Appl. Comput. Harmon. Anal. 58 (2022) 1–26

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions

Jonathan W. Siegel *, Jinchao Xu

Department of Mathematics, Pennsylvania State University, McAllister Building, University Park, 16802, PA, USA

ARTICLE INFO

Article history:
Received 24 April 2021
Received in revised form 10 December 2021
Accepted 16 December 2021
Available online 21 December 2021
Communicated by Vera Kurkova

Keywords:
Neural networks
Approximation rates
Approximation lower bounds
Finite element methods

ABSTRACT

We study the approximation properties of shallow neural networks with an activation function which is a power of the rectified linear unit. Specifically, we consider the dependence of the approximation rate on the dimension and the smoothness in the spectral Barron space of the underlying function f to be approximated. We show that as the smoothness index s of f increases, shallow neural networks with ReLU^k activation function obtain an improved approximation rate up to a best possible rate of $O(n^{-(k+1)}\log(n))$ in L^2, independent of the dimension d. The significance of this result is that the activation function ReLU^k is fixed independent of the dimension, while for classical methods the degree of polynomial approximation or the smoothness of the wavelets used would have to increase in order to take advantage of the dimension dependent smoothness of f. In addition, we derive improved approximation rates for shallow neural networks with cosine activation function on the spectral Barron space. Finally, we prove lower bounds showing that the approximation rates attained are optimal under the given assumptions.

1. Introduction

We consider approximating a function $f : \Omega \to \mathbb{C}$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain, by a superposition of ridge functions of the form

$$f_n(x) = \sum_{i=1}^{n} a_i \sigma(\omega_i \cdot x + b_i), \tag{1}$$

with activation function $\sigma = \cos(x)$ or $\sigma = [\max(0, x)]^k$ for $k \geq 0$. The latter case corresponds to a neural network with a single hidden layer and activation function given by a power of a rectified linear unit [42], or

* Corresponding author.
E-mail addresses: jus1949@psu.edu (J.W. Siegel), jxx1@psu.edu (J. Xu).

https://doi.org/10.1016/j.acha.2021.12.005
1063-5203/

rectified power unit [35]. In the case where $k = 0$, $[\max(0, x)]^0$ is interpreted as the Heaviside function and we may also use any sigmoidal activation function in its place [4]. Approximation by functions of the form (1) has received a significant amount of attention in the literature. For instance, it has been shown that as long as σ is not a polynomial, functions of the form (1) are dense in $C(\Omega)$ [2,34,20] and $C^k(\Omega)$ [23,24], and in [13,12] explicit operators realizing the approximation for any absolutely continuous f are constructed.

Beyond the problem of density, we are interested in determining rates of approximation for a given class of functions $f : \Omega \to \mathbb{C}$ by an expression of the form (1) with respect to the Sobolev norm $H^m(\Omega)$. Typical results of this type consider functions f which either satisfy classical smoothness assumptions, such as membership in a suitable Sobolev space, or non-standard smoothness assumptions coming from the theory of non-linear approximation by a dictionary.

For example, results for functions f in high-order Sobolev spaces have been obtained in [44]. Here it is shown that if the activation function σ achieves approximation order $O(n^{-r})$ for one-dimensional functions $f \in H^r([0, 1])$, then an approximation rate of $O(n^{-\frac{1}{2}-\frac{2r-1}{2d}})$ can be attained for high dimensional functions $f \in H^{\frac{d-1}{2}+r}(B^d)$ on the unit ball with respect to L^2. In [39] the case where each term in (1) may have a different profile σ is considered and the optimal rates in this case are derived for all Sobolev spaces H^r in $\mathbb{R}^d$. This result is generalized in [22,40] to general L^p spaces.

A typical example of a non-standard smoothness assumption is membership in the closed convex hull of a suitable bounded dictionary $\mathbb{D} \subset H^m(\Omega)$, specifically one considers

$$f \in B_1(\mathbb{D}) = \overline{\left\{ \sum_{i=1}^{n} a_i d_i,\ d_i \in \mathbb{D},\ n \in \mathbb{N},\ \sum_{i=1}^{n} |a_i| = 1 \right\}}. \tag{2}$$

One can also characterize this set using the guage norm (see, for instance [46]) of $B_1(\mathbb{D})$, which we denote by $\mathscr{K}_1(\mathbb{D})$ (following the notation from [16])

$$\|f\|_{\mathscr{K}_1(\mathbb{D})} = \inf\{c > 0 :\ f \in cB_1(\mathbb{D})\}. \tag{3}$$

Here $B_1(\mathbb{D})$ is exactly the unit ball in the norm $\mathscr{K}_1(\mathbb{D})$. For this class of functions, one considers non-linear approximation by finite dictionary expansions, i.e. approximation form the set

$$\Sigma_n(\mathbb{D}) = \left\{ \sum_{i=1}^{n} a_i d_i,\ d_i \in \mathbb{D} \right\}. \tag{4}$$

When the dictionary $\mathbb{D}$ is of the form

$$\mathbb{D}_\sigma = \{\sigma(\omega \cdot x + b),\ \omega \in \mathbb{R}^d,\ b \in \mathbb{R}\}, \tag{5}$$

this exactly corresponds to an expansion of the form (1). For some activation functions, such as $[\max(0, x)]^k$ for $k > 0$, the dictionary $\mathbb{D}_\sigma$ is not bounded. In this case, the dictionary must be modified (details can be found in [48,50]) and the resulting space, called the Barron space [19] or variation space corresponding to shallow ReLU networks [3] when $k = 1$, is closely related to the ridgelet spaces introduced in [11] (to be precise, it is sandwiched between $R^{1+k+(d-1)/2}_{1,1}$ and $R^{1+k+(d-1)/2}_{1,\infty}$, see section 4.2 in [11]).

Using a classical probabilistic argument of Maurey [45], an approximation rate of $O(n^{-\frac{1}{2}})$ can be obtained for the class $B_1(\mathbb{D})$ using non-linear dictionary expansions. Moreover, Jones [27] gave a constructive proof of this fact using the relaxed greedy algorithm and applied this result to shallow neural networks with a cosine activation function. Improvements upon this rate of dictionary approximation under an assumption about the behavior of the relaxed greedy algorithm appear in [31,33]. These results yield exponential rates of convergence for individual functions in the convex hull of $\mathbb{D}$ (but not necessarily its closure), which

are however not uniform over the class $B_1(\mathbb{D})$. Further, under compactness [41,29] or smoothness [50] assumptions on the dictionary $\mathbb{D}$ improved rates can also be obtained, although for general dictionaries the Maurey-Jones rate is the best one can expect [30].

The application of the Jones-Maurey result to neural networks with sigmoidal activation function is due to Barron [4]. In this work, the relevant class of functions is

$$\mathscr{B}^s(\Omega) = \left\{ f : \Omega \to \mathbb{C} : \ \|f\|_{\mathscr{B}^s(\Omega)} := \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} (1 + |\xi|)^s |\hat{f}_e(\xi)| d\xi < \infty \right\}, \tag{6}$$

where the infimum above is over extensions $f_e \in L^1(\mathbb{R}^d)$. Barron introduced this class for $s = 1$ and showed that for a sigmoidal activation function σ we have $\mathscr{K}_1(\mathbb{D}_\sigma) \supset \mathscr{B}^1(\Omega)$. This shows that shallow neural networks with sigmoidal activation function can approximate functions satisfying a certain Fourier integrability condition with a rate of $O(n^{-\frac{1}{2}})$. This convergence rate for $f \in \mathscr{B}^1(\Omega)$ has been extended to a very general class of activation functions in [25,47]. A variety of other results for functions in the spectral Barron space (6) have been obtained in [29,9,19,51,41], for instance.

It has been shown that the space $\mathscr{B}^s(\Omega)$ is exactly equivalent (with identical norm) to $B_1(\mathbb{F}_s^d)$ [48] for the dictionary

$$\mathbb{F}_s^d = \{(1 + |\omega|)^{-s} e^{2\pi i \omega \cdot x} : \ \omega \in \mathbb{R}^d\} \tag{7}$$

of decaying Fourier modes. There is a slight subtlety here. Let $\Omega = [0,1]^d$, which is the case we are primarily interested in. Note that the frequency $\omega \in \mathbb{R}^d$ is allowed to be arbitrary. This results in a larger space than restricting $\omega \in \mathbb{Z}^d$, i.e. considering the dictionary

$$\overline{\mathbb{F}_s^d} = \{(1 + |\omega|)^{-s} e^{2\pi i \omega \cdot x} : \ \omega \in \mathbb{Z}^d\}, \tag{8}$$

which is done for instance in [38]. Restricting ω to the integer lattice $\mathbb{Z}^d$ results in a much simpler space, since we have (see [11], section 7.2)

$$\|f\|_{\mathscr{K}_1(\overline{\mathbb{F}_s^d})} \eqsim \sum_{n \in \mathbb{Z}^d} (1 + |n|)^s |\hat{f}(n)|. \tag{9}$$

However, by considering a pure frequency with $\omega \notin \mathbb{Z}^d$ and calculating its Fourier series, we can easily see that the above characterization does not hold for $\mathscr{B}^s(\Omega)$ and that the space $\mathscr{B}^s(\Omega)$ is strictly larger.

An interesting fact, first observed by Makovoz [41], is that for certain activation functions σ, the rate of approximation $O(n^{-\frac{1}{2}})$ derived by Barron [4] can be improved. In particular, Makovoz shows that for the Heaviside activation function

$$\sigma = [\max(0, x)]^0 := \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases} \tag{10}$$

the rate can be improved to $O(n^{-\frac{1}{2} - \frac{1}{2d}})$. Furthermore, when $\sigma = \max(0, x)^k$ for $k \geq 1$, this can be improved to a rate of $O(n^{-\frac{1}{2} - \frac{2k+1}{2d}})$ [29,51,50] for the class $\mathscr{B}^{k+1}(\Omega)$. For functions $f \in \mathscr{B}^1(\Omega)$, improved rates have also been obtained for more general activation functions [47]. It has also been shown that the orthogonal greedy algorithm [43] can constructively obtain such improved approximation rates [49].

In this work, we study how much further the rate of $O(n^{-\frac{1}{2}})$ can be improved given stronger assumptions on the smoothness of f, i.e. assuming that $f \in \mathscr{B}^s(\Omega)$ for larger values of s. By showing that the continuous Fourier transform of a function f on the bounded set Ω can be replaced by a suitably chosen Fourier series,

we show in Theorems 1 and 2 that the approximation rates in L^2 with $\sigma = \cos(x)$ can be improved to $O(n^{-\frac{1}{2}-\frac{s}{d}})$ for $f \in \mathscr{B}^s(\Omega)$ with increasing s, and even that exponential convergence can be attained if the Fourier transform of f decays rapidly enough.

To compare with the results in [44], we observe that $\sigma = \cos(x)$ attains approximation order $O(n^{-r})$ in $H^r([0,1])$ for any $r > 0$. This means that a rate of approximation of $O(n^{-\frac{1}{2}-\frac{s}{d}})$ for cosine networks already appear in [44] for the Sobolev spaces $H^{\frac{d}{2}+s}(\Omega)$. However, our results apply to the spectral Barron space, which is not quite comparable, although we have $H^{\frac{d}{2}+s+\varepsilon}(\Omega) \subset \mathscr{B}^s(\Omega)$ (see [51] Lemma 2.5, for instance).

Further, comparing with approximation by ridgelets [11], we see from the results of Barron [4] and section 4.2 in [11] that the space $B^s(\Omega)$ is contained in $R^{1+k+(d-1)/2}_{1,\infty}$ if $s \geq k+1$. Thus, using ridgelets one can obtain a rate of $O(n^{-\frac{1}{2}-\frac{2s-1}{2d}})$, which is not quite as good as the rate attainable using cosine networks. We wish to emphasize that this approximation rate is not entirely trivial since we are considering the full spectral Barron space $B^s(\Omega)$ and not the space with frequencies restricted to a lattice, as in [38,11].

Next, we consider the problem with activation function $\sigma(x) = \max(0,x)^k$. In Theorems 3 and 4 we show that in this case the convergence rate in $H^m(\Omega)$ can also be continuously improved with increasing s, up to the limit of

$$\|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}^s} n^{m-(k+1)} \log n, \tag{11}$$

which is achieved when $s > (d+1)(k-m+\frac{1}{2}) + m + \frac{1}{2}$. This maximal rate is the same as the rate attained by piecewise polynomials of degree k in dimension $d = 1$, and is significantly higher than the maximal rate obtained in [44] in this case. In particular, this shows that regardless of the dimension d, for smooth enough functions f (here the necessary amount of smoothness depends upon d), neural networks of the form (1) attain approximation rates which match the best possible rates in one dimension.

High-order approximation rates for deeper networks have been studied in [52–54], however these results sometimes involve architectures which depend on the desired accuracy or even the function to be approximated. A theory of approximation by deeper networks in one dimension has also been developed in [15]. In addition, approximation by deep networks with ReLU^k, or RePU, activation function has been studied in [35]. In contrast, our results apply already to shallow networks and show that high order approximation rates can be obtained for a class of sufficiently smooth functions even in high dimensions. A further interesting consequence, which is collected in Theorem 5, is that the approximation results obtained for sigmoidal activation functions by Barron [4] actually hold under weaker regularity conditions on the function f. In particular, instead of $f \in \mathscr{B}^1(\Omega)$ we only need $f \in \mathscr{B}^{\frac{1}{2}}(\Omega)$, although this comes at the cost of a constant which depends exponentially upon the dimension.

Compared with other results in the literature, the main significance of our results is that the smoothness of the activation function is fixed independently of the dimension. In particular, if the target class of functions is very smooth, say $f \in H^k(\Omega)$ with k growing linearly in the dimension d, then finite elements [8] of a sufficiently high degree or wavelets with a sufficiently high degree of smoothness can attain approximation rates which do not depend upon d (since k grows with d) [16,14]. For a space of functions perhaps more comparable to the spectral Barron class, we may also consider the ridgelet space $R^s_{p,q}$ for large s depending linearly upon the dimension. In this case as well, dimension independent (again since s depends upon d) can be obtained using ridgelets [11].

Another method which is particularly effective for high-dimensional problems is the sparse grids method [10]. The sparse grid method approximates functions f from the class

$$\|f\|^2_{H^{k+1}_{mix}} := \sum_{|\alpha|_\infty \leq k+1} \int_\Omega \left| \frac{\partial^{|\alpha|_1}}{\partial \alpha_1 \cdots \partial \alpha_n} f(x) \right|^2 dx < \infty \tag{12}$$

by linear combinations of tensor products of piecewise degree k polynomials on one-dimensional grids with different resolutions in each coordinate direction. As such, the sparse grid method approximates f with a piecewise polynomial function of degree kd. Using this method, an approximation rate in $L^2(\Omega)$ of

$$\|f - f_n\|_{L^2(\Omega)} \lesssim n^{-(k+1)}(\log n)^{(k+2)(d-1)} \tag{13}$$

can be attained, where n is the number of degrees of freedom [10]. Thus, using polynomials of degree kd, the sparse grids method is able to attain similar approximation rates as shallow neural networks under high order smoothness assumptions. One additional potential advantage of the shallow networks is that the spectral Barron space is isotropic, while the sparse grids space is not.

However, in all of these examples the degree of the polynomials or the smoothness of the wavelets and ridgelets must grow with the dimension, which may be inconvenient, since for instance constructing conforming finite element spaces of high degree is a difficult problem [32,55,1,7]. In contrast, for our results the activation function σ is fixed independently of the dimension and so we are able to achieve these approximation rates using piecewise polynomials of a fixed degree. The reason this is possible is that the number of components of the piecewise polynomial represented by an expression of the form (1) grows as n^d (the number of components obtained by cutting $\mathbb{R}^d$ by n hyperplanes) while the number of parameters grows as dn. This suggests that perhaps (shallow) neural networks are particularly effective at approximating highly smooth functions in high dimensions.

This is the sense in which our results show that shallow neural networks overcome the "curse of dimensionality". It is by enabling the use of fixed degree polynomials even in high dimensions. The space of functions which we approximate, $\mathscr{B}^s(\Omega)$, is indeed very small and consists of highly smooth functions.

However, in high dimensions d the metric entropy of classical smoothness spaces with smoothness degree s decays very slowly, specifically like $O(n^{-\frac{s}{d}})$ (see for instance [37], chapter 15). Consequently this class of functions fundamentally cannot be approximated efficiently in high-dimensions. To overcome this, many learning and approximation methods consider classes of functions whose smoothness either grows with dimension or takes a non-standard form. Examples of such spaces include the sparse grids spaces H^{k+1}_{mix}, the reproducting kernel spaces associated with kernel methods [5], the ridgelet spaces [11], and convex hulls of dictionaries. In this sense, the curse of dimensionality is overcome in a similar manner to other methods, namely by considering sufficiently smooth functions.

The preceding analysis suggests that the adaptive nature of the grid underlying a ReLU^k network allows for a significantly better approximation rate than existing linear methods based on fixed grids. This suggests the potential of using such spaces for the solution of differential equations, which has been investigated in [51]. In light of these results, the space of functions represented by shallow ReLU^k networks may also be useful in understanding non-linear approximation by piecewise polynomials more generally, which has been a challenging problem [16].

The paper is organized as follows. In the next section, we consider approximation by networks with a cosine activation function. In addition to being of independent interest, key results in this section concerning the representation of functions in $\mathscr{B}^s(\Omega)$ will be used throughout the paper. Then, in section 3 we prove the main results concerning approximation by ReLU^k networks. In section 4 we show that the results obtained in the previous sections are optimal. Finally, we give concluding remarks and further research directions.

2. Approximation rates for cosine networks

To begin, we remark that throughout this manuscript, we use the following convention for the Fourier transform

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i \xi \cdot x} dx, \tag{14}$$

for which the inverse transform is given by

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi) e^{2\pi i \xi \cdot x} d\xi. \tag{15}$$

We find that this convention results in the cleanest arguments, avoiding the necessity to keep track of normalizing constants.

In this section, we analyze the approximation properties of networks with a cosine activation function on the spectral Barron space $\mathscr{B}^s(\Omega)$. Specifically, consider approximating a function $f \in \mathscr{B}^s(\Omega)$ by a superposition of finitely many complex exponentials with coefficients that are bounded in ℓ^1, i.e. by an element of the set

$$\Sigma_{n,M} = \left\{ \sum_{j=1}^{n} a_j e^{2\pi i \theta_j \cdot x} : \ \theta_j \in \mathbb{R}^d, \ a_j \in \mathbb{C}, \ \sum_{i=1}^{n} |a_i| \leq M \right\}. \tag{16}$$

Alternatively, one can view this as the set of neural networks with a single hidden layer containing n neurons with activation function $\sigma(x) = e^{2\pi i x}$, whose weights are bounded in ℓ^1.

Equivalently, we can consider approximation by networks with a cosine activation function

$$\Sigma_{n,M}^{\cos} = \left\{ \sum_{i=1}^{n} a_i \cos(2\pi \theta_i \cdot x + b_i) : \ \theta_i \in \mathbb{R}^d, \ b_i \in \mathbb{R}, \ \sum_{i=1}^{n} |a_i| \leq M \right\}. \tag{17}$$

This is because

$$e^{2\pi i \theta_i \cdot x} = \cos(2\pi \theta_i \cdot x) + i \cos\left(2\pi \theta_i \cdot x - \frac{\pi}{2}\right) \in \Sigma_{2,2}^{\cos} \tag{18}$$

and

$$\cos(2\pi \theta_i \cdot x) = \frac{1}{2} e^{2\pi i \theta_i \cdot x} + \frac{1}{2} e^{-2\pi i \theta_i \cdot x} \in \Sigma_{2,1}^{d}. \tag{19}$$

Thus we have $\Sigma_{n,M} \subset \Sigma_{2n,2M}^{\cos}$ and $\Sigma_{n,M}^{\cos} \subset \Sigma_{2n,M}$ and so the rates obtained for both sets will be the same. In what follows, we consider $\Sigma_{n,M}$ for convenience in dealing with the Fourier transform.

Let us first state the following simple result that can be obtained by following a calculation given in Section 3 of [26].

Lemma 1. *Given $\alpha > 1$, consider*

$$g(t) = \begin{cases} e^{-(1-t^2)^{1-\alpha}} & t \in (-1, 1) \\ 0 & \text{otherwise}, \end{cases} \tag{20}$$

then there is a constant c_α such that

$$|\hat{g}(\xi)| \lesssim e^{-c_\alpha |\xi|^{1-\alpha^{-1}}}, \tag{21}$$

We begin with a key lemma showing that we only need frequencies lying on a lattice to represent functions f with decaying Fourier transform on a bounded set.

Lemma 2. *Let $\Omega = [0, 1]^d$ and $\mu : \mathbb{R}^d \to \mathbb{R}_+$ be a continuous weight function. Suppose that μ satisfies the following conditions*

- $\mu(\xi+\omega) \leq \mu(\xi)\mu(\omega)$
- *There exists a* $0 < \beta < 1$ *and a* $c > 0$ *such that* $\mu(\xi) \lesssim e^{c|\xi|^\beta}$.

Suppose that f *satisfies*

$$\int_{\mathbb{R}^d} \mu(\xi)|\hat{f}(\xi)|d\xi = C_f < \infty. \tag{22}$$

Then for any $L > 1$, *there exists an* $a \in L^{-1}[0,1]^d$ *(which may depend on* f*) and coefficients* c_ξ, *such that for* $x \in \Omega$

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} c_\xi e^{2\pi i(a+\xi)\cdot x} \tag{23}$$

and

$$\sum_{\xi \in L^{-1}\mathbb{Z}^d} \mu(a+\xi)|c_\xi| \lesssim C_f. \tag{24}$$

Note that the suppressed constant in the above lemma only depends upon d, μ and L, but not on f or a. Furthermore, we note that from the proof below it follows that the suppressed constant depends exponentially on the dimension d.

Proof. Since by assumption $L > 1$, there exists an ε such that $\Omega \subset [0, L-2\varepsilon]^d$. We begin by constructing a cutoff function ϕ_Ω, which is identically 1 on Ω and 0 outside of $[-\varepsilon, L-\varepsilon]^d$. It will be important that the Fourier transform $\hat{\phi}_\Omega$ has sufficiently fast decay, so that

$$\int_{\mathbb{R}^d} \mu(\xi)|\hat{\phi}_\Omega(\xi)| < \infty. \tag{25}$$

To construct this function, we follow closely the calculation made in [26]. Choose $\alpha > 1$ such that $\beta < 1 - \alpha^{-1} < 1$ and consider the smooth one-dimensional bump function g by (20). Let g_d denote the n-dimensional function

$$g_d(x) = \frac{1}{C}\prod_{i=1}^d g(x_i), \tag{26}$$

where the normalization constant C is chosen so that $\int_{\mathbb{R}^d} g_d(x) = 1$. Then by (21) we see that

$$|\hat{g}_d(\xi)| \lesssim e^{-c_\alpha \sum_{i=1}^d |\xi_i|^{1-\alpha^{-1}}} \lesssim e^{-c_{\alpha,d}\alpha|\xi|^{1-\alpha^{-1}}}, \tag{27}$$

for a new constant $c_{\alpha,d}$.

Finally, let $\Omega' = [-\frac{\varepsilon}{2}, L - \frac{3\varepsilon}{2}]^d$ and define

$$\phi_\Omega = \left(4^d\varepsilon^{-d} g_d\left(4\varepsilon^{-1}x\right)\right) * \chi_{\Omega'}(x). \tag{28}$$

The compact support and normalization of g_d implies that $\phi_\Omega|_\Omega = 1$ and $\phi_\Omega = 0$ outside of $[-\varepsilon, L-\varepsilon]^d$. Furthermore, we calculate

$$|\hat{\phi}_\Omega(\xi)| = \left|\hat{g}_d\left(\frac{\varepsilon}{4}\xi\right)\widehat{\chi}_{\Omega'}\right| \lesssim e^{-c_{\alpha,\Omega}|\xi|^{1-\alpha^{-1}}} \tag{29}$$

for a constant $c_{\alpha,\Omega}$, since $\widehat{\chi}_{\Omega'}$ is bounded. The growth condition on μ, combined with $\beta < 1 - \alpha^{-1} < 1$ means that

$$\int_{\mathbb{R}^d} \mu(\xi)|\hat{\phi}_\Omega(\xi)| < \infty. \tag{30}$$

Now consider the function $h_f = \phi_\Omega f$. Evidently $h_f = f$ on Ω and h_f is supported on $[-\varepsilon, L-\varepsilon]^d$. Notice further that $\hat{h}_f = \hat{\phi}_\Omega * \hat{f}$ and we calculate

$$\begin{aligned}\int_{\mathbb{R}^d} \mu(\xi)|\hat{h}_f(\xi)|d\xi &\leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mu(\xi)|\hat{\phi}_\Omega(\xi-\omega)||\hat{f}(\omega)|d\omega d\xi \\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \mu(\xi+\omega)|\hat{\phi}_\Omega(\xi)|d\xi\right)|\hat{f}(\omega)|d\omega.\end{aligned} \tag{31}$$

Now $\mu(\xi+\omega) \leq \mu(\xi)\mu(\omega)$, so that we get

$$\int_{\mathbb{R}^d} \mu(\xi)|\hat{h}_f(\xi)|d\xi \leq \left(\int_{\mathbb{R}^d} \mu(\xi)|\hat{\phi}_\Omega(\xi)|d\xi\right)\left(\int_{\mathbb{R}^d} \mu(\omega)|\hat{f}(\omega)|d\omega\right) \lesssim C_f, \tag{32}$$

where the implied constant depends the value of the integral in (30).

We now rewrite the integral in (32) as

$$\int_{\mathbb{R}^d} \mu(\xi)|\hat{h}_f(\xi)|d\xi = \int_{[0,L^{-1}]^d}\left(\sum_{\xi\in L^{-1}\mathbb{Z}^d} \mu(a+\xi)|\hat{h}_f(a+\xi)|\right)da \lesssim C_f. \tag{33}$$

Certainly this means that there must exist an $a \in [0, L^{-1}]^d$ (depending on f) such that

$$\left(\sum_{\xi\in L^{-1}\mathbb{Z}^d} \mu(a+\xi)|\hat{h}_f(a+\xi)|\right) \lesssim C_f, \tag{34}$$

where the implied constant only depends upon L and d.

We proceed to apply the Poisson summation formula and the fact that h_f is supported in $[-\varepsilon, L-\varepsilon]^d$ to conclude that for a.e. $x \in \Omega \subset [-\varepsilon, L-\varepsilon]^d$ we have

$$f(x) = h_f(x) = \sum_{\nu\in L\mathbb{Z}^d} h_f(x+\nu)e^{2\pi i a\cdot\nu} = \sum_{\xi\in L^{-1}\mathbb{Z}^d} \hat{h}_f(a+\xi)e^{2\pi i(a+\xi)\cdot x}. \tag{35}$$

Here we have applied the Poisson summation formula to the function $g(\nu) = h_f(x+\nu)e^{2\pi i a\cdot\nu}$, whose Fourier transform is easily seen to be $\hat{g}(\xi) = \hat{h}_f(a+\xi)e^{2\pi i(a+\xi)\cdot x}$.

Setting $c_\xi = \hat{h}_f(a+\xi)$ we obtain the desired result. □

We now apply Lemma 2 with $\mu(\xi) = (1+|\xi|)^s$ to obtain the following corollary concerning the spectral Barron space $\mathscr{B}^s(\Omega)$.

Corollary 1. *Let $\Omega = [0,1]^d$ and $s \geq 0$. Let $f \in \mathscr{B}^s(\Omega)$. Then for any $L > 1$, there exists an $a \in L^{-1}[0,1]^d$ (potentially depending upon f) and coefficients c_ξ such that for $x \in \Omega$*

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} c_\xi e^{2\pi \mathrm{i}(a+\xi)\cdot x} \tag{36}$$

and

$$\sum_{\xi \in L^{-1}\mathbb{Z}^d} (1 + |a+\xi|)^s |c_\xi| \lesssim \|f\|_{\mathscr{B}^s(\Omega)}. \tag{37}$$

Proof. This follows immediately from Lemma 2 given the characterization of $\mathscr{B}^s(\Omega)$ and the elementary fact that $(1 + |\xi + \omega|) \leq (1 + |\xi| + |\omega|) \leq (1 + |\xi|)(1 + |\omega|)$. □

Corollary 1 can be used to improve upon the $O(n^{-\frac{1}{2}})$ approximation rate of cosine networks obtained in [27] when $f \in \mathscr{B}^s(\Omega)$ for $s > 0$.

Theorem 1. *Let $\Omega = [0,1]^d$, $0 \leq m \leq s$, and $f \in \mathscr{B}^s(\Omega)$. Then there is an $M \lesssim \|f\|_{\mathscr{B}^s(\Omega)}$ such that*

$$\inf_{f_n \in \Sigma_{n,M}} \|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}^s(\Omega)} n^{-\frac{1}{2} - \frac{s-m}{d}}. \tag{38}$$

Note that the implied constant in the above theorem depends only upon s, m, and d, but not on f. Comparing this with the results in [27], we obtain a dimension dependent improvement similar to what can be obtained using stratified sampling [29] for rectified linear networks. However, the improvement in Theorem 1, which is obtained via an entirely different argument, is greater and holds for cosine networks. We will consider rectified linear networks in the next section. Also, we note that for the Sobolev spaces $H^{\frac{d}{2}+s}(\Omega)$, this result already appears in [44]. However, our results apply to the spectral Barron space $\mathscr{B}^s(\Omega)$, which is not quite comparable, but we do have $H^{\frac{d}{2}+s+\varepsilon}(\Omega) \subset \mathscr{B}^s(\Omega)$ (see [51] Lemma 2.5, for instance). Finally, as shown in Theorem 6, the rate in Theorem 1 is actually sharp.

Proof. Choose $L > 1$. Note that all of the implied constants in what follows depend only upon s, m, d and L, but not upon f.

By Corollary 1, there exists an $a \in L^{-1}[0,1]^d$ and coefficients c_ξ such that

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} c_\xi e^{2\pi \mathrm{i}(a+\xi)\cdot x}, \tag{39}$$

and (here the first estimate follows since $|a| \leq L^{-d}\sqrt{d}$)

$$\sum_{\xi \in L^{-1}\mathbb{Z}^d} (1 + |\xi|)^s |c_\xi| \eqsim \sum_{\xi \in L^{-1}\mathbb{Z}^d} (1 + |a+\xi|)^s |c_\xi| \lesssim \|f\|_{\mathscr{B}^s(\Omega)}. \tag{40}$$

Consider the slightly enlarged set $\Omega' = [0, L]^d \supset \Omega$. On this larger set, we have for $\xi \neq \nu \in L^{-1}\mathbb{Z}^d$

$$\langle e^{2\pi \mathrm{i}(a+\xi)\cdot x}, e^{2\pi \mathrm{i}(a+\nu)\cdot x} \rangle_{H^k(\Omega')} = 0, \tag{41}$$

so that the frequencies in the expansion (39) form an orthogonal basis in $H^k(\Omega')$. Moreover, their lengths satisfy

$$\|e^{2\pi \mathrm{i}(a+\xi)\cdot x}\|_{H^m(\Omega')} \lesssim (1 + |a+\xi|)^m \eqsim (1 + |\xi|)^m. \tag{42}$$

Order the frequencies $\xi \in L^{-1}\mathbb{Z}^d$ such that

$$(1+|\xi_1|)^{2m-s}|c_{\xi_1}| \geq (1+|\xi_2|)^{2m-s}|c_{\xi_2}| \geq (1+|\xi_3|)^{2m-s}|c_{\xi_3}| \geq \cdots . \tag{43}$$

For $n \geq 1$, let $S_n = \{\xi_1, \xi_2, ..., \xi_n\}$ and set

$$f_n = \sum_{\xi \in S_n} c_\xi e^{2\pi \mathrm{i}(a+\xi)\cdot x} \in \Sigma_{n,M} \tag{44}$$

for $M \lesssim \|f\|_{\mathscr{B}^s}$ by (40).

We now estimate, using (41) and (42),

$$\begin{aligned} \|f - f_n\|^2_{H^m(\Omega')} &= \left\| \sum_{\xi \in S_n^c} c_\xi e^{2\pi \mathrm{i}(a+\xi)\cdot x} \right\|^2_{H^m(\Omega')} \\ &= \sum_{\xi \in S_n^c} |c_\xi|^2 \|e^{2\pi \mathrm{i}(a+\xi)\cdot x}\|^2_{H^m(\Omega')} \\ &\lesssim \sum_{\xi \in S_n^c} |c_\xi|^2 (1+|\xi|)^{2m}. \end{aligned} \tag{45}$$

Using Hoelder's inequality, we get

$$\sum_{\xi \in S_n^c} |c_\xi|^2 (1+|\xi|)^{2m} \leq \left(\sup_{\xi \in S_n^c} |c_\xi|(1+|\xi|)^{2m-s} \right) \left(\sum_{\xi \in S_n^c} |c_\xi|(1+|\xi|)^s \right) \tag{46}$$

By (40), the second term above is $\lesssim \|f\|_{\mathscr{B}^s(\Omega)}$. For the first term, we note that (40) implies that

$$\sum_{\nu \in S_n} |c_\nu|(1+|\nu|)^{2m-s}(1+|\nu|)^{2(s-m)} \lesssim \|f\|_{\mathscr{B}^s(\Omega)}. \tag{47}$$

Now, by the definition of S_n, we have for every $\nu \in S_n$

$$\left(\sup_{\xi \in S_n^c} |c_\xi|(1+|\xi|)^{2m-s} \right) \leq |c_\nu|(1+|\nu|)^{2m-s}, \tag{48}$$

so that

$$\begin{aligned} &\left(\sup_{\xi \in S_n^c} |c_\xi|(1+|\xi|)^{2m-s} \right) \sum_{\nu \in S_n} (1+|\nu|)^{2(s-m)} \\ &\leq \sum_{\nu \in S_n} |c_\nu|(1+|\nu|)^{2m-s}(1+|\nu|)^{2(s-m)}. \end{aligned} \tag{49}$$

By (47), we thus have

$$\left(\sup_{\xi \in S_n^c} |c_\xi|(1+|\xi|)^{2m-s} \right) \lesssim \|f\|_{\mathscr{B}^s(\Omega)} \left(\sum_{\nu \in S_n} (1+|\nu|)^{2(s-m)} \right)^{-1}. \tag{50}$$

The sum $\sum_{\nu \in S_n} (1+|\nu|)^{2(s-m)}$ is over n elements of the lattice $L^{-1}\mathbb{Z}^d$, from which it easily follows by comparison with an integral (see, for instance, [21]) that

$$\sum_{\nu \in S_n} (1 + |\nu|)^{2(s-m)} \gtrsim n^{1+\frac{2(s-m)}{d}}, \tag{51}$$

and we obtain

$$\left(\sup_{\xi \in S_n^c} |c_\xi| (1 + |\xi|)^{2m-s} \right) \lesssim \|f\|_{\mathscr{B}^s(\Omega)} n^{-1-\frac{2(s-m)}{d}}. \tag{52}$$

Combining this with (45) and (46), we get

$$\|f - f_n\|_{H^m(\Omega')}^2 \lesssim \|f\|_{\mathscr{B}^s(\Omega)}^2 n^{-1-\frac{2(s-m)}{d}}. \tag{53}$$

Finally, since $\Omega' \supset \Omega$, we get

$$\|f - f_n\|_{H^m(\Omega)} \leq \|f - f_n\|_{H^m(\Omega')} \lesssim \|f\|_{\mathscr{B}^s(\Omega)} n^{-\frac{1}{2}-\frac{(s-m)}{d}}, \tag{54}$$

which completes the proof. □

In Theorem 1, we obtained arbitrarily high polynomial rates of convergence for sufficiently smooth functions. Next we generalize this by showing that if the Fourier transform decays at a superpolynomial rate, then we can obtain spectral (i.e. superpolynomial) convergence as well. We begin by introducing an exponential version of the spectral Barron spaces.

Definition 1. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and let $0 < \beta < 1$ and $c > 0$. The exponential spectral Barron space with parameters β and c is defined by

$$\mathscr{B}_{\beta,c}(\Omega) := \left\{ f : \Omega \to \mathbb{R} : \|f\|_{\mathscr{B}_{\beta,c}(\Omega)} := \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} e^{c|\xi|^\beta} |\hat{f}_e(\xi)| d\xi < \infty \right\}, \tag{55}$$

where the infimum is taken over all extension $f_e \in L^1(\mathbb{R}^d)$.

The space $\mathscr{B}_{\beta,c}(\Omega)$ is quite restrictive, however there it still contains a relatively large class of functions. For example, it contains satisfied by any linear combination of Gaussians or any band-limited function, i.e. any function whose Fourier transform is compactly supported.

For elements of $\mathscr{B}_{\beta,c}(\Omega)$, we can prove a superpolynomial convergence rate.

Theorem 2. *Let $\Omega = [0,1]^d$, $0 < \beta < 1$, and $c > 0$. Then for any $m \geq 0$, there exists a $c' > 0$ such that for $f \in \mathscr{B}_{\beta,c}(\Omega)$ and $M \lesssim \|f\|_{\mathscr{B}_{\beta,c}(\Omega)}$ we have*

$$\inf_{f_n \in \Sigma_{n,M}} \|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}_{\beta,c}(\Omega)} e^{-c' n^{d^{-1}\beta}}. \tag{56}$$

Note that in this theorem the implied constant and the constant c' only depend upon β, c, d and m, but not on f or n.

Proof. We use a similar argument to the proof of Theorem 1. First, we apply Lemma 2 to the weight $\mu(\xi) = e^{c|\xi|^\beta}$, to obtain an $a \in L^{-1}[0,1]^d$ and coefficients c_ξ such that

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} c_\xi e^{2\pi i(a+\xi)\cdot x} \tag{57}$$

and

$$\sum_{\xi\in L^{-1}\mathbb{Z}^d} e^{c|a+\xi|^\beta}|c_\xi| \lesssim \|f\|_{\mathscr{B}_{\beta,c}(\Omega)}. \tag{58}$$

As in the proof of Theorem 1, we note that the frequencies $e^{2\pi i(a+\xi)\cdot x}$ are orthogonal on the enlarger set $\Omega' = [0, L]^d$ and their norms are bounded by (42).

This time, we order the frequencies $\xi \in L^{-1}\mathbb{Z}^d$ such that

$$\begin{aligned}(1+|\xi_1|)^{2k}e^{-c|a+\xi_1|^\beta}|c_{\xi_1}| &\geq (1+|\xi_2|)^{2k}e^{-c|a+\xi_2|^\beta}|c_{\xi_2}| \\ &\geq (1+|\xi_3|)^{2k}e^{-c|a+\xi_3|^\beta}|c_{\xi_3}| \geq \cdots.\end{aligned} \tag{59}$$

Choosing $S_n = \{\xi_1, ..., \xi_n\}$ and setting

$$f_n(x) = \sum_{\xi\in S_n} c_\xi e^{2\pi i(a+\xi)\cdot x} \in \Sigma_{n,M}, \tag{60}$$

with $M \lesssim \|f\|_{\mathscr{B}_{\beta,c}(\Omega)}$, we obtain, using the argument between equations (45) and (46), that

$$\|f - f_n\|^2_{H^m(\Omega')} \leq \left(\sup_{\xi\in S_n^c} |c_\xi|(1+|\xi|)^{2m}e^{-c|a+\xi|^\beta}\right)\left(\sum_{\xi\in S_n^c}|c_\xi|e^{c|a+\xi|^\beta}\right). \tag{61}$$

By (58), the second factor is $\lesssim \|f\|_{\mathscr{B}_{\beta,c}(\Omega)}$.

For the first factor, the argument between equations (40) and (47) implies that

$$\left(\sup_{\xi\in S_n^c} |c_\xi|(1+|\xi|)^{2m}e^{-c|a+\xi|^\beta}\right) \lesssim C_f\left(\sum_{\nu\in S_n} e^{2c|a+\nu|^\beta}(1+|a+\nu|)^{-2m}\right)^{-1}. \tag{62}$$

We now proceed to lower bound the sum on the right by considering its largest term. Since the sum is over n elements of the lattice $a + L^{-1}\mathbb{Z}^d$, the longest vector, i.e. the largest length $|a+\xi|$ which occurs in the sum, must be $\gtrsim n^{\frac{1}{d}}$. In addition $(1+|a+\xi|)^{2m} \lesssim e^{2\varepsilon|a+\xi|^\beta}$ for any $\varepsilon > 0$, so we see that there must exist a $c' > 0$ such that

$$\left(\sum_{\nu\in S_n} e^{2c|a+\nu|^\beta}(1+|a+\nu|)^{-2m}\right) \gtrsim e^{2c'n^{\frac{\beta}{d}}}. \tag{63}$$

Plugging this into (62) and (61) and using the fact that $\Omega' \subset \Omega$, we get

$$\inf_{f_n\in\Sigma_{n,M}} \|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}_{\beta,c}(\Omega)}e^{-c'n^{\frac{\beta}{d}}}, \tag{64}$$

as desired. □

3. Approximation rates for ReLU^k networks

In this section, we consider approximation by neural networks with activation function

$$\sigma_k(x) = [\max(0, x)]^k$$

for $k = \mathbb{Z}_{\geq 0}$ (here we set $0^0 = 0$, i.e. $\sigma_0(x)$ is the Heaviside function). Specifically, we consider approximating a function f by elements of the set

$$\Sigma_n^k = \left\{ \sum_{i=1}^n a_i \sigma_k(\omega_i \cdot x + b_i) : \ \omega_i \in S^{d-1}, \ b_i \in \mathbb{R}, \ a_i \in \mathbb{C} \right\}, \tag{65}$$

where we allow the coefficients a_i to have arbitrarily large ℓ^1-norm.

We will use Lemma 1 to obtain an improved approximation rate for such networks on the spectral Barron space $\mathscr{B}^m(\Omega)$. To do this, we introduce a multiscale approximation of the complex exponentials $e^{2\pi i x}$ using splines. We begin by recalling some facts about spline interpolation which will be important in the following analysis. We will refer to [17] for most of this material. Note also that similar arguments have been used to study the approximation properties of neural networks in one dimension [13,12].

Instead of working directly with σ_k it is more convenient to introduce the cardinal B-splines

$$N_k(x) = \frac{1}{k!} \sum_{i=0}^{k+1} (-1)^i \binom{k+1}{i} \sigma_k(x - i) \in \Sigma_{k+2}^k, \tag{66}$$

which are compactly supported on $[0, k+1]$.

Let $\mathscr{S}_\lambda^k$ denote the Schoenberg space of piecewise degree k splines on $\mathbb{R}$ with knots at $\lambda\mathbb{Z}$. It is well known that every spline $S \in \mathscr{S}_1^k$ can be written as

$$S(x) = \sum_{j=-\infty}^{\infty} c_j(S) N_k(x - j), \tag{67}$$

where c_j are the de Boor-Fix functionals (see [17], section 5.3). Since the knots of the spline are all evenly spaced, the functionals $c_j(S)$ are all translations of the functional c_0, i.e.

$$c_j(S) = c_0(S(\cdot - j)). \tag{68}$$

Moreover, consider change the spacing between the knots, i.e. consider $\mathscr{S}_\lambda^k$. Then, if $S \in \mathscr{S}_\lambda^k$, $S(\lambda x) \in \mathscr{S}_1^k$ and equations (67) and (68) imply that

$$S(\lambda x) = \sum_{j=-\infty}^{\infty} c_j(S(\lambda \cdot)) N_k(x - j), \tag{69}$$

so that

$$S(x) = \sum_{j=-\infty}^{\infty} c_{j,\lambda}(S) N_k(\lambda^{-1} x - j), \tag{70}$$

where the functionals $c_{j,\lambda}$ are given by $c_{j,\lambda}(S) = c_0(S(\lambda(\cdot - j)))$.

Now, we see from [17], Lemma 4.1 of Chapter 5, that

$$|c_{j,\lambda}(S)| \leq C \|S\|_{L^\infty([\lambda j, \lambda(j+k+1)])}, \tag{71}$$

for a fixed constant C. Thus, by the Hahn-Banach theorem, we can extend the de Boor-Fix functionals $c_{j,\lambda}$ to functionals $\gamma_{j,\lambda}$ on $L^\infty([\lambda j, \lambda(j+k+1)])$ which satisfy the same bound. This allows us to define the quasi-interpolation operators

$$Q_\lambda(f) = \sum_{j=-\infty}^{\infty} \gamma_{j,\lambda}(f) N_k(\lambda^{-1}x - j), \tag{72}$$

which are bounded in L^∞ (uniformly in λ) and satisfy $Q(S) = S$ for all splines $S \in \mathscr{S}_\lambda^k$ (see [17], section 5.4). Note that here and in what follows, we suppress the dependence on k of the operators Q_λ and the de Boor-Fix functions $\gamma_{j,\lambda}$ to simplify notation.

We are now in the position to introduce the following multiscale piecewise degree k approximation to $e^{2\pi \mathrm{i} x}$. We write

$$e^{2\pi \mathrm{i} x} = Q_{2^{-1}}(e^{2\pi \mathrm{i} x}) + \sum_{l=2}^{\infty} [Q_{2^{-l}}(e^{2\pi \mathrm{i} x}) - Q_{2^{-(l-1)}}(e^{2\pi \mathrm{i} x})] = \sum_{l=1}^{\infty} h_l(x), \tag{73}$$

where

$$h_l(x) = Q_{2^{-l}}(e^{2\pi \mathrm{i} x}) - Q_{2^{-(l-1)}}(e^{2\pi \mathrm{i} x}), \tag{74}$$

for $l > 1$ and $h_1(x) = Q_{2^{-1}}(e^{2\pi \mathrm{i} x})$. Since we clearly have $\mathscr{S}^k_{2^{-(l-1)}} \subset \mathscr{S}^k_{2^{-l}}$, we see that

$$Q_{2^{-l}}(Q_{2^{-(l-1)}}(e^{2\pi \mathrm{i} x})) = Q_{2^{-(l-1)}}(e^{2\pi \mathrm{i} x}),$$

so that we can rewrite h_l as

$$h_l(x) = Q_{2^{-l}}\left(e^{2\pi \mathrm{i} x} - Q_{2^{-(l-1)}}(e^{2\pi \mathrm{i} x})\right) = Q_{2^{-l}}(e_{l-1}(x)) = \sum_{j=-\infty}^{\infty} \alpha_{j,l} N_k(2^l x - j), \tag{75}$$

where the error e_{l-1} is given by $e_{l-1}(x) = e^{2\pi \mathrm{i} x} - Q_{2^{-(l-1)}}(e^{2\pi \mathrm{i} x})$ and the coefficients $\alpha_{j,l}$ are given by $\alpha_{j,l} = \gamma_{j,2^{-l}}(e_{l-1})$.

We have the following lemma concerning this piecewise degree k approximation of $e^{2\pi \mathrm{i} x}$.

Lemma 3. *The above expansion of $e^{2\pi \mathrm{i} x}$ has the following properties.*

- $\|e_l\|_\infty \lesssim 2^{-(k+1)l}$.
- *The coefficients $\alpha_{j,l}$ in equation (75) satisfy $|\alpha_{j,l}| \lesssim 2^{-(k+1)l}$.*
- *The series in (73) converges in $W^{m,\infty}(\mathbb{R})$ for $0 \leq m \leq k$.*

Note that the implied constants in the above lemma and the following proof only depend upon k and not upon l or j.

Proof. The first statement follows immediately from Theorem 4.5 in [17], since $e_l(x) = e^{2\pi \mathrm{i} x} - Q_{2^{-l}}(e^{2\pi \mathrm{i} x})$ and $e^{2\pi \mathrm{i} x} \in W^{k+1,\infty}(\mathbb{R})$.

For the second statement, we note that

$$|\alpha_{j,l}| = |\gamma_{j,2^{-l}}(e_{l-1})| \leq C\|e_{l-1}\|_{L^\infty(\mathbb{R})} \lesssim 2^{-(k+1)(l-1)} \lesssim 2^{-(k+1)l}, \tag{76}$$

where the first inequality is due to the fact that $\gamma_{j,2^{-j}}$ is a Hahn-Banach extension of the de Boor-Fix functional $c_{j,2^{-j}}$ which satisfies (71).

Finally, note that since $\|e_l\|_{L^\infty(\mathbb{R})} \lesssim 2^{-(k+1)l} \to 0$, we have that the series in (73) converges in $L^\infty(\mathbb{R})$ to $e^{2\pi \mathrm{i} x}$. We now claim that

$$\|h_l\|_{W^{m,\infty}(\mathbb{R})} \lesssim 2^{-(k+1-m)l}. \tag{77}$$

First, we note that simply by taking derivatives, we get

$$\|N_p(2^l x - j)\|_{W^{m,\infty}(\mathbb{R},dx)} \lesssim 2^{ml}. \tag{78}$$

Second, the B-splines $N_k(2^l x - j)$ are compactly supported and each point x is covered by at most $p+1$ of them. Hence

$$\begin{aligned} \|h_l\|_{W^{m,\infty}} &= \left\| \sum_{j=-\infty}^{\infty} \alpha_{j,l} N_k(2^l x - j) \right\|_{W^{m,\infty}} \\ &\leq (k+1) \sup_j |\alpha_{j,l}| \|N_k(2^l x - j)\|_{W^{m,\infty}(\mathbb{R},dx)} \\ &\lesssim 2^{-(k+1-m)l}, \end{aligned} \tag{79}$$

since $\alpha_{j,l} \lesssim 2^{-(k+1)l}$.

This means that if $m \leq k$, then $\sum_{l=1}^{\infty} \|h_l\|_{W^{m,\infty}}$ is summable and hence the sum in (73) converges in $W^{m,\infty}(\mathbb{R})$. Clearly, its limit must be the same as the limit in L^∞ and thus it converges to $e^{2\pi i x}$. □

Combining the multiscale expansion (73) with Lemma 1 and some ideas from [41], we obtain the following theorem.

Theorem 3. *Let $\Omega = [0,1]^d$ and $f \in \mathscr{B}^s(\Omega)$ for $s \geq \frac{1}{2}$. Let $k \in \mathbb{Z}_{\geq 0}$ and $m \geq 0$, with $m \leq s - \frac{1}{2}$ and $m < k + \frac{1}{2}$. Then for $n \geq 2$,*

$$\inf_{f_n \in \Sigma_n^k} \|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}^s} n^{-t} \log(n)^q, \tag{80}$$

where the exponent t is given by

$$\begin{aligned} t &= \frac{1}{2} + \min\left(\frac{2(s-m)-1}{2(d+1)}, k - m + \frac{1}{2} \right) \\ &= \begin{cases} \frac{1}{2} + \frac{2(s-m)-1}{2(d+1)} & \text{if } s < (d+1)\left(k - m + \frac{1}{2}\right) + m + \frac{1}{2} \\ k - m + 1 & \text{if } s \geq (d+1)\left(k - m + \frac{1}{2}\right) + m + \frac{1}{2} \end{cases} \end{aligned} \tag{81}$$

and q is given by

$$q = \begin{cases} 0 & \text{if } s < (d+1)\left(k - m + \frac{1}{2}\right) + m + \frac{1}{2} \\ 1 & \text{if } s > (d+1)\left(k - m + \frac{1}{2}\right) + m + \frac{1}{2} \\ 1 + (k - m + \frac{1}{2}) & \text{if } s = (d+1)\left(k - m + \frac{1}{2}\right) + m + \frac{1}{2} \end{cases}.$$

Before beginning the proof, we remark that all of the implied constants in the $\eqsim$, $\gtrsim$, and $\lesssim$ can be seen to depend only on s, k, m, d, L and δ (L and δ chosen during the course of the proof), but not on f or n. Further, we remark that the suppressed constant may depend exponentially on the dimension, i.e. as A^d for some A. Finally, note that the maximal possible rate of $s - m + 1$, which is achieved for sufficiently large s, is exactly the best achievable rate in one dimension. In Theorem 7 we use this fact to show that the rate of $s - m + 1$ cannot be improved upon no matter how large s is. It is an open problem whether such a rate can be obtained with less smoothness.

Comparing with other results in the literature, we see for instance that the results in [29] apply to the cases $k = 1, m = 0$ (ReLU) and $k = 2, m = 0$ (ReLU^2). Furthermore, in [51], the general case $0 \leq m \leq k$ is

considered. In all of these cases the rate previously obtained was $O(n^{-\frac{1}{2}-\frac{1}{d}})$, while the rates in Theorem 3 are $O(n^{-\frac{1}{2}-\frac{2k+1}{2(d+1)}})$, which are significantly better for large k and large d. However, the rates in Theorem 3 were obtained without the ℓ^1-norm bound on the coefficients as in [29] and [51]. It is open whether the same rates can also be obtained with ℓ^1-bounded coefficients.

Proof. Choose $L > 1$. Using Corollary 1, we see that there exists an $a \in L^{-1}[0,1]^d$ and coefficients a_ξ such that

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} a_\xi (1 + |a + \xi|)^{-s} e^{2\pi i (a+\xi)\cdot x} \tag{82}$$

and $\sum |a_\xi| \lesssim \|f\|_{\mathscr{B}^s}$. Here the suppressed constant depends potentially exponentially on the dimension, by the remarks in the previous section.

We expand $e^{2\pi i(a+\xi)\cdot x}$ using (73) to get

$$e^{2\pi i(a+\xi)\cdot x} = \sum_{l=1}^{\infty} h_l((a+\xi)\cdot x), \tag{83}$$

which holds in $W^{m,\infty}(\mathbb{R}^d)$ and thus in $H^m(\Omega)$ since Ω is bounded.

Expanding h_l using equation (75) and plugging this into equation (82), we obtain (in $H^m(\Omega)$)

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} \sum_{l=1}^{\infty} \sum_{j=-\infty}^{\infty} a_\xi \alpha_{j,l} (1 + |a+\xi|)^{-s} N_k(2^l(a+\xi)\cdot x - j). \tag{84}$$

Now, since $x \in \Omega$, Ω is a bounded set, and N_k is compactly supported, the number of non-zero terms in the inner-most sum above is finite. Indexing the values of j for which $N_k(2^l(a+\xi)\cdot x - j)$ is non-zero for $x \in \Omega$ as $j_1, ..., j_{n_{\xi,l}}$, we get

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} \sum_{l=1}^{\infty} \sum_{p=1}^{n_{\xi,l}} a_\xi \alpha_{j_p,l} (1 + |a+\xi|)^{-s} \psi_{\xi,l,p}(x), \tag{85}$$

where

$$\psi_{\xi,l,p}(x) = N_k(2^l(a+\xi)\cdot x - j_p). \tag{86}$$

A straightforward calculation utilizing the compact support of N_k implies that

$$\|\psi_{\xi,l,p}\|_{H^m(\Omega)} \lesssim 2^{l(m-\frac{1}{2})} (1 + |\xi|)^{(m-\frac{1}{2})}. \tag{87}$$

Further, note that the number of terms $n_{\xi,l}$ satisfies

$$n_{\xi,l} \lesssim 2^l (1 + |\xi|). \tag{88}$$

This follows since for $x \in \Omega$, $y = 2^l(a+\xi)\cdot x$ takes on values in an interval of length at most $2^l|a+\xi|\mathrm{diam}(\Omega)$ and N_k has compact support of size $k+1$.

Let $\delta > 0$ to be specified later. We proceed to write

$$f(x) = \sum_{\xi \in L^{-1}\mathbb{Z}^d} \sum_{l=1}^{\infty} \sum_{p=1}^{n_{\xi,l}} a_\xi 2^{-l(1+\delta)} (1 + |\xi|)^{-1} \phi_{\xi,l,p}(x), \tag{89}$$

where

$$\phi_{\xi,l,p}(x) = 2^{l(1+\delta)}(1+|\xi|)\alpha_{j_p,l}(1+|a+\xi|)^{-s}\psi_{\xi,l,p}(x). \tag{90}$$

Using (90) and (87) combined with the bound on $|\alpha_{j,l}|$ from Lemma 3, we calculate

$$\|\phi_{\xi,l,p}\|_{H^m(\Omega)} \lesssim 2^{-l(k-m+\frac{1}{2}-\delta)}(1+|\xi|)^{m-s+\frac{1}{2}}. \tag{91}$$

We now observe that by (88) the ℓ^1 norm of the coefficients of the $\phi_{\xi,l,p}$ in (89) is bounded, namely

$$\begin{aligned}\sum_{\xi\in L^{-1}\mathbb{Z}^d}\sum_{l=1}^{\infty}\sum_{p=1}^{n_{\xi,l}}|a_\xi 2^{-l(1+\delta)}(1+|\xi|)^{-1}| &= \sum_{\xi\in L^{-1}\mathbb{Z}^d}|a_\xi|\sum_{l=1}^{\infty}n_{\xi,l}2^{-l(1+\delta)}(1+|\xi|)^{-1}\\ &\lesssim \sum_{\xi\in L^{-1}\mathbb{Z}^d}|a_\xi|\sum_{l=1}^{\infty}2^{-l\delta}\\ &\lesssim \delta^{-1}\|f\|_{\mathscr{B}^m(\Omega)}.\end{aligned} \tag{92}$$

We can now apply Theorem 1 in [41] (note that this theorem still applies even though the coefficients in (89) are potentially complex) to f to conclude that there exists an

$$f_n = \sum_{i=1}^{n} a_i\phi_{\xi_i,l_i,p_i}(x) \tag{93}$$

with $\sum_{i=1}^n |a_i| \lesssim \|f\|_{\mathscr{B}^s(\Omega)}$ such that

$$\|f-f_n\|_{H^m(\Omega)} \lesssim \delta^{-1}\|f\|_{\mathscr{B}^s(\Omega)}\varepsilon_n(\Phi)n^{-\frac{1}{2}}, \tag{94}$$

where $\Phi = \{\phi_{\xi,l,p}\}$ and $\varepsilon_n(\Phi) = \inf\{\varepsilon > 0 : \Phi \text{ is covered by } n \text{ balls of diameter } \varepsilon\}$ is the n-covering width of Φ.

By choosing $\delta = k - m + \frac{1}{2} > 0$ we obtain the result at the endpoint $s = m + \frac{1}{2}$ (where the desired rate is $O(n^{-\frac{1}{2}})$) since by (91)

$$\varepsilon_n(\Phi) \le \varepsilon_1(\Phi) \le \sup\|\phi_{\xi,l,p}\|_{H^m(\Omega)} \lesssim 1.$$

For larger s we need to obtain a sharper bound on $\varepsilon_n(\Phi)$. We do this by considering the covering number

$$N_\Phi(\varepsilon) = \min\{n : \text{ there is a covering of } \Phi \text{ by } n \text{ balls of diameter } \varepsilon\}, \tag{95}$$

and noting that by definition $\varepsilon_n(\Phi) = \inf\{\varepsilon > 0 : N_\Phi(\varepsilon) \le n\}$.

Given $\varepsilon > 0$, we cover the set Φ by a single ball of radius $\frac{\varepsilon}{2}$ centered at the origin, and cover each of the remaining elements with additional balls. This implies that

$$N_\Phi(\varepsilon) \le 1 + \left|\left\{\phi_{\xi,l,p} : \|\phi_{\xi,l,p}\|_{H^m(\Omega)} > \frac{\varepsilon}{2}\right\}\right|. \tag{96}$$

We proceed to count the number of $\phi_{\xi,l,p}$ with large norm. This process is messy but relatively straightforward.

By (91) we must count the indices $\xi \in L^{-1}\mathbb{Z}^d, l \in \mathbb{Z}_{>0}$ and $s = 1, ..., n_{\xi,l}$ for which

$$\varepsilon \lesssim 2^{-l(k-m+\frac{1}{2}-\delta)}(1+|\xi|)^{m-s+\frac{1}{2}}. \tag{97}$$

We observe that this condition implies that we must choose ξ so that $(1+|\xi|)^{m-s+\frac{1}{2}} \gtrsim \varepsilon$ and l so that

$$2^{l(k-m+\frac{1}{2}-\delta)} \lesssim \varepsilon^{-1}(1+|\xi|)^{m-s+\frac{1}{2}}. \tag{98}$$

In addition, for each of these values of ξ and l, we get $n_{\xi,l} \lesssim 2^l(1+|\xi|)$ different values of p. Combining these observations, we see that

$$\left|\left\{\phi_{\xi,l,p} : \|\phi_{\xi,l,p}\|_{H^m(\Omega)} > \frac{\varepsilon}{2}\right\}\right| \lesssim \sum_{\substack{\xi \in L^{-1}\mathbb{Z}^d \\ |\xi| \leq R}} (1+|\xi|) \sum_{l \in L(\varepsilon,\xi)} 2^l, \tag{99}$$

where $R \lesssim \varepsilon^{\frac{1}{m-s+\frac{1}{2}}}$ (note that here we require $m-s+\frac{1}{2} < 0$) and $L(\varepsilon,\xi)$ consists of indices l which satisfy (98). Taking a logarithm, the set $L(\varepsilon,\xi)$ can be characterized by

$$l \leq \left(k-m+\frac{1}{2}-\delta\right)^{-1}\left(-\log(\varepsilon) + \left(m-s+\frac{1}{2}\right)\log(1+|\xi|)\right) + C \tag{100}$$

for some constant C. Using this bound on l, combined with the fact that $\sum_{l=1}^{k} 2^l \lesssim 2^k$, we get

$$\sum_{l \in L(\varepsilon,\xi)} 2^l \lesssim \varepsilon^{\frac{-1}{k-m+\frac{1}{2}-\delta}} (1+|\xi|)^{\frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta}}. \tag{101}$$

So we get

$$\left|\left\{\phi_{\xi,l,p} : \|\phi_{\xi,l,p}\|_{H^m(\Omega)} > \frac{\varepsilon}{2}\right\}\right| \lesssim \varepsilon^{\frac{-1}{k-m+\frac{1}{2}-\delta}} \sum_{\substack{\xi \in L^{-1}\mathbb{Z}^d \\ |\xi| \leq R}} (1+|\xi|)^{1+\frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta}}. \tag{102}$$

For the final sum, we distinguish between two cases.

First, if $m-s+\frac{1}{2} > -(d+1)\left(k-m+\frac{1}{2}\right)$, then we can choose a fixed $\delta = \delta(s,m,k,d) > 0$ small enough so that

$$1 + \frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta} > -d. \tag{103}$$

In this case, by comparing the sum over the lattice $L^{-1}\mathbb{Z}^d$ to an integral, the sum in (102) satisfies

$$\sum_{\substack{\xi \in L^{-1}\mathbb{Z}^d \\ |\xi| \leq R}} (1+|\xi|)^{1+\frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta}} \lesssim R^{d+1+\frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta}} \lesssim \varepsilon^{\frac{d+1}{m-s+\frac{1}{2}}+\frac{1}{k-m+\frac{1}{2}-\delta}}, \tag{104}$$

since $R \lesssim \varepsilon^{\frac{1}{m-s+\frac{1}{2}}}$. Combining this with (102) we get

$$\left|\left\{\phi_{\xi,l,p} : \|\phi_{\xi,l,p}\|_{H^m(\Omega)} > \frac{\varepsilon}{2}\right\}\right| \lesssim \varepsilon^{\frac{d+1}{m-s+\frac{1}{2}}}. \tag{105}$$

This implies that for small ε, $N_\Phi(\varepsilon) \lesssim \varepsilon^{\frac{d+1}{m-s+\frac{1}{2}}}$ and so

$$\varepsilon_n(\Phi) \lesssim n^{\frac{m-s+\frac{1}{2}}{d+1}}. \tag{106}$$

Plugging this into (94), we get

$$\|f - f_n\|_{H^m(\Omega)} \lesssim \delta(s,m,k,d)^{-1}\|f\|_{\mathscr{B}^s(\Omega)} n^{\frac{m-s+\frac{1}{2}}{d+1}} n^{-\frac{1}{2}} \lesssim \|f\|_{\mathscr{B}^s(\Omega)} n^{-\frac{1}{2}-\frac{s-m-\frac{1}{2}}{d+1}}. \tag{107}$$

Next, if $m - s + \frac{1}{2} \leq -(d+1)\left(k - m + \frac{1}{2}\right)$, then for any $\delta > 0$ we get

$$1 + \frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta} < -d. \tag{108}$$

In this case, the sum in (102) is summable and we get

$$\sum_{\substack{\xi \in L^{-1}\mathbb{Z}^d \\ |\xi| \leq R}} (1+|\xi|)^{1+\frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta}} \lesssim 1 \tag{109}$$

if $m - s + \frac{1}{2} < -(d+1)\left(k - m + \frac{1}{2}\right)$, and in the special case where $m - s + \frac{1}{2} = -(d+1)\left(k - m + \frac{1}{2}\right)$, we get

$$\sum_{\substack{\xi \in L^{-1}\mathbb{Z}^d \\ |\xi| \leq R}} (1+|\xi|)^{1+\frac{m-s+\frac{1}{2}}{k-m+\frac{1}{2}-\delta}} \lesssim \delta^{-1}. \tag{110}$$

Combining this with (102) we get

$$\left|\left\{\phi_{\xi,l,p} : \|\phi_{\xi,l,p}\|_{H^m(\Omega)} > \frac{\varepsilon}{2}\right\}\right| \lesssim \varepsilon^{\frac{-1}{k-m+\frac{1}{2}-\delta}}, \tag{111}$$

where we need an extra factor of δ^{-1} in the special case where $m - s + \frac{1}{2} = -(d+1)\left(k - m + \frac{1}{2}\right)$. This implies that up to a factor of $\delta^{-(k-m+\frac{1}{2}-\delta)}$ in this special case, we have

$$\varepsilon_n(\Phi) \lesssim n^{-(k-m+\frac{1}{2}-\delta)}. \tag{112}$$

Using (94), we get

$$\|f - f_n\|_{H^m(\Omega)} \lesssim \delta^{-1}\|f\|_{\mathscr{B}^s(\Omega)} n^{-(k-m+\frac{1}{2}-\delta)} n^{-\frac{1}{2}}, \tag{113}$$

where the power of δ is replaced by $-1-(k-m+\frac{1}{2}-\delta)$ in the endpoint case. Finally, optimizing over δ, we get

$$\|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}^s(\Omega)} n^{-(k-m+1)} \log(n), \tag{114}$$

where in the endpoint case the logarithm is taken to the power $1 + (k - m + \frac{1}{2})$.

Combining the results of (114) and (107) with the previously discussed result at $s = m + \frac{1}{2}$, we get

$$\|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}^s(\Omega)} n^{-t} \log(n)^q, \tag{115}$$

where $t = \min\left(\frac{1}{2} + \frac{s-m-\frac{1}{2}}{d+1}, k - m + 1\right)$ and q is given by

$$q = \begin{cases} 0 & \text{if } t < k - m + 1 \\ 1 & \text{if } t < \frac{1}{2} + \frac{2(s-m)-1}{2(d+1)} \\ 1 + (k - m + \frac{1}{2}) & \text{otherwise} \end{cases}. \tag{116}$$

This completes the proof since (86), (90), and (66) imply that $\phi_{\xi,l,p} \in \Sigma^k_{k+2}$. □

In the case where f is highly smooth, i.e. $f \in \mathscr{B}^s(\Omega)$ with $s > (d+1)(k - m - \frac{1}{2}) + m + \frac{1}{2}$, we obtain, up to a logarithmic factor, an approximation rate of $O(n^{-k-1+m})$ in $H^m(\Omega)$. We state this as a separate theorem.

Theorem 4. *Let $\Omega = [0,1]^d$, $k \in \mathbb{Z}_{\geq 0}$ and $m \geq 0$, with $m < k + \frac{1}{2}$. Suppose that $f \in \mathscr{B}^s(\Omega)$ for s sufficiently large, specifically*

$$s > (d+1)\left(k - m + \frac{1}{2}\right) + m + \frac{1}{2}.$$

Then for $n \geq 2$,

$$\inf_{f_n \in \Sigma^k_n} \|f - f_n\|_{H^m(\Omega)} \lesssim \|f\|_{\mathscr{B}^s} n^{m-(k+1)} \log(n). \tag{117}$$

Further, the special case $s = \frac{1}{2}$, $m = 0$ shows that the approximation rates obtained in [4] apply to a significantly larger class of functions if the ℓ^1-coefficient bound on the neural network is dropped.

Theorem 5. *Let $\Omega = [0,1]^d$ and $f \in B^s(\Omega)$ for $s = \frac{1}{2}$. Suppose that σ is an arbitrary sigmoidal function. Then*

$$\inf_{f_n \in \Sigma^\sigma_n} \|f - f_n\|_{L^2(\Omega)} \lesssim \|f\|_{\mathscr{B}^s} n^{-\frac{1}{2}}, \tag{118}$$

where

$$\Sigma^\sigma_n = \left\{ \sum_{i=1}^n a_i \sigma(\omega_i \cdot x + b_i) : \ \omega_i \in \mathbb{R}^d, \ b_i \in \mathbb{R}, \ a_i \in \mathbb{C} \right\}. \tag{119}$$

This result is obtained for $f \in B^1(\Omega)$ by Barron [4]. Here we show that in fact the condition $f \in B^{\frac{1}{2}}(\Omega)$ is sufficient.

Proof. By Theorem 3, the result holds if $\sigma = \sigma_0$ is the Heaviside function. For general sigmoidal σ, the result follows by noting that $\lim_{t\to\infty} \|\sigma(t(\omega \cdot x - b)) - \sigma_0(\omega \cdot x - b)\|_{L^2(\Omega)} \to 0$ for any b and ω. □

4. Lower bounds for cosine networks

In this section, we derive lower bounds which complement Theorems 1 and 3. We begin with lower bounds on the approximation rate of cosine networks on $\mathscr{B}^s(\Omega)$. In particular, we show that the approximation rate of Theorem 1 cannot be substantially improved when $m = 0$, i.e. when we are approximating in $L^2(\Omega)$. We prove this even when the coefficients are only required to be bounded in ℓ^∞, i.e. when approximating from the set

$$\Sigma^\infty_{n,M} = \left\{ \sum_{j=1}^n a_j e^{2\pi i \theta_j \cdot x} : \ \theta_j \in \mathbb{R}^d, \ a_j \in \mathbb{C}, \ |a_i| \leq M \right\}. \tag{120}$$

We have the following result.

Theorem 6. *Let $\Omega = [0,1]^d$ and $s \geq 0$. Then we have*

$$\limsup_{n\to\infty} \left[\sup_{\|f\|_{\mathscr{B}^s(\Omega)} \leq 1} \inf_{f_n \in \Sigma_{n,M}^{\infty}} \|f - f_n\|_{L^2(\Omega)} \right] n^{\frac{1}{2}+\frac{s}{d}+\varepsilon} = \infty \tag{121}$$

for any $M, \varepsilon > 0$.

Lower bounds for the σ_k activation function were obtained for $k = 0$ obtained in [41] and for $k \geq 1$ in [29]. However, for σ_k the lower bounds obtained do not match the best known rates. This gap has recently been closed in [50]. In contrast, Theorem 1 combined with Theorem 6 gives the optimal approximation rate for cosine networks on the spectral Barron space $\mathscr{B}^s(\Omega)$.

We also remark that a similar lower bound is obtained in section 7.2 of [11]. However, the lower bound here is in terms of approximation by dictionary expansions where the size of the dictionary depends polynomially upon the number of terms. In this case the correct tool is to find large hypercubes with the given class [18]. In contrast, the result we prove applies to the infinite, even non-compact dictionary of Fourier modes and the tool used is the metric entropy. As a consequence, we must assume that the coefficients are bounded. We are not quite sure how to remove this assumption completely although it can be relaxed to a bound which grows polynomially with the number of terms n.

Proof. The argument is a modification of the methods in [29,41]. The main new difficulty is in dealing with the non-compactness of the set $\{e^{2\pi i\theta\cdot x} : \theta \in \mathbb{R}^d\} \subset L^2(\Omega)$.

Suppose to the contrary that for some $M, \varepsilon > 0$, we have

$$\sup_{\|f\|_{\mathscr{B}^s(\Omega)} \leq 1} \inf_{f_n \in \Sigma_{n,M}} \|f - f_n\|_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{s}{d}-\varepsilon}. \tag{122}$$

For $R > 0$ consider the set

$$S(R) = \{\phi_\omega(x) := (1+|\omega|)^{-s} e^{2\pi i\omega\cdot x} : \ \omega \in \mathbb{Z}^d, \ |\omega|_\infty \leq R\}.$$

In the following proof all of the implied constants are independent of R.

The elements $\phi_\omega \in S(R)$ are orthogonal in $L^2(\Omega)$ and satisfy $\|\phi_\omega\|_{L^2(\Omega)} = (1+|\omega|)^{-s} \gtrsim R^{-s}$. In addition, it is clear that $\|\phi_\omega(x)\|_{\mathscr{B}^s(\Omega)} \leq 1$.

We now make use of the following combinatorial fact which follows from Berge's theorem (see [6,28]): given a set S of size n, there exist at least 2^{cn} subsets of S whose pairwise symmetric differences are at least $\frac{n}{4}$, where $c > 0$ is a universal constant (i.e. independent of n).

We apply this to the set $S(R)$ to see that there are subsets $S_1, ..., S_N \subset S(R)$ with $N = 2^{cR^d}$, such that for any $i \neq j$, we have $|S_i - S_j| \geq \frac{1}{4}R^d$. Consider the elements $\phi_i \in \mathscr{B}^s(\Omega)$ defined by

$$\phi_i(x) = \frac{1}{R^d} \sum_{\phi_\omega \in S_i} \phi_\omega(x). \tag{123}$$

We clearly have $\|\phi_i(X)\|_{\mathscr{B}^s(\Omega)} \leq 1$. Moreover, since $|S_i - S_j| \geq \frac{1}{4}R^d$, $\|\phi_\omega\|_{L^2(\Omega)} \gtrsim R^{-s}$, and the ϕ_ω are orthogonal, we see that for $i \neq j$

$$\begin{aligned}\|\phi_i(x) - \phi_j(x)\|_{L^2(\Omega)} &= \frac{1}{R^d}\left\|\sum_{\phi_\omega \in S_i - S_j} \phi_\omega(x)\right\|_{L^2(\Omega)} \\ &\gtrsim \frac{R^{-s}}{R^d}\sqrt{|S_i - S_j|} \geq \frac{R^{-s-\frac{d}{2}}}{2}.\end{aligned} \tag{124}$$

Thus, we have at least $N = 2^{cR^d}$ elements ϕ_i which satisfy $\|\phi_i(x)\|_{\mathscr{B}^s(\Omega)} \leq 1$, and such that every pair differs by at least $\delta \gtrsim R^{-s-\frac{d}{2}}$ in $L^2(\Omega)$. Note that we could also have obtained this from [29], Lemma 1 in section 2.2.

By (122) there exist $\phi_{i,n} \in \Sigma_{n,M}$ which satisfy

$$\|\phi_{i,n} - \phi_i\|_{L^2(\Omega)} \leq \frac{\delta}{6}, \tag{125}$$

for an n which satisfies

$$n \lesssim \delta^{-\frac{2d}{d+2s+2d\varepsilon}} \lesssim R^{\frac{d(2s+d)}{2s+d+2d\varepsilon}} = R^{d-t}, \tag{126}$$

where $t(s,d,\varepsilon) > 0$.

Let P_R denote the projection onto the space spanned by $S(R)$, i.e. onto the space spanned by the frequencies $e^{2\pi i\omega\cdot x}$ for $\omega \in \mathbb{Z}^d$, $|\omega|_\infty \leq R$. Consider the projection $P_R(e^{2\pi i\theta\cdot x})$ for $\theta \in \mathbb{R}^d$. We calculate

$$\begin{aligned}\|P_R(e^{2\pi i\theta\cdot x})\|^2_{L^2(\Omega)} &= \sum_{\substack{\omega\in\mathbb{Z}^d \\ |\omega|_\infty \leq R}} \left|\int_{[0,1]^d} e^{2\pi i(\theta-\omega)x} dx\right|^2 \\ &\leq \frac{1}{(2\pi)^{2d}} \sum_{\substack{\omega\in\mathbb{Z}^d \\ |\omega|_\infty \leq R}} \prod_{i=1}^d \frac{1}{|\theta_i - \omega_i|^2}.\end{aligned} \tag{127}$$

Choose K large enough such that $\|P_R(e^{2\pi i\theta\cdot x})\|_{L^2(\Omega)} \leq \frac{\delta}{6nM}$ as long as $|\theta|_\infty \geq K$. By (127) and (126), this will be guaranteed if

$$K \geq R + \frac{6}{(2\pi)^d}\delta^{-1}MnR^{\frac{d}{2}} \lesssim R^{s+2d-t}, \tag{128}$$

and so we can choose $K \lesssim R^{s+2d-t}$.

We proceed to truncate the $\phi_{i,n} \in \Sigma_{n,M}$ at frequencies with magnitude K. In particular, if

$$\phi_{i,n} = \sum_{j=1}^n a_{i,j} e^{2\pi i\theta_{i,j}\cdot x}, \tag{129}$$

we set $T_i^K = \{j : |\theta_{i,j}|_\infty \leq K\}$ and

$$\phi_{i,n}^K = \sum_{j\in T_i^K} a_{i,j} e^{2\pi i\theta_{i,j}\cdot x}. \tag{130}$$

Our choice of K guarantees that

$$\|P_R(\phi_{i,n}^K) - P_R(\phi_{i,n})\|_{L^2(\Omega)} \leq \frac{\delta}{6}, \tag{131}$$

which implies that

$$\begin{aligned}\|P_R(\phi_{i,n}^K) - \phi_i\|_{L^2(\Omega)} &\leq \|P_R(\phi_{i,n}^K) - P_R(\phi_{i,n})\|_{L^2(\Omega)} + \|P_R(\phi_{i,n} - \phi_i)\|_{L^2(\Omega)} \\ &\leq \frac{\delta}{6} + \frac{\delta}{6} = \frac{\delta}{3},\end{aligned} \tag{132}$$

since $P_r(\phi_i) = \phi_i$.

We now conclude that for $i \neq j$, we have

$$\begin{aligned}\|\phi_{i,n}^K - \phi_{j,n}^K\|_{L^2(\Omega)} &\geq \|P_R(\phi_{i,n}^K) - P_R(\phi_{j,n}^K)\|_{L^2(\Omega)} \\ &\geq \|\phi_j - \phi_i\|_{L^2(\Omega)} - \|P_R(\phi_{j,n}^K) - \phi_j\|_{L^2(\Omega)} - \|P_R(\phi_{i,n}^K) - \phi_i\|_{L^2(\Omega)} \\ &\geq \delta - \frac{\delta}{3} - \frac{\delta}{3} = \frac{\delta}{3}.\end{aligned} \tag{133}$$

However, on the other hand, we calculate that

$$\|e^{2\pi \mathrm{i}\theta_1 \cdot x} - e^{2\pi \mathrm{i}\theta_2 \cdot x}\|_{L^2(\Omega)}^2 = \int\limits_{[0,1]^d} |1 - e^{2\pi \mathrm{i}(\theta_1 - \theta_2)\cdot x}|^2 dx \lesssim |\theta_1 - \theta_2|^2. \tag{134}$$

We now cover the cube $C_K = \{\theta : |\theta|_\infty \leq K\}$ with N_1 frequencies $\nu_1, ..., \nu_{N_1}$ such that for every $\theta \in C_K$, there exists an i with

$$\|e^{2\pi \mathrm{i}\theta \cdot x} - e^{2\pi \mathrm{i}\nu_i \cdot x}\|_{L^2(\Omega)} \leq \frac{\delta}{18nM}. \tag{135}$$

By the above calculation, this can be done with

$$N_1 \lesssim (KnM\delta^{-1})^d \lesssim R^{(2s + \frac{5}{2}d - t)d}, \tag{136}$$

where here we have taken into account the dependence of K, n and δ on R.

Further, we consider the cube

$$A_M = \{\vec{a} = (a_1, ..., a_n) : |a_i| \leq M\},$$

which we can cover with N_2 elements $\vec{a}_1, ..., \vec{a}_{N_2}$ such that for every $\vec{a} \in A_M$, there is an index i with $|\vec{a} - \vec{a}_i|_1 \leq \frac{\delta}{18}$. We can do this with

$$N_2 \lesssim (Mn\delta^{-1})^{2n} \lesssim M^{2R^{d-t}} R^{2(s + \frac{3d}{2} - t)R^{d-t}}, \tag{137}$$

where the $2n$ is because the components of $\vec{a}$ can be complex, we have expanded δ and n in terms of R, and used that fact that if each component differs by $\delta/18n$, then the ℓ^1-norm differs by at most $\delta/18$ as well.

Given a

$$\phi_{i,n}^K = \sum_{j \in T_i^K} a_{i,j} e^{2\pi \mathrm{i}\theta_{i,j} \cdot x}, \tag{138}$$

we proceed to perturb each of the $\theta_{i,j}$ to one of the frequencies ν_i and the coefficients $a_{i,j}$ to one of the $\vec{a}_j$. By the preceding analysis, we can thus land at one of

$$\bar{N} = N_2 N_1^n \lesssim R^{(2s + \frac{5}{2}d - t)dR^{d-t}} M^{2R^{d-t}} R^{2(s + \frac{3d}{2} - t)R^{d-t}} \tag{139}$$

elements by perturbing $\phi_{i,n}^K$ by at most $M\frac{\delta}{18M} + \frac{\delta}{18} = \frac{\delta}{9}$. Since for $i \neq j$, $\phi_{i,n}^K$ and $\phi_{j,n}^K$ differ by at least $\frac{\delta}{3}$ from (133), we see that they must all land at distinct elements after this perturbation. This implies that

$$N = 2^{cR^d} \leq \bar{N} \lesssim R^{(2s+\frac{5}{2}d-t)dR^{d-t}} M^{2R^{d-t}} R^{2(s+\frac{3d}{2}-t)R^{d-t}}. \tag{140}$$

Taking the logarithm and keeping only the dependence on R, we get

$$R^d \lesssim (\log(R)+1)R^{d-t}, \tag{141}$$

which yields a contradiction by taking $R \to \infty$, since $t > 0$. □

Finally, we consider lower bounds for ReLU^k notworks. In the following theorem, we show that the maximal rate of $k-m+1$ obtained in Theorem 3 cannot be improved upon when approximating from Σ_n^k regardless of the level of smoothness s. The argument is relatively straightforward and simply reduces to the one-dimensional case.

Theorem 7. *Let $\Omega = [0,1]^d$, $s \geq 0$, $k \in \mathbb{Z}_{\geq 0}$, and $0 \leq m \leq s$. Then there is a function $f \in \mathscr{B}^s(\Omega)$, such that*

$$\inf_{f_n \in \Sigma_n^k} \|f - f_n\|_{H^m(\Omega)} \gtrsim n^{m-(k+1)}. \tag{142}$$

Proof. Consider the function

$$f(x) = e^{2\pi i x_1} \in \mathscr{B}^s(\Omega). \tag{143}$$

This is a function of only one coordinate, which allows us to extend the lower bound in one dimension obtained in [36] to higher dimensions. Specifically, we make the following simple observation,

$$\|f - f_n\|_{H^m(\Omega)}^2 \geq \int_{\mathbb{R}^{d-1}} \|e^{2\pi i x_1} - f_n(x_1, x_{>1})\|_{H^m([0,1],dx_1)}^2 dx_{>1}. \tag{144}$$

This holds since derivatives with respect to variables other than x_1 will only increase the norm.

Now, for each possible value of $x_{>1}$, $f_n(\cdot, x_{>1})$ is a one-dimensional piecewise polynomial function with at most n breakpoints. Since $e^{2\pi i x_1}$ is not a piecewise polynomial function, the results in [36] imply that

$$\|e^{2\pi i x_1} - f_n(x_1, x_{>1})\|_{H^m([0,1],dx_1)}^2 \gtrsim n^{-2(k-m+1)}. \tag{145}$$

So we get

$$\begin{aligned} \|f - f_n\|_{H^m(\Omega)}^2 &\geq \int_{\mathbb{R}^{d-1}} \|e^{2\pi i x_1} - f_n(x_1, x_{>1})\|_{H^m([0,1],dx_1)}^2 dx_{>1} \\ &\gtrsim n^{-2(k-m+1)}, \end{aligned} \tag{146}$$

as desired. □

5. Conclusion

We have shown that the approximation rates of neural networks with a cosine activation function or powers of a rectified linear unit as an activation function can be significantly improved beyond $O(n^{-\frac{1}{2}})$ for

sufficiently smooth functions. In relation to the finite element method, this shows that a highly adaptive grid can lead to a significantly improved approximation rate for low degree piecewise polynomial functions.

Further work which remains is understanding how much the approximation rates can be further improved when utilizing deeper networks. In particular, we believe that our techniques can be combined with the methods in [9,15] to obtain better rates for approximation of higher dimensional functions by deep neural networks.

Acknowledgments

We would like to thank Professors Weinan E, Ronald DeVore, and Russel Caflisch for helpful discussions, and Professor Yeonjong Shin for helpful comments on a draft of the manuscript. This work was supported by the Verne M. Willaman Fund, and the National Science Foundation (Grant No. DMS-1819157).

References

[1] J.H. Argyris, I. Fried, D.W. Scharpf, The tuba family of plate elements for the matrix displacement method, Aeronaut. J. 72 (1968) 701–709.
[2] J.G. Attali, G. Pagès, Approximations of functions by a multilayer perceptron: a new approach, Neural Netw. 10 (1997) 1069–1081.
[3] F. Bach, Breaking the curse of dimensionality with convex neural networks, J. Mach. Learn. Res. 18 (2017) 629–681.
[4] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inf. Theory 39 (1993) 930–945.
[5] Y. Bengio, O. Delalleau, N. Le Roux, The curse of highly variable functions for local kernel machines, Adv. Neural Inf. Process. Syst. 18 (2006) 107.
[6] C. Berge, Graphs and Hypergraphs, 1973.
[7] J.H. Bramble, M. Zlámal, Triangular elements in the finite element method, Math. Comput. 24 (1970) 809–820.
[8] S. Brenner, R. Scott, The Mathematical Theory of Finite Element Methods, Vol. 15, Springer Science & Business Media, 2007.
[9] G. Bresler, D. Nagaraj, Sharp representation theorems for relu networks with precise dependence on depth, preprint, arXiv:2006.04048, 2020.
[10] H.J. Bungartz, M. Griebel, Sparse grids, Acta Numer. 13 (2004) 147–269.
[11] E.J. Candes, Ridgelets: Theory and Applications, Stanford University, 1998.
[12] D. Costarelli, R. Spigler, Approximation results for neural network operators activated by sigmoidal functions, Neural Netw. 44 (2013) 101–106.
[13] D. Costarelli, R. Spigler, Approximation by series of sigmoidal functions with applications to neural networks, Ann. Mat. Pura Appl. 1923 (194) (2015) 289–306.
[14] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.
[15] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, G. Petrova, Nonlinear approximation and (deep) relu networks, preprint, arXiv:1905.02199, 2019.
[16] R.A. DeVore, Nonlinear approximation, Acta Numer. 7 (1998) 51–150.
[17] R.A. DeVore, G.G. Lorentz, Constructive Approximation, Vol. 303, Springer Science & Business Media, 1993.
[18] D. Donoho, I. Johnstone, Empirical Atomic Decomposition, 1995, Unpublished manuscript.
[19] W. E, C. Ma, L. Wu, Barron spaces and the compositional function spaces for neural network models, preprint, arXiv: 1906.08039, 2019.
[20] S. Ellacott, Aspects of the numerical analysis of neural networks, Acta Numer. 3 (1994) 145–202.
[21] P. Erdös, P.M. Gruber, J. Hammer, Lattice Points, Longman Scientific & Technical Harlow, 1989.
[22] Y. Gordon, V. Maiorov, M. Meyer, S. Reisner, On the best approximation by ridge functions in the uniform norm, Constr. Approx. 18 (2001) 61–85.
[23] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Netw. 4 (1991) 251–257.
[24] K. Hornik, Some new results on neural network approximation, Neural Netw. 6 (1993) 1069–1072.
[25] K. Hornik, M. Stinchcombe, H. White, P. Auer, Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives, Neural Comput. 6 (1994) 1262–1275.
[26] S.G. Johnson, Saddle-point integration of C^∞ bump functions, Manuscript. Available at http://math.mit.edu/~stevenj/bump-saddle.pdf, 2015.
[27] L.K. Jones, A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Stat. 20 (1992) 608–613.
[28] P.C. Kainen, V. Kurková, Quasiorthogonal dimension of euclidean spaces, Appl. Math. Lett. 6 (1993) 7–10.
[29] J.M. Klusowski, A.R. Barron, Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls, IEEE Trans. Inf. Theory 64 (2018) 7649–7656.
[30] V. Kurková, M. Sanguineti, Bounds on rates of variable-basis and neural-network approximation, IEEE Trans. Inf. Theory 47 (2001) 2659–2665.

[31] V. Kurková, M. Sanguineti, Geometric upper bounds on rates of variable-basis approximation, IEEE Trans. Inf. Theory 54 (2008) 5681–5688.
[32] M.J. Lai, L.L. Schumaker, Spline Functions on Triangulations, Cambridge University Press, 2007, p. 110.
[33] E. Lavretsky, On the geometric convergence of neural approximations, IEEE Trans. Neural Netw. 13 (2002) 274–282.
[34] M. Leshno, V.Y. Lin, A. Pinkus, S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural Netw. 6 (1993) 861–867.
[35] B. Li, S. Tang, H. Yu, Better approximations of high dimensional smooth functions by deep neural networks with rectified power units, preprint, arXiv:1903.05858, 2019.
[36] Q. Lin, H. Xie, J. Xu, Lower bounds of the discretization error for piecewise polynomials, Math. Comput. 83 (2014) 1–13.
[37] G.G. Lorentz, M.v. Golitschek, Y. Makovoz, Constructive Approximation: Advanced Problems, Vol. 304, Springer, 1996.
[38] J. Lu, Y. Lu, M. Wang, A priori generalization analysis of the deep ritz method for solving high dimensional elliptic equations, preprint, arXiv:2101.01708, 2021.
[39] V. Maiorov, On best approximation by ridge functions, J. Approx. Theory 99 (1999) 68–94.
[40] V. Maiorov, Best approximation by ridge functions in l p-spaces, Ukr. Math. J. 62 (2010) 452–466.
[41] Y. Makovoz, Random approximants and neural networks, J. Approx. Theory 85 (1996) 98–109.
[42] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: ICML, 2010.
[43] Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, IEEE, 1993, pp. 40–44.
[44] P.P. Petrushev, Approximation by ridge functions and neural networks, SIAM J. Math. Anal. 30 (1998) 155–189.
[45] G. Pisier, Remarques sur un résultat non publié de b. maurey, in: Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz"), 1981, pp. 1–12.
[46] R.T. Rockafellar, Convex Analysis, Princeton University Press, 1970, p. 28.
[47] J.W. Siegel, J. Xu, Approximation rates for neural networks with general activation functions, Neural Netw. 128 (2020) 313–321.
[48] J.W. Siegel, J. Xu, Characterization of the variation spaces corresponding to shallow neural networks, preprint, arXiv: 2106.15002, 2021.
[49] J.W. Siegel, J. Xu, Improved convergence rates for the orthogonal greedy algorithm, preprint, arXiv:2106.15000, 2021.
[50] J.W. Siegel, J. Xu, Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks, preprint, arXiv:2101.12365, 2021.
[51] J. Xu, Finite neuron method and convergence analysis, Commun. Comput. Phys. 28 (2020) 1707–1745, https://doi.org/10.4208/cicp.OA-2020-0191, http://global-sci.org/intro/article_detail/cicp/18394.html.
[52] D. Yarotsky, Error bounds for approximations with deep relu networks, Neural Netw. 94 (2017) 103–114.
[53] D. Yarotsky, Optimal approximation of continuous functions by very deep relu networks, preprint, arXiv:1802.03620, 2018.
[54] D. Yarotsky, A. Zhevnerchuk, The phase diagram of approximation rates for deep neural networks, Adv. Neural Inf. Process. Syst. 33 (2020).
[55] A. Ženíšek, Interpolation polynomials on the triangle, Numer. Math. 15 (1970) 283–296.