf{x}_k \in S^q$ are the centers, and $a_k \in \mathbb{R}$. This work has been further generalized to the context of very general manifolds satisfying some mild conditions on the heat kernel [20]. In general, this theory is applicable only when the activation function $\phi$ is positive definite (in addition to some other conditions).

The interest in this subject is renewed recently with the advent of deep networks. Deep neural networks especially of the convolutional type (DCNNs) have started a revolution in the field of artificial intelligence and machine learning, triggering a large number of commercial ventures and practical applications. Most deep learning references these days start with Hinton’s backpropagation and with Lecun’s convolutional networks (see for a nice review [12]). We have started an investigation of approximation properties of deep networks in [24], where it is shown that one reason why deep networks are so effective is that they allow one to utilize any compositional structure in the target function. A popular activation function in the theory of deep networks is the ReLU function $\phi(t) = \max(t, 0)$ (or equivalently, $\phi(t) = |t|$). One reason for this popularity is the relative ease with which can formulate training algorithms for multiple layers with this activation function [5]. It is pointed out in [1, 24] that

\*Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711. The research of this author is supported in part by ARO Grant W911NF-15-1-0385. Email: hrushikesh.mhaskar@cgu.edu
the problem of studying function approximation on a Euclidean space by shallow networks evaluating the ReLU activation function is equivalent to the problem of studying approximation of even functions on the sphere by ZF networks evaluating this activation function.

Restricted to the sphere, the function \( \phi(t) = |t| \) is not a positive definite function. The purpose of this paper is to develop a theory analogous to that in [19] where the activation functions are certain generalizations of the ReLU function, in particular, the functions \( \phi_\gamma(t) = |t|^{2\gamma+1} \), where \( 2\gamma + 1 \) is not an even integer. As in the case of our results in [19], our constructions in this paper have several interesting properties.

1. The constructions do not require any machine learning in the classical sense. Given a data of the form \((\xi, f(\xi))\), for \( \xi \) in some finite subset of \( S^q \), we give explicit formulas for the approximating networks.

2. The centers of the networks are chosen independently of the target function or the training data.

3. Since no machine learning is involved, we need not assume any priors on the target function; the constructions are the same for every continuous function on \( S^q \), with estimates in terms of its smoothness (adjusting automatically).

4. The networks are invariant under rotations of \( \mathbb{R}^{q+1} \).

5. The coefficients are linear functionals on the target function \( f \), and if the available data consists of spherical harmonic coefficients of a sufficiently “smooth” \( f \), then the sum of absolute values of these coefficients is bounded independently of the size of the network/amount of available data.

After reviewing some preliminary facts about analysis on the sphere in Section 2, we state our main result and algorithm in Section 3. The proof of the result requires several preparatory technical details which are reviewed in Section 4. The proof itself is given in Section 5. In the Appendix, we derive an expansion of the function \( \phi_\gamma(t) = |t|^{2\gamma+1} \) in terms of certain ultraspherical polynomials.

2 Preliminaries

The statement of our main result requires some notation and background preparation. In Sub-section 2.1, we develop the necessary notation. In Sub-section 2.2, we review the notion of MZ quadrature measures on the sphere, which play a critical role in our construction. In Sub-section 2.3, we define a class of activation functions of interest in this paper, including the ReLU functions, define the associated continuous networks and smoothness classes, as well as review the properties of certain operators essential in our constructions.

2.1 Spherical harmonics

In the sequel, \( \mu_q^* \) denotes the Lebesgue surface (Riemannian volume) measure of \( S^q \). The surface area of \( S^q \) is \( \omega_q = \frac{2\pi^{(q+1)/2}}{\Gamma((q + 1)/2)} \). The geodesic distance between points \( x, y \in S^q \) is given by \( \rho(x, y) = \arccos(x \cdot y) \). For \( r > 0 \), a spherical cap with radius \( r \) and center \( x_0 \in S^q \) is defined by

\[ \mathbb{B}(x_0, r) = \{ y \in S^q : \rho(x_0, y) \leq r \}. \]

We note that for any \( x_0 \in S^q \) and \( r > 0 \),

\[ \mu_q^*(\mathbb{B}(x_0, r)) = \omega_{q-1} \int_0^r \sin^{q-1} t dt \leq \frac{\omega_{q-1}}{q} r^q. \quad (2.1) \]

In the sequel, the term measure will mean a (signed or positive), complete, sigma-finite, Borel measure on \( S^q \). The total variation measure of \( \nu \) will be denoted by \(|\nu|\). If \( \nu \) is a measure, \( 1 \leq p \leq \infty \), and \( f : S^q \rightarrow \mathbb{R} \) is \( \nu \)-measurable, we write

\[ \|f\|_{L^p(\nu)} := \left\{ \int_{S^q} |f(x)|^p d\nu(x) \right\}^{1/p}, \quad \text{if } 1 \leq p < \infty, \]

\[ |\nu| - \text{ess sup}_{x \in S^q} |f(x)|, \quad \text{if } p = \infty. \]

The space of all Lebesgue measurable functions on \( S^q \) such that \( \|f\|_{L^p(\nu)} < \infty \) will be denoted by \( L^p(\nu) \), with the usual convention that two functions are considered equal as elements of this space if they are equal \(|\nu|\)-almost everywhere.
We will omit the mention of $\nu$ if $\nu = \mu_n^*$, unless we feel a cause for any confusion on this account. For example, $L^p = L^p(\mu_n^*)$, $\|f\|_p = \|f\|_{\mu_n^*, p}$. The symbol $C(S^n)$ denotes the class of all continuous, real valued functions on $S^n$, equipped with the norm $\|\cdot\|_{\infty}$.

For a real number $x \geq 0$, let $\Pi_n^x$ denote the class of all spherical polynomials (i.e., the restrictions to $S^n$ of polynomials in $q + 1$ variables) of degree $\leq x$. (This is the same as the class $\Pi_n^q$, where $n$ is the largest integer not exceeding $x$. However, our extension of the notation allows us, for example, to use the simpler notation $\Pi_{n/2}$ rather than the more cumbersome notation $\Pi_{(n/2)}^q$.)

For a fixed integer $\ell \geq 0$, the restriction to $S^n$ of a homogeneous harmonic polynomial of exact degree $\ell$ is called a spherical harmonic of degree $\ell$. Most of the following information is based on [26, Section IV.2], and [3, Chapter XI], although we use a different notation. The class of all harmonic polynomial of exact degree $\ell$ exceeding $\ell$ is a univariate polynomial of degree $\ell$.

In particular, for $\phi \in \Pi_n^x$, we define formally

$$ \int_{S^n} \phi(t) p^{(q/2-1,q/2-1)}(t)(1 - t^2)^{q/2-1} dt = \delta_{\ell,k}, \quad \ell, k = 0, 1, \ldots. $$

In particular, for $x \in S^n$, $\ell = 0, 1, \ldots,$

$$ d_q^\ell := \dim \mathbb{H}_\ell^n = \left\{ \begin{array}{ll}
2\ell + q - 1 \left( \frac{\ell + q - 1}{\ell} \right), & \text{if } \ell \geq 1, \\
1, & \text{if } \ell = 0.
\end{array} \right. \quad (2.2) $$

and that of $\Pi_n^x$ is $\sum_{\ell=0}^n d_q^\ell = d_{q+1}^\ell$. Furthermore, $L^2 = L^2$–closure$\left( \bigoplus_{\ell=0}^\infty \mathbb{H}_\ell^n \right)$. Hence, if we choose an orthonormal basis $\{Y_{\ell,k} : k = 1, \ldots, d_q^\ell\}$ for each $\mathbb{H}_\ell^n$, then the set $\{Y_{\ell,k} : k = 1, \ldots, d_q^\ell \ell = 0, 1, \ldots\}$ is a complete orthonormal basis for $L^2$. If $1 \leq p \leq \infty$, $f \in L^p$ and $n > 0$, we define

$$ E_{n,p}(f) = \min \{ \|f - P\|_p : P \in \Pi_n^x \}, $$

and write $X^p = \{f \in L^p : \lim_{n \to \infty} E_{n,p}(f) = 0\}$. Thus, $X^p = L^p$ if $1 \leq p < \infty$ and $C(S^n)$ if $p = \infty$.

One has the well-known addition formula [26 and 3, Chapter XI, Theorem 4] connecting $Y_{\ell,k}$’s with Jacobi polynomials defined in [1.1]:

$$ \sum_{k=1}^{d_q^\ell} Y_{\ell,k}(x)Y_{\ell,k}(y) = \omega_{q-1}^{-1} p^{(q/2-1,q/2-1)}(1)p^{(q/2-1,q/2-1)}(x \cdot y), \quad \ell = 0, 1, \ldots, \quad (2.4) $$

where each $p^{(q/2-1,q/2-1)}$ is a univariate polynomial of degree $\ell$ with positive leading coefficient, and one has the orthogonality relation

$$ \int_{-1}^1 p^{(q/2-1,q/2-1)}(t)p^{(q/2-1,q/2-1)}(1) dt = \delta_{\ell,k}, \quad \ell, k = 0, 1, \ldots. $$

If $f \in L^1$, we define

$$ \hat{f}(\ell, k) = \int_{S^n} f(y) Y_{\ell,k}(y) d\mu_q^*(y), \quad k = 1, \ldots, d_q^\ell, \quad \ell = 0, 1, \ldots. \quad (2.6) $$

We note that if $f$ is an even function, then [2.4] shows that $\hat{f}(2\ell + 1, k) = 0$ for $\ell = 0, 1, \ldots$. If $\phi : [-1, 1] \to \mathbb{R}$, we define formally

$$ \hat{\phi}(\ell) = \frac{\omega_{q-1}^{-1}}{p^{(q/2-1,q/2-1)}(1)} \int_{-1}^1 \phi(t) p^{(q/2-1,q/2-1)}(t)(1 - t^2)^{q/2-1} dt, $$

so that we have the formal expansions

$$ \phi(x \cdot y) = \sum_{\ell=0}^\infty \hat{\phi}(\ell) \sum_{k=1}^{d_q^\ell} Y_{\ell,k}(x)Y_{\ell,k}(y), \quad (2.8) $$

and

$$ \int_{S^n} f(y) \phi(x \cdot y) d\mu_q^*(y) = \sum_{\ell=0}^\infty \hat{\phi}(\ell) \sum_{k=1}^{d_q^\ell} \hat{f}(\ell,k) Y_{\ell,k}(x). \quad (2.9) $$
2.2 Regular measures on the sphere

The space of all signed (or positive), complete, sigma-finite, Borel measures on $\mathbb{S}^q$ will be denoted by $\mathcal{M}$. If $\nu \in \mathcal{M}$, the set $\text{supp}(\nu)$ is the set of all $x \in \mathbb{S}^q$ with the property that $|\nu|(B(x,r)) > 0$ for all $r > 0$. It is easy to verify that $\text{supp}(\nu)$ is a compact set.

**Definition 2.1** Let $d > 0$. A measure $\nu \in \mathcal{M}$ will be called $d$–regular if

$$|\nu|(B(x,d)) \leq cd^q, \quad x \in \mathbb{S}^q. \quad (2.10)$$

The infimum of all constants $c$ which work in (2.10) will be denoted by $\|\nu\|_{R,d}$, and the class of all $d$–regular measures will be denoted by $\mathcal{R}_d$.

For example, $\mu^*_{q}$ itself is in $\mathcal{R}_d$ with $\|\mu^*\|_{R,d} \leq c$ for every $d > 0$ (see (2.1)).

The following proposition (cf. [7, Proposition 5.6]) reconciles different notions of regularity condition on measures defined in our papers.

**Proposition 2.1** Let $d \in (0,1]$, $\nu \in \mathcal{M}$.

(a) If $\nu$ is $d$–regular, then for each $r > 0$ and $x \in \mathbb{S}^q$,

$$|\nu|(B(x,r)) \leq c\|\nu\|_{R,d} \mu^*(B(x,c(r + d))) \leq c_1\|\nu\|_{R,d}(r + d)^q. \quad (2.11)$$

Conversely, if for some $A > 0$, $|\nu|(B(x,r)) \leq A(r + d)^q$ or each $r > 0$ and $x \in \mathbb{S}^q$, then $\nu$ is $d$–regular, and $\|\nu\|_{R,d} \leq 2^q A$.

(b) For each $\alpha > 0$,

$$\|\nu\|_{R,\alpha d} \leq c_1(1 + 1/\alpha)^q\|\nu\|_{R,d} \leq c_1^2(1 + 1/\alpha)^q(\alpha + 1)^q\|\nu\|_{R,\alpha d}, \quad (2.12)$$

where $c_1$ is the constant appearing in (2.11).

(c) If $\nu$ is $d$–regular, then $\|P\|_{\nu,p} \leq c_1\|\nu\|_{R,d}\|P\|_{\mu,p}$ for all $P \in \Pi_{1/d}$ and all $1 \leq p < \infty$. Conversely, if for some $A > 0$ and some $1 \leq p < \infty$, $\|P\|_{\nu,p} \leq A^{1/p}\|P\|_{\mu,p}$ for all $P \in \Pi_{1/d}$, then $\nu$ is $d$–regular, and $\|\nu\|_{R,d} \leq c_2 A$.

In this paper, we are interested in quadrature formulas exact for integrating spherical polynomials, based on the training data; i.e., without assuming any specific location of the nodes involved in such a formula, in contrast to such well known formulas as the Driscoll-Healy or Clenshaw-Curtis formulas.

**Definition 2.2** Let $n \geq 1$. A measure $\nu \in \mathcal{M}$ is called a quadrature measure of order $n$ if

$$\int_{\mathbb{S}^q} P d\nu = \int_{\mathbb{S}^q} P d\mu^*, \quad P \in \Pi_n^q. \quad (2.13)$$

An MZ (Marcinkiewicz-Zygmund) quadrature measure of order $n$ is a quadrature measure $\nu$ of order $n$ for which $\|\nu\|_{R,1/n} < \infty$.

If $C \subset \mathbb{S}^q$, we define the mesh norm $\delta(C)$ (also known as fill distance, covering radius, density content, etc.) and minimal separation $\eta(C)$ by

$$\delta(C) = \sup_{x \in \mathbb{S}^q} \inf_{y \in C} \rho(x,y), \quad \eta(C) = \inf_{x,y \in C, x \neq y} \rho(x,y). \quad (2.14)$$

The first part of the following proposition ([20, Lemma 5.3]) gives an example of a regular measure apart from the surface measure itself. The second part ([7, Theorem 5.7]) asserts the existence of a regular measure supported on a given subset $C$, that integrates polynomials of certain degree exactly, and the third part ([7, Theorem 5.8]) asserts that any positive quadrature formula necessarily defines a regular measure.
Proposition 2.2 (a) If $C$ is finite, the measure that associates the mass $\eta(C)^q$ with each point of $C$ is $\eta(C)$-regular, and $\|\nu\|_{R,\eta(C)} \leq c$.

(b) There exist constants $c_1, c_2 > 0$ with the following property: If $\nu$ is a signed measure, $\delta(\text{supp } (\nu)) < d < c_1$ and $0 < N < c_2 d^{-1}$, then there exists a simple function $W : \text{supp } (\nu) \to [0, \infty)$, satisfying

$$\int_{S^n} P(y) d\mu(y) = \int_{S^n} P(y) W(y) d\nu(y), \quad P \in \Pi_n^1.$$  \hfill (2.15)

If $\nu$ is $d$-regular, then $W(y) \geq c\|\nu\|_{R,d}^{-1}$, $y \in S^n$.

(c) There exists a constant $c > 0$ such that if $n \geq c$ and $\tau$ is a positive quadrature measure of order $2n$, then for $1 \leq p < \infty$,

$$\|P\|_{\tau,p} \sim \|P\|_{\mu,p}, \quad P \in \Pi_n^q,$$ \hfill (2.16)

where the constants involved may depend upon $p$ but not on $\tau$ or $n$. In particular, $\tau$ is $1/n$-regular, and $\|\nu\|_{R,1/n} \leq c$.

We point out a consequence of Proposition 2.2.

Remark 2.1 Let $C$ be a finite subset of $S^n$. By removing close by points, we may obtain a subset $\tilde{C} \subseteq C$ such that $\delta(C) \sim \delta(\tilde{C})$ and $\eta(C) \leq 2\delta(\tilde{C}) \leq 4\eta(\tilde{C})$ ([20, Proposition 2.1]). We rename $\tilde{C}$ to be $C = \{x_1, \cdots, x_M\}$. According to Proposition 2.2(a), the measure $\nu$ that associates the mass $\eta(C)^q$ with each $x_j$ is a $\eta(C)$-regular measure with $\|\nu\|_{R,\eta(C)} \leq c$. Proposition 2.2(b) then asserts the existence of positive numbers $w_j$ such that

$$\eta(C)^q \sum_{j=1}^M w_j P(x_j) = \int_{S^n} P(x) d\mu_q(x), \quad P \in \Pi_{\eta(C)^q},$$ \hfill (2.17)

and $w_j \geq c$ for each $j$. It can be shown, in fact, that $w_j \sim 1$. Algorithms to compute the quadrature weights are discussed in [11].

\hfill \Box

2.3 Smoothness classes and operators

First, we define a smoothness class on the sphere, comprising essentially of those functions which are a “continuous ZF network” of the form (2.20) below. In the sequel, if $a \in \mathbb{C}$, we define

$$a^{-1} = \begin{cases} 1/a, & \text{if } a \neq 0, \\ 0, & \text{if } a = 0. \end{cases} \quad (2.18)$$

we denote by $W_{q, \phi}^p$ the set of all $f \in X_p^M$ such that there exists a $D_{q, \phi}(f) \in X_p^M$ such that (cf. [10])

$$\tilde{D}_{q, \phi}(f)(\ell, k) = \hat{\phi}(\ell)^{-1} f(\ell, k), \quad k = 1, \cdots, d^q, \ell = 0, 1, \cdots .$$ \hfill (2.19)

Using (2.9), it is then clear that

$$f(x) = \int_{S^q} \phi(x \cdot y) D_{q, \phi}(f)(y) d\mu_q(y), \quad f \in W_{q, \phi}^p.$$ \hfill (2.20)

In the case when $\hat{\phi}(\ell) > 0$ for all $\ell$, then the space $W_{q, \phi}^2$ is the reproducing kernel Hilbert space with $\phi$ as the reproducing kernel. The main technical challenge of this paper is that the function $\phi_\gamma(t) = \|t\|^{2\gamma+1}$ does not satisfy this condition. In fact, it follows from Proposition A.1 (used with $\alpha = q/2 - 1$) that $\hat{\phi}_\gamma(2\ell + 1) = 0$ and with $x! = \Gamma(x + 1)$,

$$\hat{\phi}_\gamma(2\ell) = (-1)^{\ell} \frac{\cos(\pi \gamma)(q/2 - 1)!/(2\gamma + 1)!}{2^{\gamma+1} \sqrt{\pi}} \frac{(\ell - \gamma - 3/2)!}{(\ell + \gamma + q/2)!}. \quad \hfill (2.21)$$

The fact that $\phi_\gamma$ is not a positive definite function on the sphere implies that $D_{\phi_\gamma}$ is not a differential operator. As shown in Figure 1, it is not even a local operator; i.e., if a function $f$ is supported on a cap centered at a point $x_0 \in S^n$, $D_{\phi_\gamma}(f)$ is not supported on this cap, but instead on an equatorial region perpendicular to $x_0$.

We now turn our attention to the conditions that we wish to require on $\{\hat{\phi}(\ell)\}_{\ell=0}^\infty$ instead of positivity. For a sequence $a = \{a_\ell\}_{\ell \geq 0}$, the forward differences are defined by

$$\Delta a_\ell = \Delta^1 a_\ell = a_{\ell+1} - a_\ell, \quad \Delta^r a_\ell = \Delta(\Delta^{r-1} a_\ell), \quad \ell \geq 0, \ r \geq 2. \quad \hfill (2.22)$$
Definition 2.3 Let $s \in \mathbb{R}$. A sequence $b = \{b_\ell\}_{\ell=0}^\infty$ of real numbers is called an $s$-sequence (written $b \in B(s)$) if

$$\sup_{\ell,r \geq 0} (\ell + 1)^r |\Delta^r (\ell + 1)^s b_\ell| \leq c. \quad (2.23)$$

The class $A(s)$ consists of continuous, even functions $\phi : [-1, 1] \to \mathbb{R}$ such that $\{(-1)^\ell \hat{\phi}(2\ell)\}_{\ell=0}^\infty \in B(s)$.

In Corollary 4.1, we will show that for any sequence $b \in B(s)$, $s > (q + 1)/2$, there exists a continuous, even function $\phi$ such that $\hat{\phi}(2\ell) = (-1)^\ell b_\ell$, $\ell = 0, 1, 2, \ldots$, so that $\phi \in A(s)$. We note that the condition (2.23) is satisfied if $b$ satisfies an asymptotic expansion of the form

$$b_\ell = \ell^{-s} \sum_{j=0}^\infty \frac{d_j}{\ell^j}. \quad (2.24)$$

In particular, $\phi, \gamma \in A((4\gamma + 3 + q)/2)$ if $\gamma > -1/2$ and $2\gamma + 1$ is not an even integer (cf. (A.2)).

Finally, we define an operator and review its properties. Let $S > q + 1$ be an integer, $h : [0, \infty) \to [0, 1]$ be an $S$ times continuously differentiable function such that $h(t) = 1$ if $0 \leq t \leq 1/2$, $h(t) = 0$ if $t \geq 1$. We write

$$\Phi_n(x, y) = \omega_{q-1}^{-1} \sum_{\ell=0}^\infty h(\ell/n) p_{2\ell}^{(q/2-1,q/2-1)} (1 - p_{2\ell}^{(q/2-1,q/2-1)})(x \cdot y), \quad n > 0, \ x \in \mathbb{S}^q. \quad (2.25)$$

If $\nu$ is a measure on $\mathbb{S}^q$, and $f \in L^1(\nu)$ is an even function, we define

$$\sigma_n(\nu; f, x) = \int_{\mathbb{S}^q} f(y) \Phi_n(x, y) d\nu(y). \quad (2.26)$$

The following proposition summarizes important properties of the operator $\sigma_n$. The estimate (2.27) is proved in [7], the estimate (2.28) was proved in [13] (with a different notation), the rest of the assertions can be verified easily from the definitions.

Proposition 2.3 Let $1 \leq p \leq \infty$, $n \geq 1, \nu$ be an MZ quadrature measure of order $4n$, $f \in L^p$ be an even function. Then $\sigma_n(\nu; f) \in \Pi_{2n}^q$. If $P \in \Pi_{2n}^q$ then $\sigma_n(\nu; P) = P$. We have

$$||\sigma_n(\nu; f)||_p \leq c \|\nu\|^{1/p'}_{H,1/n} \|f\|_{L^p}, \quad (2.27)$$

and

$$E_{4n,p}(f) \leq ||f - \sigma_n(\nu; f)||_p \leq c \|\nu\|^{1/p'}_{H,1/n} \min_{P \in \Pi_{2n}^q} ||f - P||_{L^p}. \quad (2.28)$$

Further,

$$\sigma(\mu^*_q; D_\phi(f)) = D_\phi \left( \sigma(\mu^*_q; f) \right). \quad (2.29)$$
3 Main results

The results in this section are stated in a much greater generality than required for studying approximation on the Euclidean space, partly because the details are not any simpler in restricting ourselves to the case $\phi_\gamma(t) = ||t||^{2\gamma+1}$ and the uniform norm. Accordingly, our main theorem, Theorem 3.1 will be stated in a very general and abstract form.

In this section, let $\phi: [-1,1] \to \mathbb{R}$ be an even, continuous function in $A(s)$ for some $s > (q + 1)/2$. Let $\mu$, $\nu$ be measures on $\mathbb{S}^q$, $f \in L^p(\mu)$. We define for $n > 0$, the (abstract) ZF network approximating $f$ (cf. (3.3)) by

$$G_n(\phi, \mu, \nu; f, x) = \int_{\mathbb{S}^q} \phi(x \cdot y)D_\phi(\sigma_n(\mu; f))(y) d\nu(y).$$

(3.1)

Before stating the main theorem, we first illustrate its implementation, and make some comments, explaining in particular why (3.1) defines an abstract ZF network. In Algorithm 1 below, we illustrate the construction of the network in the case when $p = \infty$, using a training data of the form $(\xi, f(\xi))$, $\xi \in \tilde{C}$, for a finite subset $\tilde{C} \subset \mathbb{S}^q$.

Algorithm 1 Construction of ZF network

a) Input: $(\xi, f(\xi))$, $\xi \in \tilde{C}$, for a finite subset $\tilde{C} \subset \mathbb{S}^q$.
b) Find an integer $N$, and weights $\tilde{w}_\xi$, $\xi \in \tilde{C}$, by solving the under-determined system of equations

$$\sum_{\xi \in \tilde{C}} \tilde{w}_\xi Y_{\ell,j}(\xi) = \delta_{(\ell,j),(0,0)}, \quad j = 1, \ldots, d^q_\ell, \quad \ell = 0, \ldots, 4N,$

/* $N$ is chosen by trial and error so as to minimize $\sum_{\xi \in \tilde{C}} |\tilde{w}_\xi|^2$ or to control the residual error or the condition number. The measure $\mu$ in Theorem 3.1 associates the mass $\tilde{w}_\xi$ with $\xi, \xi \in \tilde{C}$. It can be shown that $\|\mu\|_{R^{1/(4N)}}$ is at most a constant times the largest singular value of the matrix involved in the above system of equations.*

c) With $N$ as above, we use any known quadrature formula i.e., a set of points $C = \{x_k\}$, and weights $w_k \geq 0$ such that

$$\sum_{x_k \in C} w_k P(x_k) = \int_{\mathbb{S}^q} P d\mu^*_q, \quad P \in \Pi_4^{2N}.$$

(3.2)

/* The measure $\nu$ in Theorem 3.1 associates the mass $w_k$ with $x_k$, $x_k \in C$. In view of Proposition 2.2(c), the measure $\nu$ is an MZ quadrature measure of order $4N$.*

d) Let

$$\Psi_N(x_k, \xi) = w_k \tilde{w}_\xi \sum_{\ell} \sum_{j=1}^N d^q_\ell h(\frac{\ell}{N}) \hat{\phi}(\frac{\ell}{N}) Y_{2\ell,j}(x_k) Y_{2\ell,j}(\xi), \quad x_k \in C, \quad \xi \in \tilde{C}.$$

e) Output:

$$G(\phi, \mu, \nu; f, x) = \sum_{x_k \in C} \left\{ \sum_{\xi \in \tilde{C}} \Psi_N(x_k, \xi)f(\xi) \right\} \phi(x \cdot x_k) = \sum_{\xi \in \tilde{C}} \left\{ \sum_{x_k \in C} \Psi_N(x_k, \xi) \phi(x \cdot x_k) \right\} f(\xi).$$

(3.3)

Remark 3.1 The first formula in (3.3) shows that $G$ is a ZF network, whose coefficients can be computed as a linear combination of the training data. Also, the centers of this network are independent of the training data. The second formula in (3.3) shows that the network can be constructed by treating the training data as coefficients of a set of fixed networks constructed independently of the target function; based entirely on the locations of the sampling points.

Remark 3.2 In view of (2.4), if $U$ is a rotation of $\mathbb{R}^{q+1}$, and $f_U(x) = f(Ux)$ for $x \in \mathbb{S}^q$, then the rotation invariance of $\mu^*_q$ implies that for integer $\ell = 0, 1, \ldots, 1$

$$\sum_{j=1}^{d^q_\ell} \int_{\mathbb{S}^q} f(Uy) p^{(q/2-1,q/2-1)}_\ell(x \cdot y) d\mu^*_q(y) = \sum_{j=1}^{d^q_\ell} \int_{\mathbb{S}^q} f(\ell, j) Y_{\ell,j}(Ux).$$
The rotation invariance implies also that if \( \{x_k\} \) satisfies (3.2), then so does the system \( \{UX_k\} \), with the same weights. Therefore, using the definition (3.3) once with \( f_U \) in place of \( f \), and once with \( \{UX_k\} \) in place of \( \{x_k\} \), it is easy to deduce that
\[
G_{2^n}(\phi, \mu_q^*; \nu; f, UX_k), \quad x \in \mathbb{S}^q.
\]
(3.4)

Our main theorem is the following (cf. [19, Theorem 3.1]).

**Theorem 3.1** Let \( s > \frac{(q + 1)}{2} \), \( \phi \in \mathcal{A}(s), \) \( 1 \leq p \leq \infty \), \( f \in W_{\mu, \nu}^p \), \( n \geq 1 \) and \( \nu \) be an MZ quadrature measure of order \( 2^{n+2} \). Then
\[
\|f - G_{2^n}(\phi, \mu_q^*; \nu, f)\|_p \leq c2^{-n(s-(q-1)/2)}(1 + \|\nu\|_{R,1/2^n})\|D_\phi(f)\|_p.
\]
(3.5)
Moreover, the coefficients in the network satisfy
\[
\int_{\mathbb{S}^q} |D_\phi (\sigma_n(\mu_q^*; f)) (y)| \|\nu\|(y) \leq c\|D_\phi(f)\|_1.
\]
(3.6)

If \( p = \infty \), and \( \mu \) is an MZ quadrature measure of order \( 2^{n+2} \) then
\[
\|f - G_{2^n}(\phi, \mu, \nu, f)\|_\infty \leq c2^{-n(s-(q-1)/2)}\|\mu\|_{R,1/2^n}(1 + \|\nu\|_{R,1/2^n})\|D_\phi(f)\|_\infty.
\]
(3.7)

We state a corollary for the special case \( \phi_\gamma(t) = |t|^{2\gamma+1} \), obtained by noting from (A.2) that \( \phi_\gamma \in \mathcal{A}((4\gamma+3+q)/2) \).

**Corollary 3.1** Let \( \gamma > -1/2, 2\gamma + 1 \) not be an even integer. Let \( 1 \leq p \leq \infty, f \in W_{\mu, \nu}^p, n \geq 1 \) and \( \nu \) be an MZ quadrature measure of order \( 2^{n+2} \). Then
\[
\|f - G_{2^n}(\phi_\gamma, \mu_q^*; \nu, f)\|_p \leq c2^{-n(\gamma+1)}(1 + \|\nu\|_{R,1/2^n})\|D_{\phi_\gamma}(f)\|_p.
\]
(3.8)
Moreover, the coefficients in the network satisfy
\[
\int_{\mathbb{S}^q} |D_\phi (\sigma_n(\mu_q^*; f)) (y)| \|\nu\|(y) \leq c\|D_{\phi_\gamma}(f)\|_1.
\]
(3.9)

If \( p = \infty \), and \( \mu \) is an MZ quadrature measure of order \( 2^{n+2} \) then
\[
\|f - G_{2^n}(\phi_\gamma, \mu, \nu, f)\|_\infty \leq c2^{-n(\gamma+1)}\|\mu\|_{R,1/2^n}(1 + \|\nu\|_{R,1/2^n})\|D_{\phi_\gamma}(f)\|_\infty.
\]
(3.10)

**Remark 3.3** Although the theorem and its corollary are stated with dyadic integers \( 2^n \), this is only for the convenience of the proof. Since for any integer \( m \), we can always find \( n \) with \( 2^{n-1} \leq m \leq 2^n \), in practice, we may use other integers as well without affecting the degree of approximation except for the constants involved.

**Remark 3.4** In the case of the ReLU activation function \( t \mapsto |t| \), the right hand sides of the estimates (3.8) and (3.10) are both \( O(2^{-2n}) \), a substantial improvement on the result announced in [24]. In terms of the number \( M \) of parameters involved in these ZF networks, the estimates are \( O(M^{-2/4}) \). We conjecture that this is the best possible.

## 4 Technical preparation

In this section, we review some facts about Jacobi polynomials, and certain kernels defined with these. In the sequel, we denote \( \Gamma(z+1) \) by \( z! \).

For \( \alpha, \beta > -1, \ x \in (-1,1) \) and integer \( \ell \geq 0 \), the Jacobi polynomials \( p^{(\alpha, \beta)}_\ell \) are defined by the Rodrigues’ formula [31, Formulas (4.3.1), (4.3.4)]
\[
(1 - x)^{\alpha}(1 + x)^{\beta} p^{(\alpha, \beta)}_\ell (x) = \frac{2\ell + \alpha + \beta + 1}{2^{\alpha+\beta+1}} \frac{\ell!}{(\ell + \alpha)!} (1 + x)^{\ell} \int_0^1 (-1)^t t^\ell \frac{dt}{dx} ((1 - x)^{\ell+\alpha}(1 + x)^{\ell+\beta}).
\]
(4.1)
Each \( p_{\ell}^{(\alpha, \beta)} \) is a polynomial of degree \( \ell \) with positive leading coefficient, one has the orthogonality relation

\[
\int_{-1}^{1} p_{\ell}^{(\alpha, \beta)}(x) p_{j}^{(\alpha, \beta)}(x) (1 - x)^{\alpha}(1 + x)^{\beta} = \delta_{\ell,j},
\]

(4.2)

and

\[
p_{\ell}^{(\alpha, \beta)}(1) = \left\{ \frac{2\ell + \alpha + \beta + 1}{2^{\alpha+\beta+1}} \left( \frac{\ell!}{(\ell+\alpha)! (\ell+\beta)!} \right)^{1/2} \right\} \sim \ell^{\alpha+1/2}.
\]

(4.3)

It follows that \( p_{\ell}^{(\alpha, \beta)}(-x) = (-1)^{\ell} p_{\ell}^{(\alpha, \beta)}(x) \). In particular, \( p_{2\ell}^{(\alpha, \alpha)} \) is an even polynomial, and \( p_{2\ell+1}^{(\alpha, \alpha)} \) is an odd polynomial. We note (cf. [31, Theorem 4.1]) that

\[
p_{2\ell}^{(\alpha, \alpha)}(x) = 2^{\alpha+1/4} p_{\ell}^{(-1/2)}(2x^2 - 1) = 2^{\alpha+1/4} (-1)^{\ell} p_{\ell}^{(-1/2)}(1 - 2x^2),
\]

(4.4)

\[
p_{2\ell+1}^{(\alpha, \alpha)}(x) = 2^{\alpha+1/2} (\cos x)^\alpha p_{\ell}^{(1/2)}(2x^2 - 1) = 2^{\alpha+1/2} (-1)^{\ell} (\cos x)^\alpha p_{\ell}^{(1/2)}(1 - 2x^2).
\]

(4.5)

Next, we discuss the localization properties of some kernels associated with Jacobi polynomials. We record the following theorem, which follows easily from [17, Lemma 4.10].

**Theorem 4.1** Let \( \alpha, \beta \geq -1/2, S > \max(\alpha + 3/2, -\alpha + 1), n \geq 1 \) be integers, \( a = \{a_\ell\}_{\ell=0}^\infty \) be a sequence, such that \( a_\ell = 0 \) for all \( \ell \geq n \). Then for \( \theta \in [0, \pi] \),

\[
\left| \sum_{\ell=0}^\infty a_\ell p_{\ell}^{(\alpha, \beta)}(1) p_{\ell}^{(\alpha, \beta)}(\cos \theta) \right| \leq c_1 \sum_{\ell=0}^\infty (\ell + 1)^{2\alpha+2} \sum_{r=1}^K \min(1, (\ell + 1)\theta)^{-S} (\ell + 1)^{-1} |\Delta^r a_\ell|,
\]

(4.6)

where \( K \geq \max(S - \alpha - 1/2, S - \alpha + \beta) \) is an integer (\( \geq 1 \)).

**Proof.** In this proof only, let

\[ \Psi(\theta) = \sum_{\ell=0}^\infty a_\ell p_{\ell}^{(\alpha, \beta)}(1) p_{\ell}^{(\alpha, \beta)}(\cos \theta). \]

In [17, Lemma 4.10] (used with \( y = \cos \theta \)), we have proved that for integer \( K \geq 1 \),

\[
|\Psi(\theta)| \leq c \left\{ \sum_{\ell=0}^\infty \min_{K-1} \left( (\ell + 1)^{2}, (1 - \cos \theta)^{-1} \right)^{\alpha/2+K/2+1/4} \times \sum_{m=0}^{K-1} (\ell + 1)^{\alpha+1/2-m-1} |\Delta^{K-m} a_\ell|, \right. \]

if \( 0 \leq \theta \leq \pi/2 \),

\[
\sum_{\ell=0}^\infty (\ell + 1)^{\alpha+\beta+1} \sum_{m=0}^{K-1} (\ell + 1)^{-m-1} |\Delta^{K-m} a_\ell|, \]

if \( \pi/2 < \theta \leq \pi \).

(4.7)

First, let \( 0 \leq \theta \leq \pi/2 \). A change in the summation index \( m \) to \( K - r \) yields

\[
\sum_{m=0}^{K-1} (\ell + 1)^{\alpha+1/2-m} |\Delta^{K-m} a_\ell| = (\ell + 1)^{\alpha+3/2-K} \sum_{r=1}^{K} (\ell + 1)^{-r} |\Delta^r a_\ell|.
\]

Since \( 1 - \cos \theta \sim \theta^2 \), and \( K + \alpha + 1/2 \geq S \), the first estimate in (4.7) leads to (4.6). If \( \pi/2 < \theta \leq \pi \), then \( 1 - \cos \theta \geq 1 \). Therefore, a similar change of the index of summation in the second estimate in (4.7) leads to (4.6). \( \square \)

We will use heavily the following consequence of Theorem 4.1

**Theorem 4.2** Let \( \alpha \geq -1/2, S > \alpha + 3/2 \) be an integer, \( G : [0, \infty) \to [0, 1] \) be \( S \) times continuously differentiable, \( G(t) = 0 \) for \( t \) in a neighborhood of \( 0 \), \( G(t) = 0 \) if \( t \geq 1 \). Let \( s \in \mathbb{R}, \{b_\ell\}_{\ell=0}^\infty \in \mathcal{B}(s) \). Then

\[
\left| \sum_{\ell=0}^\infty b_\ell G \left( \frac{\ell}{n} \right) p_{2\ell}^{(\alpha, \alpha)}(1) p_{2\ell}^{(\alpha, \alpha)}(\cos \theta) \right| \leq c_1(G) n^{-s+2\alpha+2} \min(1, (n\theta)^{-S}, (n(\pi - \theta))^{-S})
\]

(4.8)

\[
\left| \sum_{\ell=0}^\infty (-1)^{\ell} b_\ell G \left( \frac{\ell}{n} \right) p_{2\ell}^{(\alpha, \alpha)}(1) p_{2\ell}^{(\alpha, \alpha)}(\cos \theta) \right| \leq c_2(G) n^{-s+\alpha+3/2} \min(1, (n\pi/2 - \theta)^{-S})
\]

(4.9)
Proof. We note first that the assumption (2.23) implies (cf. [11] Lemma 4.3(a), (b)) that for any $t \in \mathbb{R}$,
\[
\sum_{\ell=0}^{\infty} \sum_{r=1}^{S} (\ell + 1)^{r-1} |\Delta^r(b_{t}G(\ell/n))| \leq c \sum_{\ell=0}^{\infty} \sum_{r=1}^{S} (\ell + 1)^{r-s-1} |\Delta^r(G(\ell/n))|.
\]
Using the mean value theorem, and the fact that $G$ is supported on $[a,1]$ for some $a > 0$, we conclude (cf. [17] Proof of Theorem 3.1]) that for any $t \in \mathbb{R}$,
\[
\sum_{\ell=0}^{\infty} \sum_{r=1}^{S} (\ell + 1)^{r-s-1} |\Delta^r(b_{t}G(\ell/n))| \leq cn^{t-s}.
\] (4.10)
Therefore, (4.6) in this case leads to
\[
\left| \sum_{\ell=0}^{\infty} b_{\ell}G \left( \frac{\ell}{n} \right) p_{t}^{(\alpha,\beta)}(1) p_{t}^{(\alpha,\beta)}(\cos(2\theta)) \right| \leq c(G)n^{2\alpha+2-s} \min(1,(n\theta)^{-S},(n(\pi-\theta))^{-S}).
\] (4.11)
We use this estimate with $\beta = -1/2$, and recall the first equation in (4.4) to deduce (4.8).
To obtain (4.9), we deduce first using the Leibnitz formula for differences that \( \{b_{t}p_{t}^{(\alpha,\alpha)}(1)/p_{t}^{(-1/2,2)}(1)\}_{t=0}^{\infty} \in \mathcal{B}(s-\alpha-1/2) \). Therefore, we use the second equality in (4.4), and (4.11) with $-1/2$ in place of $\alpha$, $\alpha$ in place of $\beta$, $s-\alpha-1/2$ in place of $s$, to conclude that
\[
\left| \sum_{\ell=0}^{\infty} (-1)^{t} b_{\ell}G \left( \frac{\ell}{n} \right) p_{2t}^{(\alpha,\alpha)}(1) p_{2t}^{(\alpha,\alpha)}(\cos(\theta)) \right|
\leq 2^{\alpha/2+1/4} \sum_{\ell=0}^{\infty} b_{\ell} p_{2t}^{(\alpha,\alpha)}(1) G \left( \frac{\ell}{n} \right) p_{2t}^{(-1/2,2)}(1) p_{2t}^{(-1/2,2)}(\cos(\pi-2\theta))
\leq c(G)n^{-s+\alpha/3} \min(1,(n\pi/2-\theta)^{-S}).
\]

In the remainder of this paper, we overload the notation and write
\[
\hat{\Phi}_{n}(H, b; t) = \omega_{q-1}^{-1} \sum_{\ell=0}^{\infty} (-1)^{t} b_{\ell}H \left( \frac{\ell}{n} \right) p_{2\ell}^{(\alpha,\alpha)}(1) p_{2\ell}^{(\alpha,\alpha)}(t), \quad t \in [-1,1].
\] (4.12)
We set $g(t) = h(t) - h(2t)$, and note that for all $n \geq 1$, $t \in [0,\infty)$,
\[
h(t) + \sum_{m=1}^{n} g \left( \frac{t}{2m} \right) = h \left( \frac{t}{2^{m}} \right), \quad h \left( \frac{t}{2^{m}} \right) + \sum_{m=n+1}^{\infty} g \left( \frac{t}{2^{m}} \right) = 1.
\] (4.13)

**Corollary 4.1** Let $s > \alpha + 3/2$, and $b \in \mathcal{B}(s)$. Then there exists a continuous, even function $\phi : [-1,1] \to \mathbb{R}$ such that
\[
\int_{-1}^{1} p_{2t}^{(\alpha,\alpha)}(t) \phi(t)(1-t^2)^{\alpha} dt = (-1)^{t} b_{2t}p_{2t}^{(\alpha,\alpha)}(1),
\] (4.14)
and
\[
\left\| \phi - \hat{\Phi}_{2n}(H, b; \cdot) \right\|_{\infty,[-1,1]} \leq c2^{-n(s-\alpha-3/2)}.
\] (4.15)

**Proof.**

In view of (4.9), we see that for $n \geq 0$,
\[
\sum_{j=n+1}^{\infty} \left\| \hat{\Phi}_{2^{j}}(g, b; \cdot) \right\|_{\infty,[-1,1]} \leq c \sum_{j=n+1}^{\infty} 2^{-j(s-\alpha-3/2)} < c2^{-n(s-\alpha-3/2)}.
\]
Using (4.13) (with $j$ in place of $t$), this leads easily to (4.15), and hence, to the rest of the assertions of this corollary.
5 Proof of of Theorem 3.1

The proof of Theorem 3.1 mimics that of [20, Theorem 3.1, equivalently, Theorem 6.1]. Thus, we first prove the theorem when the target function is a spherical polynomial (Lemma 5.3). The other main step is to obtain a degree of approximation of the actual target function by spherical polynomials in terms of the operator $D_\phi$ (Lemma 5.4). The proof of Theorem 3.1 is then a simple calculation combining the results of these two main steps. A crucial step in these proofs is Proposition 5.1.

We now start this program with two lemmas preparatory for Proposition 5.1.

Lemma 5.1 Let $0 < d \leq 1, \nu \in \mathcal{R}_d, s \in \mathbb{R}, b \in B(s)$. Then for $n \geq 1$,

$$
\int_{S^q} |\tilde{\Phi}_n(g, b; \mathbf{x} \cdot \mathbf{y})| d|\nu|(\mathbf{y}) \leq cn^{-s+(q-1)/2} \|\nu\|_{R_d}, \quad \mathbf{x} \in S^q.
$$

Proof. Since the kernel $\tilde{\Phi}_n$ is an even function of $\mathbf{x}$ and $\mathbf{y}$, there is no loss of generality in assuming that $\nu$ is an even measure, and $\|\nu\|_{R_d} = 1$. In this proof only, let $\mathbf{x} \in S^q_d$ be fixed, and for $r, \rho \geq 0$,

$$
A_{r, \rho} = \{y \in S^q : \mathbf{x} \cdot \mathbf{y} = \cos \theta, \quad r < \pi/2 - \theta \leq \rho, \quad \theta \geq \pi/4\}. \tag{5.2}
$$

We estimate first $|\nu|(A_{r, \rho})$. If $\rho - r > d$, then $A_{r, \rho}$ is covered by $(\rho - r)^{-q} (q-1)$ caps of radius $\sim \rho - r$ (i.e., volume $\sim (\rho - r)^q$) each. Therefore, $|\nu|(A_{r, \rho}) \leq c(\rho - r)$. If $\rho - r \leq d$, then $A_{r, \rho}$ is covered by $d^{-q} (q-1)$ caps of radius $\sim d$ each. Therefore, $|\nu|(A_{r, \rho}) \leq cd$. Thus,

$$
|\nu|(A_{r, \rho}) \leq c(\rho - r) \left(1 + \frac{d}{\rho - r}\right). \tag{5.3}
$$

Next, we recall (4.9) with $(q - 2)/2$ in place of $\alpha$, and $\mathbf{x} \cdot \mathbf{y} = \cos \theta$:

$$
|\hat{\Phi}_n(g, b; \mathbf{x} \cdot \mathbf{y})| \leq c \frac{n^{-s+(q+1)/2}}{\max(1, |n(\pi/2 - \theta)|^q)}. \tag{5.4}
$$

It follows immediately from (5.4) and (5.3) that

$$
\int_{A_{0, \pi/n}} |\hat{\Phi}_n(g, b; \mathbf{x} \cdot \mathbf{y})| d|\nu|(\mathbf{y}) \leq cn^{-s+(q+1)/2}|\nu|(A_{0, \pi/n}) \leq cn^{-s+(q-1)/2} (1 + nd), \tag{5.5}
$$

and

$$
\int_{y \in S^q : \theta \leq \pi/4} |\hat{\Phi}_n(g, b; \mathbf{x} \cdot \mathbf{y})| d|\nu|(\mathbf{y}) \leq cn^{-s+(q+1)/2-S} |\nu|(S^q) \leq cn^{-s+(q+1)/2-S}. \tag{5.6}
$$

Finally, using (5.4) and (5.3) again,

$$
\int_{A_{\pi/n, \pi/2}} |\hat{\Phi}_n(g, b; \mathbf{x} \cdot \mathbf{y})| d|\nu|(\mathbf{y}) \leq \sum_{k=0}^{\infty} |\nu|(A_{\pi/n, 2^{k+1}\pi/n}) \leq cn^{-s+(q+1)/2-S} \sum_{k=0}^{\infty} \frac{2^k |\nu|(A_{\pi/n, 2^k\pi/n})}{\|\nu\|_{R_d}} \leq cn^{-s+(q+1)/2-S} \sum_{k=0}^{\infty} \frac{2^k (S-1) (1 + nd/2^k)}{\|\nu\|_{R_d}} \leq cn^{-s+(q+1)/2-S} \leq cn^{-s+(q-1)/2} (1 + nd).
$$

Together with (5.5) and (5.6), this leads to (5.1). \hfill \Box

The following lemma follows easily from Lemma 5.1.

Lemma 5.2 Let $0 < d \leq 1, \nu \in \mathcal{R}_d, b \in B(s), 1 \leq p \leq \infty, F \in L^p(\nu)$. Then for $n \geq 1$,

$$
\left\| \int_{S^q} \hat{\Phi}_n(g, b; \mathbf{y}) F(\mathbf{y}) d\nu(\mathbf{y}) \right\|_{L^p(\nu)} \leq cn^{-s+(q-1)/2} (1 + nd)^{1/p} \|F\|_{L^p(\nu)}. \tag{5.7}
$$
Proof. The estimate (5.7) is clear from (5.1) for $F \in L^\infty(\nu)$. Applying (5.1) with $\mu^*_q$ in place of $\nu$, (so that we may choose $d = 0$), we get

$$\int_{S^q} |\hat{\Phi}_n(g, b; x \cdot y)| d\mu^*_q(y) \leq cn^{s+(q-1)/2}. \quad (5.8)$$

Using Fubini’s theorem, it is then easy to see that (5.7) holds for $F \in L^1(\nu)$. The general case follows from the Riesz-Thorin interpolation theorem [4, Theorem 1.1.1].

The following proposition is the analogue of [20, Proposition 5.2].

Proposition 5.1 Let $s > (q + 1)/2$, $\phi \in A(s)$, $b = (-1)^i \hat{\phi}(2\ell)$, $n \geq 1$ be an integer, $0 < d \leq 1$, $\nu \in R_d$, $1 \leq p \leq \infty$, $F \in L^p(\nu)$. For $m \geq \log_2(1/d)$, let

$$U_m(\nu; F, x) = \int_{S^q} \left\{ \phi(x \cdot y) - \hat{\Phi}_2^n(h, b; x \cdot y) \right\} F(y) d\nu(y), \quad x \in S^q. \quad (5.9)$$

Then

$$\|U_m(\nu; F)\|_p \leq c 2^{-m(s+1/p-(q+1)/2)}(d\|\nu\|_{R,d})^{1/p'} \|F\|_{\nu,p}. \quad (5.10)$$

Proof. Since $m \geq \log_2(1/d)$, we see that $2^jd \geq 1$ for all integer $j \geq m$. Consequently, (5.7) used with $2^j$ in place of $\nu$ shows that for $j \geq m$,

$$\int_{S^q} |\hat{\Phi}_2^n(g, b; \phi \cdot y) F(y) d\nu(y)|_p \leq c 2^{-j(s-(q-1)/2)} ((1 + 2^j d)\|\nu\|_d)^{1/p'} \|F\|_{\nu,p}.$$ 

Therefore (cf. (4.13)),

$$\|U_m(\nu; F)\|_p \leq \|U_0(\nu; F)\|_p \leq \sum_{j=m+1}^\infty \|U_j(\nu; F)\|_p \leq c 2^{-m(s+1/p-(q+1)/2)}(d\|\nu\|_{R,d})^{1/p'} \|F\|_{\nu,p}. \quad (5.11)$$

Lemma 5.3 Let $n \geq 0$, $P \in \Pi_{2n+1}^q$ be an even polynomial, $\nu$ be an MZ quadrature measure of order $2^{n+2}$, $s > (q + 1)/2$, $\phi \in A(s)$. Then for $1 \leq p \leq \infty$,

$$\|P - \int_{S^q} \phi(\phi \cdot y) d\phi(P)(y) d\nu(y)\|_p \leq c 2^{-n(s-(q-1)/2)}\|\nu\|_{R,1/2^n}^{1/p'} \|D\phi(P)\|_{\nu,p}. \quad (5.12)$$

Proof. In this proof, let $b = (-1)^i \hat{\phi}(2\ell)$, so that $b \in B(s)$. Let $x \in S^q$. Since $\nu$ is an MZ quadrature measure of order $2^{n+2}$, and $D\phi(P), \hat{\Phi}_2^n(h, b; x \cdot y) \in \Pi_{2n+1}^q$, we have

$$\int_{S^q} \hat{\Phi}_2^n(h, b; x \cdot y) d\phi(P)(y) d\nu(y) = \int_{S^q} \hat{\Phi}_2^n(h, b; x \cdot y) D\phi(P)(y) d\mu^*_q(y). \quad \text{Hence, (2.20) implies that}$$

$$P(x) - \int_{S^q} \phi(x \cdot y) D\phi(P)(y) d\nu(y) = \int_{S^q} \phi(x \cdot y) d\mu^*_q(y) - \int_{S^q} \phi(x \cdot y) D\phi(P)(y) d\nu(y) = \int_{S^q} \left\{ \phi(x \cdot y) - \hat{\Phi}_2^n(h, b; x \cdot y) \right\} D\phi(P)(y) d\mu^*_q(y) \quad \text{Thus,}$$

$$\|P - \int_{S^q} \phi(\phi \cdot y) D\phi(P)(y) d\nu(y)\|_p \leq \|U_n(\mu^*_q; D\phi(P), x) - U_n(\nu; D\phi(P), x)\|_p.$$ 

We now use Proposition 5.1 twice, once with $\mu^*_q$ in place of $\nu$, and once with $\nu$, and (in both cases), with $D\phi(P)$ in place of $F$, $2^{-n}$ in place of $d$ to deduce (5.12).
Lemma 5.4 Let $s > (q + 1)/2$, $\phi \in \mathcal{A}(s)$, $1 \leq p \leq \infty$, $f \in W_{q;\phi}^p$, $n \geq 1$ Then
\[
\|f - \sigma_{2^n}(\mu^{*}_q; f)\|_p \leq c 2^{-n(s-(q-1)/2)}\|D\phi(f)\|_p. \tag{5.13}
\]
In the case when $p = \infty$, and $\mu$ is an MZ quadrature measure of order $2^{n+2}$, we have
\[
\|f - \sigma_{2^n}(\mu; f)\|_{\infty} \leq c 2^{-n(s-(q-1)/2)}\|\mu\|_{R,2^{-n}}\|D\phi(f)\|_{\infty}. \tag{5.14}
\]
Proof. In this proof, let $b = (-1)^q \hat{\phi}(2\ell)$, so that $b \in \mathcal{B}(s)$. We note that
\[
\sigma_{2^n}(\mu^{*}_q; f, x) = \int_{\mathcal{S}^n} \phi(x \cdot y)\mathcal{D}\phi(\sigma_{2^n}(\mu^{*}_q; f))(y) d\mu^{*}_q(y) = \int_{\mathcal{S}^n} \phi(x \cdot y)\sigma_{2^n}(\mu^{*}_q; \mathcal{D}\phi(f), y) d\mu^{*}_q(y) = \int_{\mathcal{S}^n} \sigma_{2^n}(\mu^{*}_q; \phi(x \cdot, y))\mathcal{D}\phi(f)(y) d\mu^{*}_q(y) = \int_{\mathcal{S}^n} \Phi_{2^n}(h, b; x \cdot y)\mathcal{D}\phi(f)(y) d\mu^{*}_q(y).
\]
Therefore,
\[
f(x) - \sigma_{2^n}(\mu^{*}_q; f, x) = \int_{\mathcal{S}^n} \{\phi(x \cdot y) - \Phi_{2^n}(h, b; x \cdot y)\} \mathcal{D}\phi(f)(y) d\mu^{*}_q(y) = U_{2^n}(\mu^{*}_q; \mathcal{D}\phi(f)).
\]
We now deduce (5.13) using Proposition 5.1 with $\mu^{*}_q$ in place of $\nu$ and $2^{-n}$ in place of $d$.
Specializing to the case $p = \infty$, Proposition 2.3 shows that
\[
\|f - \sigma_{2^n}(\mu; f)\|_{\infty} \leq c \|\mu\|_{R,2^{-n}} E_{2^n,\infty}(f) \leq c \|\mu\|_{R,2^{-n}} \|f - \sigma_{2^n-z}(\mu^{*}_q; f)\|_{\infty}.
\]
Therefore, (5.13) implies (5.14). \qed

Proof of Theorem 3.1
We use Lemma 5.3 with $P = \sigma_{2^n}(\mu^{*}_q; f)$ in place of $P$, and recall that
\[
\|P\|_p = \|\mathcal{D}\phi(\sigma_{2^n}(\mu^{*}_q; f))\|_p = \|\sigma_{2^n}(\mu^{*}_q; \mathcal{D}\phi(f))\|_p \leq c \|\mathcal{D}\phi(f)\|_p
\]
to conclude that
\[
\|\sigma_{2^n}(\mu^{*}_q; f) - G_{2^n}(\phi, \mu^{*}_q; \nu; f)\|_p = \left\| P - \int_{\mathcal{S}^n} \phi(x \cdot y)\mathcal{D}\phi(P)(y) d\nu(y) \right\|_p \leq c 2^{-n(s-(q-1)/2)}\|\nu\|_{R,1/2^n}^{1/p'} \|\mathcal{D}\phi(P)\|_{\nu;p} \tag{5.15}
\]
In view of Proposition 2.1(c),
\[
\|\mathcal{D}\phi(P)\|_{\nu;p} \leq c \|\nu\|_{R,1/2^n}^{1/p} \|\mathcal{D}\phi(P)\|_p.
\]
Therefore, the estimate (3.5) follows from (5.13) and (5.15). The estimate (3.6) is a simple consequence of Proposition 2.3.
The estimate (3.7) is proved in the same way, using (5.14) instead of (5.13). \qed

A Appendix: Computation of coefficients

The main purpose of this appendix is to prove the following proposition.

Proposition A.1 Let $\alpha > -1$, $\gamma > -1/2$ and $2\gamma + 1$ not be an even integer. Then for $\ell = 0, 1, \cdots$,
\[
\int_{-1}^1 |t|^{2\gamma+1} p_{2\ell}^{(\alpha,\gamma)}(t)(1-t^2)^\alpha dt = (-1)^{\ell} \frac{\cos(\pi\gamma)^\alpha(2\gamma+1)!}{\ell+\gamma+\alpha+1!} \frac{(\ell - \gamma - 3/2)!}{\ell+\gamma+\alpha+1!} P_{2\ell}^{(\alpha,\gamma)}(1) = \cos(\pi)^{2\gamma+1} \frac{(2\gamma+1)!}{2\gamma+1} \frac{(\ell + \alpha)!}{(\ell - 1/2)!} P_{2\ell}^{(\alpha,\gamma)}(0). \tag{A.1}
\]
In particular, the following asymptotic expansions hold with real constants $c_{1,j}, c_{2,j}$:
\[
\int_{-1}^1 |t|^{2\gamma+1} p_{2\ell}^{(\alpha,\gamma)}(t)(1-t^2)^\alpha dt = (-1)^{\ell} \frac{p_{2\ell}^{(\alpha,\gamma)}(1)}{\ell+2\gamma+1/2} \sum_{j=0}^\infty c_{1,j}(\gamma, \alpha) \ell^j = \frac{p_{2\ell}^{(\alpha,\gamma)}(0)}{\ell+2\gamma+2} \sum_{j=0}^\infty c_{2,j}(\gamma, \alpha) \ell^j. \tag{A.2}
\]
In preparation for the proof, we make first some observations. First, we note that the duplication and complimentary formulas for gamma functions take the form

\[
\frac{(2z)!}{z!} = 2^{2z}(z - 1/2)! \quad \text{and} \quad \frac{1}{(-z)!} = \frac{\sin(\pi z)}{\pi z} \frac{1}{z!} = \frac{\sin(\pi z)}{\pi} (z - 1)!, \quad z \text{ not a negative integer},
\]

respectively.

In view of (4.1) and (4.3) we obtain

\[
(1 - x^2)\alpha I_{2\ell}^{(\alpha, \alpha)}(x) = p_2^{(\alpha, \alpha)}(1) \frac{\alpha!}{2^{2\ell}(2\ell + \alpha)! \alpha!} \frac{d^{2\ell}}{dx^{2\ell}} (1 - x^2)^{2\ell + \alpha}.
\]

Further, since

\[
(1 - x^2)^{2\ell + \alpha} = \sum_{j=0}^{\infty} \frac{(2\ell + \alpha)!}{j!(2\ell + \alpha - j)!} (-1)^j x^{2j}, \quad |x| < 1,
\]

we deduce that for integer \( m \geq 0 \),

\[
\left. \frac{d^{2m}}{dx^{2m}} (1 - x^2)^{2\ell + \alpha} \right|_{x=0} = (-1)^m \frac{(2\ell + \alpha)! (2m)!}{m! (2\ell + \alpha - m)!}.
\]

In particular, (4.1) shows that

\[
p_2^{(\alpha, \alpha)}(0) = (-1)^{\ell} \frac{(2\ell + \alpha)!}{2^{2\ell} (\ell + \alpha)!} \left\{ \frac{4\ell + \alpha + \beta + 1}{2^{\alpha + \beta + 1}} \frac{(2\ell)! (2\ell + \alpha + \beta)!}{(\ell + \alpha)! (2\ell + \beta)!} \right\}^{1/2}.
\]

Hence, using (A.3) and (4.3), we see that

\[
\frac{p_2^{(\alpha, \alpha)}(1)}{p_2^{(\alpha, \alpha)}(0)} = (-1)^{\ell} \frac{\sqrt{\pi} (\ell + \alpha)!}{(2\ell)!} = (-1)^{\ell} \frac{\sqrt{\pi} (\ell + \alpha)!}{(\ell - 1/2)!}.
\]

**Proof of Proposition A.1**

In this proof, let

\[
I_{\gamma, \ell} = \int_0^1 x^{2\gamma + 1} \frac{d^{2\ell}}{dx^{2\ell}} (1 - x^2)^{2\ell + \alpha} dx.
\]

We will prove that

\[
I_{\gamma, \ell} = (-1)^{\ell} \cos(\pi\gamma) \frac{(2\gamma + 1)! (2\ell + \alpha)!}{(\ell - \gamma - 1)!(\gamma + \ell + \alpha + 1)!} = (-1)^{\ell} \frac{\cos(\pi\gamma)}{2^{2\ell+2} \sqrt{\pi}} \frac{(2\ell + \alpha)! (\ell - 3/2)!}{(\ell + \gamma + \alpha + 1)!}.
\]

The second equation in (A.9) follows from the first using (A.3). So, we need to prove only the first equation. We distinguish two cases.

**Case I:** \( 2\ell \leq 2\gamma + 1 \), or \( 2\gamma + 1 \) is not an integer.

Integration by parts \( 2\ell \) times in (A.8) gives

\[
I_{\gamma, \ell} = \frac{(2\gamma + 1)!}{(2\gamma + 1 - 2\ell)!} \int_0^1 x^{2\gamma + 1 - 2\ell} (1 - x^2)^{2\ell + \alpha} dx = \frac{(2\gamma + 1)!}{(2\gamma + 1 - 2\ell)!} \int_0^1 y^{\gamma - \ell} (1 - y)^{2\ell + \alpha} dy
\]

where we note that neither \( \gamma - \ell \) nor \( 2\gamma + 1 - 2\ell \) is a negative integer. Using (A.4),

\[
\frac{(\gamma - \ell)!}{(2\gamma + 1 - 2\ell)!} = \frac{(2\ell - 2\gamma - 2)! \sin((2\gamma + 1)\pi)}{(\ell - \gamma - 1)! \sin(\pi(\ell - \gamma))} = 2(-1)^{\ell} \cos(\pi\gamma) \frac{(2\ell - 2\gamma - 2)!}{(\ell - \gamma - 1)!}.
\]
Together with (A.10), this leads to (A.9).

**Case II:** $2\gamma + 1$ is an odd integer (i.e., $\gamma$ is an integer), and $2\ell > 2\gamma + 1$.

Let $m = \ell - \gamma - 1$. Then $m \geq 0$ is an integer. Using Leibnitz rule, we find that

$$\frac{d^{2m}}{dx^{2m}}(1 - x^2)^{2\ell + \alpha} \bigg|_{x=1} = 0.$$ 

Therefore, an integration by parts $2\gamma + 1$ times gives (cf. (A.6))

$$I_{\gamma,\ell} = (-1)^{2\gamma+1}(2\gamma + 1)! \int_0^1 \frac{d^{2\ell-2\gamma-1}x}{dx^{2\ell-2\gamma-1}}(1 - x^2)^{2\ell + \alpha} dx = (-1)^{\ell + \gamma} \frac{(2\gamma + 1)!(2\ell + \alpha)!(2m)!}{m!(2\ell + \alpha - m)!} = (-1)^{\ell + \gamma} \frac{(2\gamma + 1)!(2\ell + \alpha)!(2\ell - 2\gamma - 2)!}{(\ell - \gamma - 1)!(\ell + \alpha + \gamma + 1)!} = (-1)^{\ell} \frac{\cos(\pi\gamma)}{(\ell - \gamma - 1)!(\gamma + \ell + \alpha + 1)!}. \quad (A.11)$$

This completes the proof of (A.9) also in this case.

Having proved (A.9), the first identity in (A.1) now follows from (A.5). The second identity in (A.1) follows from the first and (A.7). The expansions (A.2) follow from (A.1) and the asymptotic formula for the ratios of gamma functions [27, Chapter 4, formula (5.02)].

### References

[1] F. Bach. Breaking the curse of dimensionality with convex neural networks. *arXiv preprint arXiv:1412.8690*, 2014.

[2] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *Information Theory, IEEE Transactions on*, 39(3):930–945, 1993.

[3] H. Bateman, A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi. *Higher transcendental functions*, volume 2. McGraw-Hill New York, 1955.

[4] J. Bergh and J. Löfström. *Interpolation spaces: an introduction*, volume 223. Springer Berlin, 1976.

[5] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[6] G. Cybenko. Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[7] F. Filbir and H. N. Mhaskar. Marcinkiewicz–Zygmund measures on manifolds. *Journal of Complexity*, 27(6):568–596, 2011.

[8] K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

[9] B. Irie and S. Miyake. Capabilities of three-layered perceptrons. In *Neural Networks, 1988., IEEE International Conference on*, pages 641–648. IEEE, 1988.

[10] Q. T. Le Gia and H. N. Mhaskar. Polynomial operators and local approximation of solutions of pseudo-differential equations on the sphere. *Numerische Mathematik*, 103(2):299–322, 2006.

[11] Q. T. Le Gia and H. N. Mhaskar. Localized linear polynomial operators and quadrature formulas on the sphere. *SIAM Journal on Numerical Analysis*, pages 440–466, 2008.

[12] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[13] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[14] H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1):164–177, 1996.
[15] H. N. Mhaskar. Approximation theory and neural networks. In *Wavelet Analysis and Applications, Proceedings of the international workshop in Delhi*, pages 247–289, 1999.

[16] H. N. Mhaskar. On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity*, 20(4):561–590, 2004.

[17] H. N. Mhaskar. Polynomial operators and local smoothness classes on the unit interval. *Journal of Approximation Theory*, 131(2):243–267, 2004.

[18] H. N. Mhaskar. On the representation of smooth functions on the sphere using finitely many bits. *Applied and Computational Harmonic Analysis*, 18(3):215–233, 2005.

[19] H. N. Mhaskar. Weighted quadrature formulas and approximation by zonal function networks on the sphere. *Journal of Complexity*, 22(3):348–370, 2006.

[20] H. N. Mhaskar. Eignets for function approximation on manifolds. *Applied and Computational Harmonic Analysis*, 29(1):63–87, 2010.

[21] H. N. Mhaskar and C. A. Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied mathematics*, 13(3):350–373, 1992.

[22] H. N. Mhaskar and C. A. Micchelli. Dimension-independent bounds on the degree of approximation by neural networks. *IBM Journal of Research and Development*, 38(3):277–284, 1994.

[23] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward. Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature. *Mathematics of computation*, 70(235):1113–1130, 2001.

[24] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[25] M. Minsky and S. Papert. Perceptrons: an introduction to computational geometry (expanded edition), 1988.

[26] C. Müller. *Spherical harmonics*, volume 17. Springer, 2006.

[27] F. W. Olver. *Asymptotics and special functions*. Academic press, 2014.

[28] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[29] E. M. Stein and G. Weiss. *Introduction to Fourier analysis on Euclidean spaces (PMS-32)*, volume 32. Princeton university press, 2016.

[30] M. Stinchcombe and H. White. Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, pages 613–617. IEEE, 1989.

[31] G. Szegö. Orthogonal polynomials. In *Colloquium publications/American mathematical society*, volume 23. Providence, 1975.

[32] A. F. Timan. *Theory of Approximation of Functions of a Real Variable: International Series of Monographs on Pure and Applied Mathematics*, volume 34. Elsevier, 2014.