Deep Neural Network Approximation Theory

Philipp Grohs, Dmytro Perekrestenko, Dennis Elbrächter, and Helmut Bölcskei

Abstract

Deep neural networks have become state-of-the-art technology for a wide range of practical machine learning tasks such as image classification, handwritten digit recognition, speech recognition, or game intelligence. This paper develops the fundamental limits of learning in deep neural networks by characterizing what is possible if no constraints on the learning algorithm and the amount of training data are imposed. Concretely, we consider information-theoretically optimal approximation through deep neural networks with the guiding theme being a relation between the complexity of the function (class) to be approximated and the complexity of the approximating network in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights. The theory we develop educes remarkable universality properties of deep networks. Specifically, deep networks are optimal approximants for vastly different function classes such as affine systems and Gabor systems. Affine systems are generated by the affine group (scalings and translations) whereas Gabor systems are generated by the Weyl-Heisenberg group (time-shifts and modulations). This universality is afforded by a concurrent invariance property of deep networks to time-shifts, scalings, and frequency-shifts. In addition, deep networks provide exponential approximation accuracy—i.e., the approximation error decays exponentially in the number of non-zero weights in the network—of vastly different functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures and fractal functions such as the Weierstrass function, both of which do not have any known methods achieving exponential approximation accuracy. In summary, deep neural networks provide information-theoretically optimal approximation of a very wide range of functions and function classes used in mathematical signal processing. We also show that in the approximation of sufficiently smooth functions finite-width deep networks require strictly smaller connectivity than finite-depth wide networks.

I. Introduction

Triggered by the availability of vast amounts of training data and drastic improvements in computing power, deep neural networks have become state-of-the-art technology for a wide range of practical machine learning tasks

P. Grohs is with the Department of Mathematics and the Research Platform DataScience@UniVienna, University of Vienna, Austria (e-mail: philipp.grohs@univie.ac.at).

D. Elbrächter is with the Department of Mathematics, University of Vienna, Austria (e-mail: dennis.elbraechter@univie.ac.at).

D. Perekrestenko is with the Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland (e-mail: pdmytro@nari.ee.ethz.ch).

H. Bölcskei is with the Department of Information Technology and Electrical Engineering and the Department of Mathematics, ETH Zurich, Switzerland (e-mail: hboelcskei@ethz.ch).

D. Perekrestenko and H. Bölcskei were supported in part by a gift from Huawei's Future Network Theory Lab. D. Elbrächter was supported by the Austrian Science Fund via the project P 30148.

such as image classification [1], handwritten digit recognition [2], speech recognition [3], or game intelligence [4]. For an in-depth overview, we refer to the survey paper [5] and the recent book [6].

A neural network effectively implements a mapping approximating a function which is learned based on a given set of input-output value pairs, typically through the backpropagation algorithm [7]. Characterizing the fundamental limits of approximation through neural networks shows what is possible if no constraints on the learning algorithm and on the amount of training data are imposed [8].

It is well known that single-hidden-layer neural networks can approximate continuous functions on bounded domains arbitrarily well, provided that the activation function satisfies certain (mild) conditions and the number of nodes is allowed to grow arbitrarily large [9], [10], [11]. In practice one is, however, often interested in approximating functions from a given function class $\mathcal{C}$ determined by the application at hand. It is therefore natural to ask how the complexity of a neural network approximating every function in $\mathcal{C}$ to within a prescribed accuracy depends on the complexity of $\mathcal{C}$ (and on the desired approximation accuracy). The recently developed Kolmogorov rate-distortion-theoretic approach [12] formalizes this question by relating the complexity of $\mathcal{C}$—in terms of the number of bits needed to describe any element in $\mathcal{C}$ to within prescribed accuracy—to network complexity in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights.

The purpose of this paper is to provide a comprehensive, principled, and self-contained introduction to Kolmogorov rate-distortion optimal approximation through deep neural networks. The idea is to equip the reader with a working knowledge of the mathematical tools underlying the theory at a level that is sufficiently deep to enable further own research in the field. Part of this paper is based on [12], but extends the theory therein to the rectified linear unit (ReLU) activation function and to networks with depth scaling in the approximation error.

The theory we develop educes remarkable universality properties of finite-width deep networks. Specifically, deep networks are optimal approximants for vastly different function classes such as affine systems [12] and Gabor systems and local cosine bases [13], [14]. Affine systems are generated by the affine group (scalings and translations) whereas Gabor systems and local cosine bases are generated by the Weyl-Heisenberg group (time-shifts and modulations). This universality is afforded by a concurrent invariance property of deep networks to time-shifts, scalings, and frequency-shifts. In addition, deep networks provide exponential approximation accuracy—i.e., the approximation error decays exponentially in the number of parameters employed in the approximant, namely the number of non-zero weights in the network—of vastly different functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures [15] and fractal functions such as the Weierstrass function, both of which do not have any known methods achieving exponential approximation accuracy. In summary, deep neural networks provide optimal approximation of a very wide range of functions and function classes used in mathematical signal processing.

While we consider networks based on the ReLU activation function throughout, the parts of our theory not involving sinusoidal functions, i.e., everything apart from Sections IV and VIII, can be shown to also apply to

strongly sigmoidal activation functions of order $k \geq 2$ as defined in [12]. The key to this extension is to note that the result on the neural network approximation of the multiplication function according to Theorem III.2 holds for strongly sigmoidal activation functions of order $k \geq 2$ as well. The rest of the theory then follows mutatis mutandis. We do not provide the details here for the sake of conciseness.

Notation. For the function $f(x)\colon \mathbb{R}^d \to \mathbb{R}$, we define $\|f\|_{L^\infty(\Omega)} := \inf\{C \geq 0 : |f(x)| \leq C, \text{ for all } x \in \Omega\}$. $L^p(\mathbb{R}^d)$ and $L^p(\mathbb{R}^d, \mathbb{C})$ denote the space of real-valued, respectively complex-valued, L^p-functions. For a vector $b \in \mathbb{R}^d$, we let $\|b\|_\infty := \max_{i=1,\ldots,d} |b_i|$, similarly we write $\|A\|_\infty := \max_{i,j} |A_{i,j}|$ for the matrix $A \in \mathbb{R}^{m\times n}$. We denote the identity matrix of size $n \times n$ by $\mathbb{I}_n$. Throughout, $\log$ stands for the logarithm to base 2. For a set $X \in \mathbb{R}^d$, we write $|X|$ for its Lebesgue measure.

II. Setup and basic ReLU calculus

There is a plethora of neural network architectures and activation functions in the literature. Here, we restrict ourselves to the ReLU activation function and consider the following general network architecture.

Definition II.1. *Let $L, N_0, N_1, \ldots, N_L \in \mathbb{N}$, $L \geq 2$. A map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by*

$$\Phi(x) = \begin{cases} W_2(\rho\,(W_1(x))), & L = 2 \\ W_L(\rho\,(W_{L-1}(\rho\,(\ldots\rho\,(W_1(x)))))), & L \geq 3 \end{cases}, \tag{1}$$

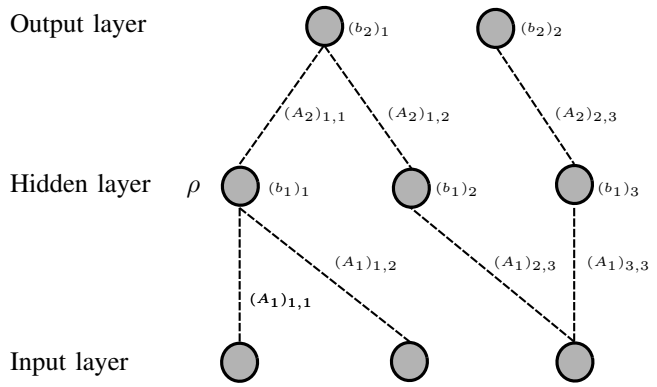
with affine linear maps $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L\}$, and the ReLU activation function $\rho(x) = \max(x, 0)$, $x \in \mathbb{R}$, acting component-wise (i.e., $\rho(x_1, \ldots, x_N) := (\rho(x_1), \ldots, \rho(x_N))$) is called a ReLU neural network*. The map W_ℓ corresponding to layer ℓ is given by $W_\ell(x) = A_\ell x + b_\ell$, with $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$. Throughout the paper, we shall write $L \in \mathbb{N}$ to mean that the corresponding statement applies to networks with $L \in \mathbb{N}, L \geq 2$ layers. We define the* network connectivity *$\mathcal{M}(\Phi)$ as the total number of non-zero entries in the matrices A_ℓ, $\ell \in \{1, 2, \ldots, L\}$, and the vectors b_ℓ, $\ell \in \{1, 2, \ldots, L\}$. The* depth of the network *or, equivalently, the number of layers is $\mathcal{L}(\Phi) := L$ and its width $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$. We further denote by $\mathcal{B}(\Phi) := \max_{\ell=1,\ldots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$ the maximum absolute value of the weights in the network.*

Remark II.2. *We note that the connectivity satisfies $\mathcal{M}(\Phi) \leq \mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi) + 1)$.*

Remark II.3. *N_0 in Definition II.1 is the* dimension of the input layer*, $N_1, \ldots, N_{L-1}$ are the* dimensions of the $L - 1$ hidden layers*, and N_L is the* dimension of the output layer*. Note that our definition of $\mathcal{L}(\Phi)$ does not take into account the input layer, which we consider to be the '0-th' layer. The matrix entry $(A_\ell)_{i,j}$ represents the* weight associated with the edge between the j-th node in the $(\ell - 1)$-th layer and the i-th node in the ℓ-th layer*, $(b_\ell)_i$ is* the weight associated with the i-th node in the ℓ-th layer*. These assignments are schematized in Figure 1. The real numbers $(A_\ell)_{i,j}$ and $(b_\ell)_i$ are referred to as the network's edge weights and node weights, respectively. Throughout the paper, we assume that every node in the input layer and in layers $1, \ldots, L - 1$ has at least one outgoing edge*

and every node in the output layer L has at least one incoming edge. These non-degeneracy assumptions are basic as nodes that do not satisfy them can be removed without changing the functional relationship realized by the network.

The term "network" stems from the interpretation of the mapping Φ as a weighted acyclic directed graph with nodes arranged in hierarchical layers and edges only between adjacent layers.



$$A_2 = \begin{pmatrix} (A_2)_{1,1} & (A_2)_{1,2} & 0 \\ 0 & 0 & (A_2)_{2,3} \end{pmatrix}$$

$$A_1 = \begin{pmatrix} (A_1)_{1,1} & (A_1)_{1,2} & 0 \\ 0 & 0 & (A_1)_{2,3} \\ 0 & 0 & (A_1)_{3,3} \end{pmatrix}$$

Fig. 1: Assignment of the weights $(A_\ell)_{i,j}$ and $(b_\ell)_i$ of a two-layer network to the edges and nodes, respectively.

We denote the class of ReLU networks $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ with no more than L layers, connectivity no more than M, input dimension d, and output dimension N_L by $\mathcal{NN}_{L,M,d,N_L}$. Moreover, we let

$$\mathcal{NN}_{\infty,M,d,N_L} := \bigcup_{L\in\mathbb{N}} \mathcal{NN}_{L,M,d,N_L}, \quad \mathcal{NN}_{L,\infty,d,N_L} := \bigcup_{M\in\mathbb{N}} \mathcal{NN}_{L,M,d,N_L}, \quad \mathcal{NN}_{\infty,\infty,d,N_L} := \bigcup_{L\in\mathbb{N}} \mathcal{NN}_{L,\infty,d,N_L}.$$

Throughout the paper, we consider almost exclusively the case $\Phi : \mathbb{R}^d \to \mathbb{R}$, i.e., $N_L = 1$. We emphasize, however, that the results readily generalize to $N_L > 1$.

Now, given a function $f : \mathbb{R}^d \to \mathbb{R}$, we are interested in the best possible approximation of f by a network Φ under various constraints on the topology and the weights of Φ. Specifically, we will need the notion of "polynomially bounded weights" of families of networks, which we make precise as follows.

Definition II.4. *We say that the (edge and node) weights of a family of neural networks Φ_z, $z = (z_1, z_2, \ldots, z_N) \in D \subseteq \mathbb{R}^N$, are* polynomially bounded *in $z_1, z_2, \ldots, z_N$ if there exists an N-variate polynomial π such that $\mathcal{B}(\Phi_z) \leq \pi(z_1, z_2, \ldots, z_N)$, for all $z \in D$. We further say that the weights of a family of neural networks $\Phi_{z,j}$, $z \in D \subseteq \mathbb{R}^N$, $j \in J$, are* uniformly (w.r.t. $j \in J$) polynomially bounded *in $z_1, z_2, \ldots, z_N$ if there exists an N-variate polynomial π such that $\mathcal{B}(\Phi_{z,j}) \leq \pi(z_1, z_2, \ldots, z_N)$, for all $z \in D$, $j \in J$.*

A neural network Φ as defined in (1) realizes a function $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$. We note, however, that for a given function there may be different choices of parameters $L, N_0, N_1, \ldots, N_L \in \mathbb{N}$ and affine mappings $W_1, W_2, ..., W_L$ that realize this function through a neural network [16]. Despite this dichotomy it makes sense to speak of

compositions and linear combinations of neural networks. To this end, we first record a technical lemma on the composition of neural networks as defined in [17].

Lemma II.5. *Let* $L_1, L_2, M_1, M_2, d_1, d_2, N_{L_1}, N_{L_2} \in \mathbb{N}$, $\Phi_1 \in \mathcal{NN}_{L_1,M_1,d_1,N_{L_1}}$, *and* $\Phi_2 \in \mathcal{NN}_{L_2,M_2,d_2,N_{L_2}}$ *with* $N_{L_1} = d_2$. *Then, there exists a network* $\Psi \in \mathcal{NN}_{L_1+L_2,2M_1+2M_2,d_1,N_{L_2}}$ *with* $\mathcal{W}(\Psi) \leq \max\{2N_{L_1}, \mathcal{W}(\Phi_1), \mathcal{W}(\Phi_2)\}$ *and* $\mathcal{B}(\Psi) = \max\{\mathcal{B}(\Phi_1), \mathcal{B}(\Phi_2)\}$, *satisfying* $\Psi(x) = \Phi_2(\Phi_1(x))$, *for all* $x \in \mathbb{R}^{d_1}$.

Proof. The proof is based on the identity $x = \rho(x) - \rho(-x)$. First, note that by Definition II.1, we can write

$$\Phi_1(x) = W^1_{L_1}(\,\rho\,(\ldots W^1_1(x))) \text{ and } \Phi_2(x) = W^2_{L_2}(\,\rho\,(\ldots W^2_1(x))).$$

Next, define the affine map given by $\widetilde{W}(x) = W^2_1\left(\begin{pmatrix}\mathbb{I}_{N_{L_1}} & -\mathbb{I}_{N_{L_1}}\end{pmatrix} x\right)$, for $x \in \mathbb{R}^{2N_{L_1}}$, and note that thanks to

$$W^2_1(\Phi_1(x)) = \widetilde{W}\left(\rho\left(\begin{pmatrix} W^1_{L_1} \\ -W^1_{L_1}\end{pmatrix}(\,\rho\,(\ldots W^1_1(x)))\right)\right),$$

the map

$$\Psi(x) = W^2_{L_2}\left(\rho\left(\ldots W^2_2\left(\rho\left(\widetilde{W}\left(\rho\left(\begin{pmatrix} W^1_{L_1} \\ -W^1_{L_1}\end{pmatrix}(\,\rho\,(\ldots W^1_1(x)))\right)\right)\right)\right)\right)\right)$$

satisfies $\Psi(x) = \Phi_2(\Phi_1(x))$, for all $x \in \mathbb{R}^{d_1}$, with $\mathcal{L}(\Psi) = L_1 + L_2$, $\mathcal{M}(\Psi) \leq 2M_1 + 2M_2$, $\mathcal{W}(\Psi_1) \leq \max\{2N_{L_1}, \mathcal{W}(\Phi_1), \mathcal{W}(\Phi_2)\}$, and $\mathcal{B}(\Psi) \leq \max\{\mathcal{B}(\Phi_1), \mathcal{B}(\Phi_2)\}$. □

Before we can formalize the concept of a linear combination of neural networks, we need a result that shows how to augment network depth while retaining the network's input-output relation.

Lemma II.6. *Let* $L, M, K, d \in \mathbb{N}$, $\Phi_1 \in \mathcal{NN}_{L,M,d,1}$, *and* $K > L$. *Then, there exists a corresponding network* $\Phi_2 \in \mathcal{NN}_{K,M+\mathcal{W}(\Phi_1)+2(K-L)+1,d,1}$ *such that* $\Phi_2(x) = \Phi_1(x)$, *for all* $x \in \mathbb{R}^d$. *Moreover,* $\mathcal{W}(\Phi_2) = \max\{2, \mathcal{W}(\Phi_1)\}$ *and the weights of* Φ_2 *consist of the weights of* Φ_1 *and* ± 1*'s.*

Proof. The proof is based on the identity $x = \rho(x) - \rho(-x)$. First, note that by (1) we can write $\Phi_1(x) = W_L(\,\rho\,(\ldots W_1(x)))$. For $K = L + 1$, Φ_2 is given by

$$\Phi_2(x) = \begin{pmatrix}1 & -1\end{pmatrix}\rho\left(\begin{pmatrix} W_L \\ -W_L\end{pmatrix}(\,\rho\,(\ldots W_1(x)))\right) \in \mathcal{NN}_{L+1,M+\mathcal{W}(\Phi_1)+3,d,1}. \tag{2}$$

For $K > L + 1$, consider the network

$$\Phi'_1(x) = \begin{pmatrix}\rho(\Phi_1(x)) \\ \rho(-\Phi_1(x))\end{pmatrix} = \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\rho\left(\begin{pmatrix} W_L \\ -W_L\end{pmatrix}(\,\rho\,(\ldots W_1(x)))\right) \in \mathcal{NN}_{L+1,M+\mathcal{W}(\Phi_1)+3,d,2}, \tag{3}$$

which satisfies $\mathcal{W}(\Phi'_1) = \max\{2, \mathcal{W}(\Phi_1)\}$. Next, we note that for every network of the form $\Psi(x) = \mathbb{I}_2\,\rho\,(\ldots)$, the network

$$\Psi'(x) := \mathbb{I}_2\,\rho\,(\Psi(x)), \tag{4}$$

satisfies $\Psi'(x) = \Psi(x)$, for all $x \in \mathbb{R}^d$, $\mathcal{L}(\Psi') = \mathcal{L}(\Psi) + 1$, and $\mathcal{M}(\Psi') = \mathcal{M}(\Psi) + 2$. Moreover, the weights of Ψ' consist of the weights of Ψ and $\{1\}$. Noting that Φ_1' in (3) is of the form $\mathbb{I}_2 \, \rho(\dots)$ and iteratively applying the operation in (4) $K - L - 2$ times to Φ_1', we obtain a network $\Phi_1'' \in \mathcal{NN}_{K-1, M+\mathcal{W}(\Phi_1)+2(K-L)-1, d, 2}$. The proof is concluded by noting that $\Phi_2 = (1 \;\; -1)\rho(\Phi_1'') \in \mathcal{NN}_{K, M+\mathcal{W}(\Phi_1)+2(K-L)+1, d, 1}$ satisfies $\Phi_2(x) = \Phi_1(x)$, for all $x \in \mathbb{R}^d$. □

The next result formalizes the concept of a linear combination of neural networks.

Lemma II.7. *Let $N, L_i, M_i, d_i \in \mathbb{N}$, $a_i \in \mathbb{R}$, $\Phi_i \in \mathcal{NN}_{L_i, M_i, d_i, 1}$, $i = 1, 2, \dots, N$, $d = \sum_{i=1}^N d_i$. Then, there exist networks $\Phi^1 \in \mathcal{NN}_{L, M, d, N}$ and $\Phi^2 \in \mathcal{NN}_{L, M+N, d, 1}$ with $L = \max_i L_i$, $\mathcal{W}(\Phi^1) = \mathcal{W}(\Phi^2) \leq \sum_{i=1}^N \max\{2, \mathcal{W}(\Phi_i)\}$, and $M = \sum_{i=1}^N (M_i + \mathcal{W}(\Phi_i) + 2(L - L_i) + 1)$ satisfying*

$$\Phi^1(x) = (a_1 \Phi_1(x_1) \quad a_2 \Phi_2(x_2) \quad \dots \quad a_N \Phi_N(x_N))^T \quad \text{and}$$
$$\Phi^2(x) = \sum_{i=1}^N a_i \Phi_i(x_i),$$

for all $x = (x_1^T, x_2^T, \dots, x_N^T)^T \in \mathbb{R}^d$ with $x_i \in \mathbb{R}^{d_i}$, $i = 1, 2, \dots, N$. Moreover, the weights of Φ^1, Φ^2 consist of the weights of the Φ_i, $i = 1, 2, \dots, N$, $\{a_1, a_2, ..., a_N\}$, and ± 1's.

Proof. Apply Lemma II.6 to the networks Φ_i to get corresponding networks $\tilde{\Phi}_i$ of depth L and set $\Phi^1(x) := \big(a_1 \tilde{\Phi}_1(x_1), \, a_2 \tilde{\Phi}_2(x_2), \, \dots, \, a_N \tilde{\Phi}_N(x_N)\big)^\top$, $\Phi^2(x) := (1, \, 1, \, \dots, \, 1) \Phi^1(x)$. □

III. Approximation of multiplication, polynomials, and smooth functions

We start with the approximation of the multiplication operation by deep ReLU networks and then proceed to the approximation of polynomials. Specifically, we shall be interested in networks that approximate to within error ε, are of finite width, and have depth scaling poly-logarithmically in ε^{-1} (i.e., as a polynomial in $\log(\varepsilon^{-1})$) and (edge and node) weights that do not grow faster than polynomially in the size of the domain over which approximation takes place. It will be shown in Section V that this combination of requirements leads to exponential approximation accuracy for individual signals, i.e., the approximation error decays exponentially in the number of nodes in the network, and to rate-distortion optimal approximation of signal classes.

The proof ideas for the results in this section are partly inspired by [18] and the "sawtooth" construction of [19]. In contrast to [18], we consider networks without "skip connections" and of finite and explicitly specified width.

Proposition III.1. *There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{NN}_{\infty, \infty, 1, 1}$ satisfying $\mathcal{L}(\Phi_\varepsilon) \leq C \log(\varepsilon^{-1})$, $\mathcal{W}(\Phi_\varepsilon) = 4$, $\mathcal{B}(\Phi_\varepsilon) \leq 4$, $\Phi_\varepsilon(0) = 0$, and*

$$\|\Phi_\varepsilon(x) - x^2\|_{L^\infty([0,1])} \leq \varepsilon. \tag{5}$$

Proof. Consider the function $g : [0, 1] \to [0, 1]$,

$$g(x) = \begin{cases} 2x, & \text{if } x < \frac{1}{2}, \\ 2(1-x), & \text{if } x \geq \frac{1}{2}, \end{cases} \tag{6}$$

along with the "sawtooth" functions given by its s-fold composition

$$g_s := \underbrace{g \circ g \circ \cdots \circ g}_{s}, \quad s \geq 2, \tag{7}$$

and set $g_0(x) := x, g_1(x) := g(x)$. We next briefly review a fundamental result from [18] showing how the function $f(x) := x^2, x \in [0,1]$, can be approximated by linear combinations of "sawtooth" functions g_s. Specifically, let f_m be the piecewise linear interpolation of f with $2^m + 1$ uniformly spaced "knots" according to

$$f_m\Big(\frac{k}{2^m}\Big) = \Big(\frac{k}{2^m}\Big)^2, \quad k = 0, \ldots, 2^m, \quad m \in \mathbb{N}_0.$$

The function f_m approximates f with error $\varepsilon_m = 2^{-2m-2}$ in the sense of

$$\|f_m(x) - x^2\|_{L^\infty[0,1]} \leq 2^{-2m-2}.$$

Next, note that we can refine interpolation in the sense of going from f_{m-1} to f_m by adjustment with a sawtooth function according to

$$f_m(x) = f_{m-1}(x) - \frac{g_m(x)}{2^{2m}}. \tag{8}$$

This leads to the representation

$$f_m(x) = x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}. \tag{9}$$

While Yarotsky's construction [18] is finalized by realizing (9) through a deep ReLU network of width 3 with the help of skip connections [20], i.e., connections between nodes in non-consecutive layers, we proceed by constructing an equivalent (in terms of input-output relation) network without skip connections and of width 4. As $g(x) = 2\rho(x) - 4\rho(x - 1/2) + 2\rho(x - 1)$, it follows that

$$g_m = 2\rho(g_{m-1}) - 4\rho(g_{m-1} - 1/2) + 2\rho(g_{m-1} - 1), \tag{10}$$

and since $f_m = \rho(f_m), \forall m \in \mathbb{N}_0$, (8) can be rewritten as

$$f_m = \rho(f_{m-1}) - 2^{-2m}\Big(2\rho(g_{m-1}) - 4\rho(g_{m-1} - 1/2) + 2\rho(g_{m-1} - 1)\Big). \tag{11}$$

Equivalently, (10) and (11) can be cast as a composition of affine linear maps and a ReLU nonlinearity according to

$$\begin{pmatrix} g_m \\ f_m \end{pmatrix} = W_1 \left(\rho\left(W_2 \begin{pmatrix} g_{m-1} \\ f_{m-1} \end{pmatrix} \right) \right), \tag{12}$$

with

$$W_1(x) = \begin{pmatrix} 2 & -4 & 2 & 0 \\ -2^{-2m+1} & 2^{-2m+2} & -2^{-2m+1} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \quad W_2(x) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 0 \\ 1/2 \\ 1 \\ 0 \end{pmatrix}. \tag{13}$$

Applying (12) iteratively initialized with $g_0(x) = x$, $f_0(x) = x$ yields

$$\begin{pmatrix} g_m \\ f_m \end{pmatrix} = W_1 \left(\rho \left(W_2 \left(W_1 \left(\rho \left(\dots \rho \left(W_2 \left(W_1 \left(\rho \left(W_2 \begin{pmatrix} x \\ x \end{pmatrix} \right) \right) \right) \right) \right) \right) \right) \right) \right), \tag{14}$$

and hence shows that f_m can be realized through a network in $\mathcal{NN}_{m+1,13m,1,1}$ of width 4 and with weights bounded (in magnitude) by 4. Since $\varepsilon_m = 2^{-2m-2}$ and hence $\log(1/\varepsilon_m) = 2m + 2$, the statement follows upon noting that $f_m(0) = 0, \forall m \in \mathbb{N}_0$. □

With Proposition III.1 we are now ready to show how the multiplication operation can be realized through deep ReLU networks. This will then lead us to the approximation of arbitrary polynomials.

Proposition III.2. *There exists a constant $C > 0$ such that for all $D \in \mathbb{R}_+$ and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,2,1}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}) \leq C \log(\lceil D \rceil^2 \varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}) \leq 12$, $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \max\{4, 2\lceil D \rceil^2\}$, $\Phi_{D,\varepsilon}(0, x) = \Phi_{D,\varepsilon}(x, 0) = 0$, for all $x \in \mathbb{R}$, and*

$$\|\Phi_{D,\varepsilon}(x, y) - xy\|_{L^\infty([-D,D]^2)} \leq \varepsilon. \tag{15}$$

Proof. The proof is based on the identity

$$xy = \frac{1}{2}((x + y)^2 - x^2 - y^2), \tag{16}$$

which shows how multiplication can be implemented through the squaring operation. We need a ReLU network realization of x^2 and y^2 over $[-D, D]$ and of $(x + y)^2$ over $[-D, D]^2$. To this end, take $\Psi_\delta(x)$ to be a neural network approximating x^2 according to Proposition III.1, i.e., $\|\Psi_\delta(x) - x^2\|_{L^\infty([0,1])} \leq \delta$, $\Psi_\delta(0) = 0$. Next, note that

$$\left\| 4\lceil D \rceil^2 \Psi_\delta \left(\frac{|x|}{2\lceil D \rceil} \right) - x^2 \right\|_{L^\infty([-D,D])} \leq 4\lceil D \rceil^2 \delta, \tag{17}$$

and likewise

$$\left\| 4\lceil D \rceil^2 \Psi_\delta \left(\frac{|x + y|}{2\lceil D \rceil} \right) - (x + y)^2 \right\|_{L^\infty([-D,D]^2)} \leq 4\lceil D \rceil^2 \delta. \tag{18}$$

The network $x \mapsto \Psi_\delta(|x|)$ has one layer more than the network $x \mapsto \Psi_\delta(x)$ as it implements $|x| = \rho(x) + \rho(-x)$ in its first layer. Next, we define for $D \in \mathbb{R}_+$, $\delta \in (0, 1/2)$,

$$\Phi^*_{D,\delta}(x, y) := 2\lceil D \rceil^2 \left(\Psi_\delta \left(\frac{|x + y|}{2\lceil D \rceil} \right) - \Psi_\delta \left(\frac{|x|}{2\lceil D \rceil} \right) - \Psi_\delta \left(\frac{|y|}{2\lceil D \rceil} \right) \right), \tag{19}$$

and observe that by Lemma II.7, there exists a constant $C > 0$ such that $\mathcal{L}(\Phi^*_{D,\delta}) \leq C \log(\delta^{-1})$, $\mathcal{W}(\Phi^*_{D,\delta}) \leq 12$, $\mathcal{B}(\Phi^*_{D,\delta}) \leq \max\{4, 2\lceil D \rceil^2\}$, and $\Phi^*_{D,\delta}(x, 0) = \Phi^*_{D,\delta}(0, x) = 0$, for all $D \in \mathbb{R}_+$, $\delta \in (0, 1/2)$, $x \in \mathbb{R}$. Using (16)

in combination with (17) and (18), we get

$$\left\| \Phi^*_{D,\delta}(x,y) - \frac{1}{2}\Big((x+y)^2 - x^2 - y^2\Big) \right\|_{L^\infty([-D,D]^2)} \leq 6\lceil D\rceil^2 \delta.$$

The proof is completed by setting, for $D \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, $\Phi_{D,\varepsilon} = \Phi^*_{D,\delta_{D,\varepsilon}}$ with $\delta_{D,\varepsilon} := \frac{\varepsilon}{6\lceil D\rceil^2}$. □

An approach similar to that used in the proof of Proposition III.2 was developed previously in [21] to show that the multiplication operation and the gradient of multiplication can both be approximated by networks of finite width. The networks in [21] are of the same width and exhibit the same depth scaling as those constructed here. Proposition III.2 also makes the dependence of the approximating network's depth on D explicit and provides a bound on the absolute value of the weights in the network.

Now that we know how to approximate the squaring operation and multiplication by deep ReLU networks, we can realize arbitrary powers of x through the composition of squaring and multiplication networks and arbitrary polynomials by taking weighted linear combinations of powers of x according to Lemma II.7. Specifically, we shall show how polynomials can be approximated by ReLU networks of finite width and of depth growing logarithmically in the inverse of the approximation error.

Proposition III.3. *There exists a constant $C > 0$ such that for all $m \in \mathbb{N}$, $A \in \mathbb{R}_+$, $p_m(x) = \sum_{i=0}^m a_i x^i$ with $\max_{i=0,\ldots,m} |a_i| = A$, $D \in \mathbb{R}_+$, and $\varepsilon \in (0,1/2)$, there is a network $\Phi_{p_m,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Phi_{p_m,D,\varepsilon}) \leq Cm(\log(\lceil A\rceil) + \log(\varepsilon^{-1}) + m\log(\lceil D\rceil) + \log(m))$, $\mathcal{W}(\Phi_{p_m,D,\varepsilon}) \leq 16$, $\mathcal{B}(\Phi_{p_m,D,\varepsilon}) \leq \max\{A, 8\lceil D\rceil^{2m-2}\}$, and*

$$\|\Phi_{p_m,D,\varepsilon} - p_m\|_{L^\infty([-D,D])} \leq \varepsilon. \tag{20}$$

Proof. We start by noting that for $m = 1$ the resulting affine function $p_1(x) = a_0 + a_1 x$ can be realized exactly, i.e., with $\varepsilon = 0$, by a network of depth $L = 2$ with

$$W_1(x) = \begin{pmatrix} a_1 \\ -a_1 \end{pmatrix} x + \begin{pmatrix} a_0 \\ -a_0 \end{pmatrix}$$

and $A_2 = (1 \ \ -1), b_2 = 0$. The proof for $m \geq 2$ will be effected by realizing the monomials $x^k, k \geq 2$, through iterative composition of multiplication networks and combining this with a construction which uses the network realizing x^k not only as a building block in the network realizing x^{k+1} but also to construct the network approximating the partial sum $\sum_{i=0}^k a_i x^i$ in parallel.

We start by setting $H^k_{D,\eta} := \lceil D\rceil^k + \eta\sum_{s=0}^{k-2}\lceil D\rceil^s$, $k \in \mathbb{N}$, and let $\Phi_{H^k_{D,\eta},\eta}$, $D \in \mathbb{R}_+$, $k \in \mathbb{N}$, $\eta \in (0,1/2)$, be multiplication networks according to Proposition III.2. For $D \in \mathbb{R}_+$, $k \in \mathbb{N}$, $\eta \in (0,1/2)$, we then recursively define $\Psi^k_{D,\eta}$ according to $\Psi^0_{D,\eta}(x) = 1$, $\Psi^1_{D,\eta}(x) = x$, and $\Psi^k_{D,\eta}(x) = \Phi_{H^{k-1}_{D,\eta},\eta}(x, \Psi^{k-1}_{D,\eta}(x)), k \geq 2$. Note that $\Psi^k_{D,\eta}(x)$ can be realized through a neural network for all $k \in \mathbb{N}$ thanks to Lemma II.5 and the fact that, as already noted above for the case $m = 1$, any affine function can be realized through a neural network.

We first show by induction that

$$\|\Psi_{D,\eta}^k(x) - x^k\|_{L^\infty([-D,D])} \leq \eta \sum_{s=0}^{k-2} \lceil D \rceil^s, \tag{21}$$

for all $\eta \in (0, 1/2)$, $k \geq 2$. The base case $k = 2$ follows from

$$\|\Psi_{D,\eta}^2(x) - x^2\|_{L^\infty([-D,D])} = \|\Phi_{H_{D,\eta}^1,\eta}(x, x) - x^2\|_{L^\infty([-D,D])} \leq \eta.$$

We proceed to establishing the induction step $(k-1) \to k$. The induction assumption is

$$\|\Psi_{D,\eta}^{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \leq \eta \sum_{s=0}^{k-3} \lceil D \rceil^s. \tag{22}$$

Since $\|\Psi_{D,\eta}^{k-1}\|_{L^\infty([-D,D])} \leq \|x^{k-1}\|_{L^\infty([-D,D])} + \|\Psi_{D,\eta}^{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \leq H_{D,\eta}^{k-1}$, Proposition III.2 implies that

$$\begin{aligned}
\|\Psi_{D,\eta}^k(x) - x^k\|_{L^\infty([-D,D])} &\leq \|\Phi_{H_{D,\eta}^{k-1},\eta}(x, \Psi_{D,\eta}^{k-1}(x)) - x\Psi_{D,\eta}^{k-1}(x)\|_{L^\infty([-D,D])} \\
&\quad + \max_{[-D,D]} |x| \|\Psi_{D,\eta}^{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \\
&\leq \eta + \lceil D \rceil \eta \sum_{s=0}^{k-3} \lceil D \rceil^s = \eta \sum_{s=0}^{k-2} \lceil D \rceil^s,
\end{aligned}$$

which completes the proof of the induction step.

We are now ready to proceed to the construction of the network $\Phi_{p_m,D,\varepsilon}$ approximating the polynomial $p_m(x) = \sum_{i=0}^m a_i x^i$. To this end, we first note that the identity mapping $x \mapsto x$ and the linear combination $x, y \mapsto x + a_{i-1}y$ are affine transformations and can thus be realized by a network of depth $L = 2$. By Lemma II.7 there hence exists a constant C_2 such that for every $m \geq 2$, $p_m(x) = \sum_{\ell=0}^m a_\ell x^\ell$, $i \in \{2, 3, \ldots, m\}$, $\eta \in (0, 1/2)$ there is a network $\varphi_{p_m,D,\eta}^i \in \mathcal{NN}_{\infty,\infty,3,3}$ with $\mathcal{L}(\varphi_{p_m,D,\eta}^i) \leq C_2 \log(\lceil H_{D,\eta}^{i-1} \rceil^2 \eta^{-1})$, $\mathcal{W}(\varphi_{p_m,D,\eta}^i) \leq 16$, and $\mathcal{B}(\varphi_{p_m,D,\eta}^i) \leq \max\{4, 2\lceil H_{D,\eta}^{i-1} \rceil^2, \max_{i \in \{0,\ldots,m\}} |a_i|\}$ realizing the map

$$(x \quad s \quad y)^\top \to \begin{pmatrix} x & s + a_{i-1}y & \Phi_{H_{D,\eta}^{i-1},\eta}(x, y) \end{pmatrix}^\top.$$

The statements in the following apply for all $m \in \mathbb{N}$, $A \in \mathbb{R}_+$, $p_m(x) = \sum_{i=0}^m a_i x^i$ with $\max_{i=0,\ldots,m} |a_i| \leq A$, $D \in \mathbb{R}_+$, and $\varepsilon \in (0, 1/2)$. The network $\Phi_{p_m,D,\varepsilon}$ approximating the polynomial $p_m(x) = \sum_{i=0}^m a_i x^i$ is now constructed according to

$$\Phi_{p_m,D,\varepsilon}(x) := \begin{pmatrix} 0 & 1 & a_m \end{pmatrix} \varphi_{p_m,D,\eta_\varepsilon}^m \left(\varphi_{p_m,D,\eta_\varepsilon}^{m-1} \left(\cdots \varphi_{p_m,D,\eta_\varepsilon}^2 \left(\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} x + \begin{pmatrix} 0 \\ a_0 \\ 0 \end{pmatrix} \right) \right) \right),$$

with $\eta_\varepsilon := (\lceil A \rceil m^2 \lceil D \rceil^m)^{-1} \varepsilon$. This yields

$$\Phi_{p_m,D,\varepsilon}(x) = \sum_{i=0}^m a_i \Psi_{D,\eta_\varepsilon}^i(x), \quad \text{for all } x \in \mathbb{R}.$$

Hence (21) implies

$$\left\|\Phi_{p_m,D,\varepsilon}(x)-p_m\right\|_{L^\infty([-D,D])} \leq \sum_{i=0}^{m}|a_i|\|\Psi^i_{D,\eta_\varepsilon}(x)-x^i\|_{L^\infty([-D,D])} \leq \sum_{i=2}^{m}|a_i|\Big(\eta_\varepsilon\sum_{s=0}^{i-2}\lceil D\rceil^s\Big)$$
$$\leq \eta_\varepsilon \max_{i\in\{2,\dots,m\}}|a_i|\sum_{i=2}^{m}(i-1)\lceil D\rceil^{i-2} \leq Am^2\lceil D\rceil^{m-2}\eta_\varepsilon \leq \varepsilon.$$

Thanks to its compositional structure, the width of $\Phi_{p_m,D,\varepsilon}$ equals the maximum width of the individual networks in the composition, i.e., $\mathcal{W}(\Phi_{p_m,D,\varepsilon})\leq 16$. Since $H^{i-1}_{D,\eta_\varepsilon}\leq 2\lceil D\rceil^{m-1}$, for $i\leq m$, we further have

$$\begin{aligned}\mathcal{L}(\Phi_{p_m,D,\varepsilon}) &\leq \sum_{i=2}^{m}\mathcal{L}(\varphi^i_{p_m,D,\eta_\varepsilon}) \leq \sum_{i=2}^{m} C_2\log(\lceil H^{i-1}_{D,\eta_\varepsilon}\rceil^2\eta_\varepsilon^{-1})\\ &\leq C_2\, m\,(\log(\lceil A\rceil)+\log(\varepsilon^{-1})+(3m-2)\log(\lceil D\rceil)+2\log(m)+2)\\ &\leq 4\,C_2 m\,(\log(\lceil A\rceil)+\log(\varepsilon^{-1})+m\log(\lceil D\rceil)+\log(m)).\end{aligned}$$

Finally, we note that

$$\mathcal{B}(\Phi_{p_m,D,\varepsilon}) = \max\{1,|a_0|,|a_m|,\max_{i\in\{2,3,\dots,m\}}\mathcal{B}(\varphi^i_{p_m,D,\eta_\varepsilon})\}\leq \max\{A,8\lceil D\rceil^{2m-2}\}.$$

This finalizes the proof. □

Next, we recall that the Weierstrass approximation theorem states that every continuous function on a closed interval can be approximated to within arbitrary accuracy by a polynomial.

Theorem III.4 ([22])**.** *Let $[a,b]\subseteq\mathbb{R}$ and $f\in C([a,b])$. Then, for every $\varepsilon>0$, there exists a polynomial π such that*

$$\|f-\pi\|_{L^\infty([a,b])}\leq\varepsilon.$$

Proposition III.3 hence allows to conclude that every continuous function on a closed interval can be approximated to within arbitrary accuracy by a deep ReLU network of width no more than 16. This amounts to a variant of the universal approximation theorem [9], [10] for finite-width deep ReLU networks. We note, however, that the Weierstrass approximation theorem is non-quantitative. A quantitative statement can be obtained for smooth functions defined as follows.

Definition III.5. *For $D\in\mathbb{R}_+$, let the set $\mathcal{S}_D\subseteq C^\infty([-D,D],\mathbb{R})$ be given by*

$$\mathcal{S}_D=\left\{f\in C^\infty([-D,D],\mathbb{R})\colon \|f^{(n)}(x)\|_{L^\infty([-D,D])}\leq n!,\ \text{ for all } n\in\mathbb{N}_0\right\}. \tag{23}$$

Lemma III.6. *There exist a constant $C>0$ and a polynomial π such that for all $D\in\mathbb{R}_+$, $f\in\mathcal{S}_D$, and $\varepsilon\in(0,1/2)$, there is a network $\Psi_{f,\varepsilon}\in\mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Psi_{f,\varepsilon})\leq C\lceil D\rceil(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Psi_{f,\varepsilon})\leq 23$, $\mathcal{B}(\Psi_{f,\varepsilon})\leq\max\{1/D,\lceil D\rceil\}\pi(\varepsilon^{-1})$, and*

$$\|\Psi_{f,\varepsilon}-f\|_{L^\infty([-D,D])}\leq\varepsilon. \tag{24}$$

Proof. We first consider the case $D = 1$. A fundamental result on Chebyshev interpolation, see e.g. [23, Lemma 3], guarantees, for all $f \in \mathcal{S}_1$, $n \in \mathbb{N}$, the existence of a polynomial $P_{f,n}$ of degree n such that

$$\|f - P_{f,n}\|_{L^\infty([-1,1])} \leq \tfrac{1}{2^n(n+1)!}\|f^{(n+1)}\|_{L^\infty([-1,1])} \leq \tfrac{1}{2^n}. \tag{25}$$

Writing the polynomials $P_{f,n}$ as $P_{f,n} = \sum_{j=0}^n a_{f,n,j}x^j$, crude—but sufficient for our purposes—estimates show that there exists a constant $c > 0$ such that for all $f \in \mathcal{S}_1$, $n \in \mathbb{N}$ it holds that

$$A_{f,n} := \max_{j=0,\dots,n} |a_{f,n,j}| \leq 2^{cn}.$$

Application of Proposition III.3 to $P_{f,n}$ establishes the existence of a constant $C_1 > 0$ such that for all $f \in S_1$, $n \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{P_{f,n},1,\varepsilon/2} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{W}(\Phi_{P_{f,n},1,\varepsilon/2}) \leq 16$, $\mathcal{B}(\Phi_{P_{f,n},1,\varepsilon/2}) \leq \max\{A_{f,n}, 8\} \leq \max\{2^{cn}, 8\}$,

$$\mathcal{L}(\Phi_{P_{f,n},1,\varepsilon/2}) \leq C_1 n(cn + \log(2/\varepsilon) + \log(n)), \tag{26}$$

and

$$\|\Phi_{P_{f,n},1,\varepsilon/2} - P_{f,n}\|_{L^\infty([-1,1])} \leq \tfrac{\varepsilon}{2}. \tag{27}$$

In the following, we set $n_\varepsilon = \lceil \log(2/\varepsilon) \rceil$ and $\Psi_{f,\varepsilon} = \Phi_{P_{f,n_\varepsilon},1,\varepsilon/2}$. Combining (25) and (27) establishes that for all $f \in \mathcal{S}_1$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}\|\Psi_{f,\varepsilon} - f\|_{L^\infty([-1,1])} &\leq \|\Psi_{f,\varepsilon} - P_{f,n_\varepsilon}\|_{L^\infty([-1,1])} + \|P_{f,n_\varepsilon} - f\|_{L^\infty([-1,1])} \\ &\leq \tfrac{\varepsilon}{2} + \tfrac{1}{2^{n_\varepsilon}} \leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Using $\lceil \log(2/\varepsilon) \rceil \leq 2\log(2/\varepsilon)$ and $\log(2/\varepsilon) \leq 2\log(1/\varepsilon)$, for all $\varepsilon \in (0, 1/2)$, in (26) implies the existence of a constant C_2 such that for all $f \in \mathcal{S}_1$, $\varepsilon \in (0, 1/2)$,

$$\mathcal{L}(\Psi_{f,\varepsilon}) = \mathcal{L}(\Phi_{P_{f,n_\varepsilon},1,\varepsilon/2}) \leq C_2(\log(\varepsilon^{-1}))^2. \tag{28}$$

By the same token there exists a polynomial π_1 such that

$$\mathcal{B}(\Psi_{f,\varepsilon}) = \mathcal{B}(\Phi_{P_{f,n_\varepsilon},1,\varepsilon/2}) \leq \max\{2^{cn_\varepsilon}, 8\} \leq \pi_1(\varepsilon^{-1}).$$

This completes the proof for the case $D = 1$.

We next prove the statement for $D \in (0, 1)$. To this end, we start by noting that for $g \in \mathcal{S}_D$, with $D \in (0, 1)$, the function $f_g \colon [-1, 1] \to \mathbb{R}, x \mapsto g(Dx)$ is in $\mathcal{S}_1$. Hence, there exists, for every $g \in S_D$, $\varepsilon \in (0, 1/2)$, a network $\Psi_{f_g,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\sup_{x\in[-1,1]} |\Psi_{f_g,\varepsilon}(x) - f_g(x)| \leq \varepsilon$, $\mathcal{L}(\Psi_{f_g,\varepsilon}) \leq C_2(\log(1/\varepsilon))^2$, $\mathcal{W}(\Psi_{f_g,\varepsilon}) \leq 16$, and $\mathcal{B}(\Psi_{f_g,\varepsilon}) \leq \pi_1(\varepsilon^{-1})$. The claim is established by taking the network approximating $g(x)$ to be $\Psi'_{f_g,\varepsilon}(x) := \Psi_{f_g,\varepsilon}(\frac{x}{D})$ and noting that

$$\sup_{x\in[-D,D]}|\Psi'_{f_g,\varepsilon}(x)-g(x)| = \sup_{x\in[-D,D]}|\Psi_{f_g,\varepsilon}(\tfrac{x}{D})-f_g(\tfrac{x}{D})|$$
$$= \sup_{x\in[-1,1]}|\Psi_{f_g,\varepsilon}(x)-f_g(x)|\leq\varepsilon,$$

$\mathcal{L}(\Psi'_{f_g,\varepsilon})\leq C_2(\log(1/\varepsilon))^2$, $\mathcal{W}(\Psi'_{f_g,\varepsilon})\leq 16$, and $\mathcal{B}(\Psi_{f_g,\varepsilon})\leq(1/D)\pi_1(\varepsilon^{-1})$.

It remains to prove the statement for the case $D>1$. This will be accomplished by approximating f on intervals of length 2 (or less) and stitching the resulting approximations together using a localized partition of unity. To this end consider $a,b\in\mathbb{R}$ such that $1\leq b-a\leq 2$, and let $h\in C^\infty([a,b],\mathbb{R})$ with $\|h^{(n)}\|_{L^\infty([a,b])}\leq n!$, for all $n\in\mathbb{N}_0$. Next, note that the function $x\mapsto h\left(\frac{b-a}{2}x+\frac{b+a}{2}\right)$ is in $\mathcal{S}_1$. Hence, there exists, for every $\varepsilon\in(0,1/2)$, a network $\Psi'_{h,\varepsilon}\in\mathcal{NN}_{\infty,\infty,1,1}$ such that $\sup_{x\in[-1,1]}|\Psi'_{h,\varepsilon}(x)-h\left(\frac{b-a}{2}x+\frac{b+a}{2}\right)|\leq\varepsilon$, $\mathcal{L}(\Psi'_{h,\varepsilon})\leq C_2(\log(1/\varepsilon))^2$, $\mathcal{W}(\Psi'_{h,\varepsilon})\leq 16$, and $\mathcal{B}(\Psi'_{h,\varepsilon})\leq\pi_1(\varepsilon^{-1})$. The networks $\Psi_{h,\varepsilon}(x):=\Psi'_{h,\varepsilon}\left(\frac{2}{b-a}x-\frac{b+a}{b-a}\right)$ then satisfy

$$\sup_{x\in[a,b]}|\Psi_{h,\varepsilon}(x)-h(x)| = \sup_{y\in[-1,1]}|\Psi'_{h,\varepsilon}(y)-h\left(\tfrac{b-a}{2}y+\tfrac{b+a}{2}\right)|\leq\varepsilon, \tag{29}$$

$\mathcal{L}(\Psi_{h,\varepsilon})\leq C_2(\log(1/\varepsilon))^2$, $\mathcal{W}(\Psi_{h,\varepsilon})\leq 16$, and $\mathcal{B}(\Psi_{h,\varepsilon})\leq\max\{2,|b|+|a|\}\pi_1(\varepsilon^{-1})$. Now, for $D>1$, let $N_D\in\mathbb{N}$ be such that $1\leq\frac{2D}{N_D}\leq 2$ and consider the intervals

$$I_{D,k}:=\left[\tfrac{(k-1)D}{N_D},\tfrac{(k+1)D}{N_D}\right],\ k\in\{-N_D,\dots,N_D\}.$$

By (29) it follows that, for all $D>1$, $f\in\mathcal{S}_D$, $k\in\{-N_D,\dots,N_D\}$, and $\varepsilon\in(0,1/2)$, there exists a network $\Psi_{f,k,\varepsilon}\in\mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\sup_{x\in I_{D,k}}|\Psi_{f,k,\varepsilon}(x)-f(x)|\leq\tfrac{\varepsilon}{4}, \tag{30}$$

$\mathcal{L}(\Psi_{f,k,\varepsilon})\leq C_2(\log(4/\varepsilon))^2$, $\mathcal{W}(\Psi_{f,k,\varepsilon})\leq 16$, and $\mathcal{B}(\Psi_{f,k,\varepsilon})\leq\max\{2,2|k|\}\pi_1(\varepsilon^{-1})$. We next build a partition of unity through ReLU networks. Specifically, let $\chi(x)=\rho(x+1)-2\rho(x)+\rho(x-1)$, set $\chi_{D,k}(x)=\chi(\frac{N_D}{D}x-k)$, $D>1$, $k\in\mathbb{Z}$, and note that $\chi_{D,k}\in\mathcal{NN}_{2,8,1,1}$. This yields a partition of unity according to

$$\sum_{k\in\mathbb{Z}}\chi_{D,k}(x)=1,\quad\text{for all}\,x\in\mathbb{R}. \tag{31}$$

For $D>1$, $f\in\mathcal{S}_D$, $\varepsilon\in(0,1/2)$, let $f_\varepsilon\colon\mathbb{R}\to\mathbb{R}$ be given by

$$f_\varepsilon(x):=\sum_{k=-N_D}^{N_D}\Phi_{2,\varepsilon/4}(\chi_{D,k}(x),\Psi_{f,k,\varepsilon}(x)), \tag{32}$$

where $\Phi_{2,\varepsilon/4}$ is the multiplication network from Proposition III.2. Note that $|f(x)|\leq 1$, for all $x\in[-D,D]$, and $|\chi_{D,k}(x)|\leq 1$, for all $x\in[-D,D]$, $k\in\{-N_D,\dots,N_D\}$. Observe further that, for each $x\in[-D,D]$, there are no more than 2 indices k such that $\chi_{D,k}(x)\neq 0$. Proposition III.2 therefore implies that the sum in (32) has no more than 2 non-zero terms for each $x\in[-D,D]$. Combining (30), (31), and Proposition III.2, and noting that $\operatorname{supp}(\chi_{D,k})=I_{D,k}$, hence yields

$$\|f_\varepsilon-f\|_{L^\infty([-D,D])}\leq\varepsilon,$$

for all $D > 1$, $f \in \mathcal{S}_D$, $\varepsilon \in (0, 1/2)$. It remains to show that the functions f_ε can be realized by networks with the desired properties. To this end, consider for every $D > 1$, $f \in \mathcal{S}_D$, $k \in \{1, \dots, 2N_D + 1\}$, $\varepsilon \in (0, 1/2)$, the network $\alpha_{f,k,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ given by

$$\alpha_{f,k,\varepsilon}(x) := \Phi_{2,\varepsilon/4}(\chi_{D,k-(N_D+1)}(x), \Psi_{f,k-(N_D+1),\varepsilon}(x)),$$

and the network $\beta_{f,k,\varepsilon} \in \mathcal{NN}_{\infty,\infty,3,3}$ according to

$$\beta_{f,k,\varepsilon}(x_1, x_2, x_3) := \begin{pmatrix} x_1 \\ \alpha_{f,k,\varepsilon}(x_2) \\ x_3 \end{pmatrix}.$$

Further, set $\beta_0(x) := (x, 0, 0)^T$ and let $A \in \mathbb{R}^{3\times 3}$ be such that $A(y_1, y_2, y_3)^T = (y_1, y_1, y_2 + y_3)^T$, for all $y_1, y_2, y_3 \in \mathbb{R}$. We can now define, for every $D > 1$, $f \in \mathcal{S}_D$, $\varepsilon \in (0, 1/2)$, the network $\Psi_{f,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ given by

$$\Psi_{f,\varepsilon}(x) := (0 \;\; 1 \;\; 1)\, \beta_{f,2N_D+1,\varepsilon}(A\beta_{f,2N_D,\varepsilon}(\dots(A\beta_{f,1,\varepsilon}(A\beta_0(x))))).$$

Direct calculation shows that $f_\varepsilon(x) = \Psi_{f,\varepsilon}(x)$, for all $D > 1$, $f \in \mathcal{S}_D$, $\varepsilon \in (0, 1/2)$, $x \in \mathbb{R}$. Furthermore, thanks to Proposition III.2, there exists a constant $C_3 > 0$ such that, for all $D > 1$, $f \in \mathcal{S}_D$, $\varepsilon \in (0, 1/2)$,

$$\mathcal{W}(\Psi_{f,\varepsilon}) \leq 4 + \max_{k\in\{1,\dots,2N_D+1\}} \mathcal{W}(\alpha_{f,k,\varepsilon}) \leq 23,$$

$$\begin{aligned}\mathcal{L}(\Psi_{f,\varepsilon}) = 2 + \sum_{k=1}^{2N_D+1} \mathcal{L}(\beta_{f,k,\varepsilon}) &= 2 + \sum_{k=1}^{2N_D+1} \big(\mathcal{L}(\Phi_{2,\varepsilon/4}) + \max\{\mathcal{L}(\chi_{k-(N_D+1)}), \mathcal{L}(\Psi_{f,k-(N_D+1),\varepsilon})\}\big) \\ &\leq 2 + (2N_D + 1)(C_1 \log(16\varepsilon^{-1}) + \max\{2, C_2(\log(4\varepsilon^{-1}))^2\}) \leq C_3 D(\log(\varepsilon^{-1}))^2,\end{aligned}$$

and for all $f \in \mathcal{S}_D$, $D \in \mathbb{R}_+$, there exists a polynomial π_2 so that

$$\mathcal{B}(\Psi_{f,\varepsilon}) = \max_{k\in\{1,\dots,2N_D+1\}} \mathcal{B}(\alpha_{f,k,\varepsilon}) \leq \max\{8, 2D, 4D\pi_1(\varepsilon^{-1})\} \leq \lceil D \rceil \pi_2(\varepsilon^{-1}). \tag{33}$$

This concludes the proof. □

Remark III.7. *Lemma III.6 was formulated for symmetric intervals $[-D, D]$ for the sake of simplicity. The extension to functions $f \in C^\infty([a, b], \mathbb{R})$ with $\|f^{(n)}\|_{L^\infty([a,b])} \leq n!$, for all $n \in \mathbb{N}_0$, supported on arbitrary intervals $[a, b]$ is obtained by symmetrizing the support of f according to $g(x) = f(x + \frac{b+a}{2})$ and then applying Lemma III.6 to $g(x)$ with $D = \frac{b-a}{2}$. Note that this shift adds a weight of magnitude $|\frac{b+a}{2}|$, the bounds on $\mathcal{L}$ and $\mathcal{W}$ remain unaffected.*

Remark III.8. *The weights of the (finite-width) networks $\Psi_{f,\varepsilon}$ in Lemma III.6 depending on ε may be undesirable in practice. Proposition A.1 allows, however, to convert the $\Psi_{f,\varepsilon}$ into (finite-width) networks with depth still scaling poly-logarithmically in $1/\varepsilon$, weights bounded (in absolute value) by 2, and realizing the exact same function. We*

conclude by noting that the conversion result Proposition A.1 is interesting in its own right as often bounded weights at the expense of network size are desirable [24].

Remark III.9. *The results in this section all have approximating networks of finite width and depth scaling polylogarithmically in $1/\varepsilon$. Owing to*

$$\mathcal{M}(\Phi) \leq \mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1)$$

this implies that the connectivity scales no faster than polylogarithmic in $1/\varepsilon$. It therefore follows that the approximation error ε decays (at least) exponentially fast in the connectivity or equivalently in the number of parameters the approximant (i.e., the neural network) employs. We say that the network provides exponential approximation accuracy.

IV. Approximation of sinusoidal functions

We are now ready to proceed to the approximation of sinusoidal functions.

Theorem IV.1. *There exists a constant $C > 0$ such that for every $a, D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{a,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Psi_{a,D,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil aD \rceil))$, $\mathcal{W}(\Psi_{a,D,\varepsilon}) \leq 16$, $\mathcal{B}(\Psi_{a,D,\varepsilon}) \leq C$, and*

$$\|\Psi_{a,D,\varepsilon} - \cos(a\,\cdot)\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Proof. We start by approximating $x \mapsto \cos(2\pi x)$ on $[0,1]$. To this end note the MacLaurin series representation

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}, \quad \forall x \in \mathbb{R}.$$

Thanks to the Taylor theorem with remainder in Lagrange form, we have, for all $x \in [0,1]$,

$$\left|\cos(2\pi x) - \sum_{n=0}^{N} \frac{(-1)^n}{(2n)!}(2\pi x)^{2n}\right| \leq \left|\frac{(2\pi x)^{2N+1}}{(2N+1)!}\right| \sup_{t\in[0,1]} |\cos^{(2N+1)}(2\pi t)| \leq \frac{(2\pi)^{4N+2}}{(2N+1)!}. \tag{34}$$

Next observe that $n! \geq (\frac{n}{e})^n e$, for all $n \in \mathbb{N}$, which implies,

$$\frac{(2\pi)^{4N+2}}{(2N+1)!} \leq \frac{(4\pi^2)^{2N+1}}{(\frac{2N+1}{e})^{2N+1} e} \leq \Big(\frac{4\pi^2 e}{2N+1}\Big)^{2N+1}, \quad \text{for all } N \in \mathbb{N}. \tag{35}$$

With $N_\varepsilon := \lceil 2\pi^2 e \log(2/\varepsilon) \rceil$, we get, for all $\varepsilon \in (0, 1/2)$,

$$\Big(\frac{4\pi^2 e}{2N_\varepsilon+1}\Big)^{2N_\varepsilon+1} = \Big(\frac{4\pi^2 e}{2\lceil 2\pi^2 e \log(2/\varepsilon)\rceil + 1}\Big)^{2\lceil 2\pi^2 e \log(2/\varepsilon)\rceil+1} \leq 2^{-\lceil 2\pi^2 e \log(2/\varepsilon)\rceil} \leq 2^{-\log(2/\varepsilon)} = \frac{\varepsilon}{2}. \tag{36}$$

Noting that $C_1 := \left\lceil \max_{n\in\mathbb{N}_0} \left(\frac{(2\pi)^{2n}}{(2n)!}\right)\right\rceil < \infty$ and $N_\varepsilon \leq C_2 \log(\varepsilon^{-1})$, for all $\varepsilon \in (0,1/2)$, with $C_2 := 4\pi^2 e + 1$, application of Proposition III.3 to

$$p_m(x) = p_{N_\varepsilon}(x) := \sum_{n=0}^{N_\varepsilon} \frac{(-1)^n}{(2n)!}(2\pi x)^{2n},$$

with $D = 1$, establishes the following: There is a constant C_3 such that, for all $\varepsilon \in (0,1/2)$, there is a network $\Phi_{\varepsilon/2}$ satisfying

$$\left\|\Phi_{\varepsilon/2} - p_{N_\varepsilon}\right\|_{L^\infty([-1,1])} \leq \frac{\varepsilon}{2}, \tag{37}$$

with $\mathcal{W}(\Phi_{\varepsilon/2}) \leq 16$, $\mathcal{B}(\Phi_{\varepsilon/2}) \leq C_3$, and

$$\mathcal{L}(\Phi_{\varepsilon/2}) \leq C_3 N_\varepsilon(\log(C_1) + \log(2/\varepsilon) + N_\varepsilon \log(1) + \log(N_\varepsilon)) \leq C_4(\log(\varepsilon^{-1}))^2,$$

where $C_4 := C_2C_3(3 + \log(C_1) + \log(C_2))$. Combining (34), (35), (36), and (37), it follows that the network $\Phi_{\varepsilon/2}$ approximates the function $x \mapsto \cos(2\pi x)$ on $[0,1]$ to within accuracy ε, i.e., for all $\varepsilon \in (0, 1/2)$, we have

$$\|\Phi_{\varepsilon/2} - \cos(2\pi\,\cdot)\|_{L^\infty([0,1])} \leq \varepsilon. \tag{38}$$

We next extend this result to the approximation of $x \mapsto \cos(ax)$ on the interval $[-1,1]$ for arbitrary $a \in \mathbb{R}_+$. This will be accomplished by exploiting that $x \mapsto \cos(2\pi x)$ is 1-periodic and even. First recall the "sawtooth" functions $g_s \colon [0,1] \to [0,1]$, $s \in \mathbb{N}$, as defined in (7). It is straightforward, albeit somewhat tedious, to see that, for all $s \in \mathbb{N}_0$, $x \in [0,1]$,

$$\cos(2\pi 2^s x) = \cos(2\pi g_s(x)).$$

Fig. 2 illustrates this relation. Similarly, it follows that $\cos(2\pi 2^s x) = \cos(2\pi g_s(|x|))$, for all $s \in \mathbb{N}_0$, $x \in [-1,1]$. Next, note that for every $a \in \mathbb{R}_+$, there exists a $C_a \in (1/2, 1]$ such that $a/(2\pi) = C_a 2^{\lceil \log(a) - \log(2\pi) \rceil}$; we thus have, for all $a \in \mathbb{R}_+$, $x \in [-1,1]$,

$$\cos(ax) = \cos(2\pi 2^{\lceil \log(a) - \log(2\pi) \rceil} C_a x) = \cos\left(2\pi g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a|x|)\right).$$

Since $g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a|x|) \in [0,1]$, for all $a \in \mathbb{R}_+$, $x \in [-1,1]$, it follows from (38) that

$$\begin{aligned} &\left\|\Phi_{\varepsilon/2}\Big(g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a|x|)\Big) - \cos\big(2\pi g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a|x|)\big)\right\|_{L^\infty([-1,1])} \\ &= \left\|\Phi_{\varepsilon/2}\Big(g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a|x|)\Big) - \cos(ax)\right\|_{L^\infty([-1,1])} \leq \varepsilon. \end{aligned} \tag{39}$$

Now recall that $x \mapsto |x| = \rho(x) + \rho(-x)$ can be implemented by a 2-layer network and consider the realization of $x \mapsto g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a x)$, $a \in \mathbb{R}_+$, as developed in the proof of Proposition III.1. Applying Lemma II.5 twice, then establishes, thanks to (39), the existence of a constant C_5 such that the network

$$\Psi_{a,\varepsilon} := \Phi_{\varepsilon/2}(g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a|x|))$$

approximates $x \mapsto \cos(ax)$ on $[-1,1]$ with accuracy ε, while satisfying $\mathcal{L}(\Psi_{a,\varepsilon}) \leq C_5((\log(1/\varepsilon))^2 + \log(\lceil a \rceil))$, $\mathcal{W}(\Psi_{a,\varepsilon}) \leq 16$, and $\mathcal{B}(\Psi_{a,\varepsilon}) \leq C_5$.

Finally, we consider the approximation of $x \mapsto \cos(ax)$ on intervals $[-D, D]$, for arbitrary $D \geq 1$. To this end, we define, for all $a \in \mathbb{R}_+$, $D \in [1, \infty)$, $\varepsilon \in (0, 1/2)$, the network $\Psi_{a,D,\varepsilon}(x) := \Psi_{aD,\varepsilon}(\frac{x}{D})$ and observe that

$$\sup_{x \in [-D,D]} |\Psi_{a,D,\varepsilon}(x) - \cos(ax)| = \sup_{y \in [-1,1]} |\Psi_{a,D,\varepsilon}(Dy) - \cos(aDy)| = \sup_{y \in [-1,1]} |\Psi_{aD,\varepsilon}(y) - \cos(aDy)| \leq \varepsilon.$$

This concludes the proof. □

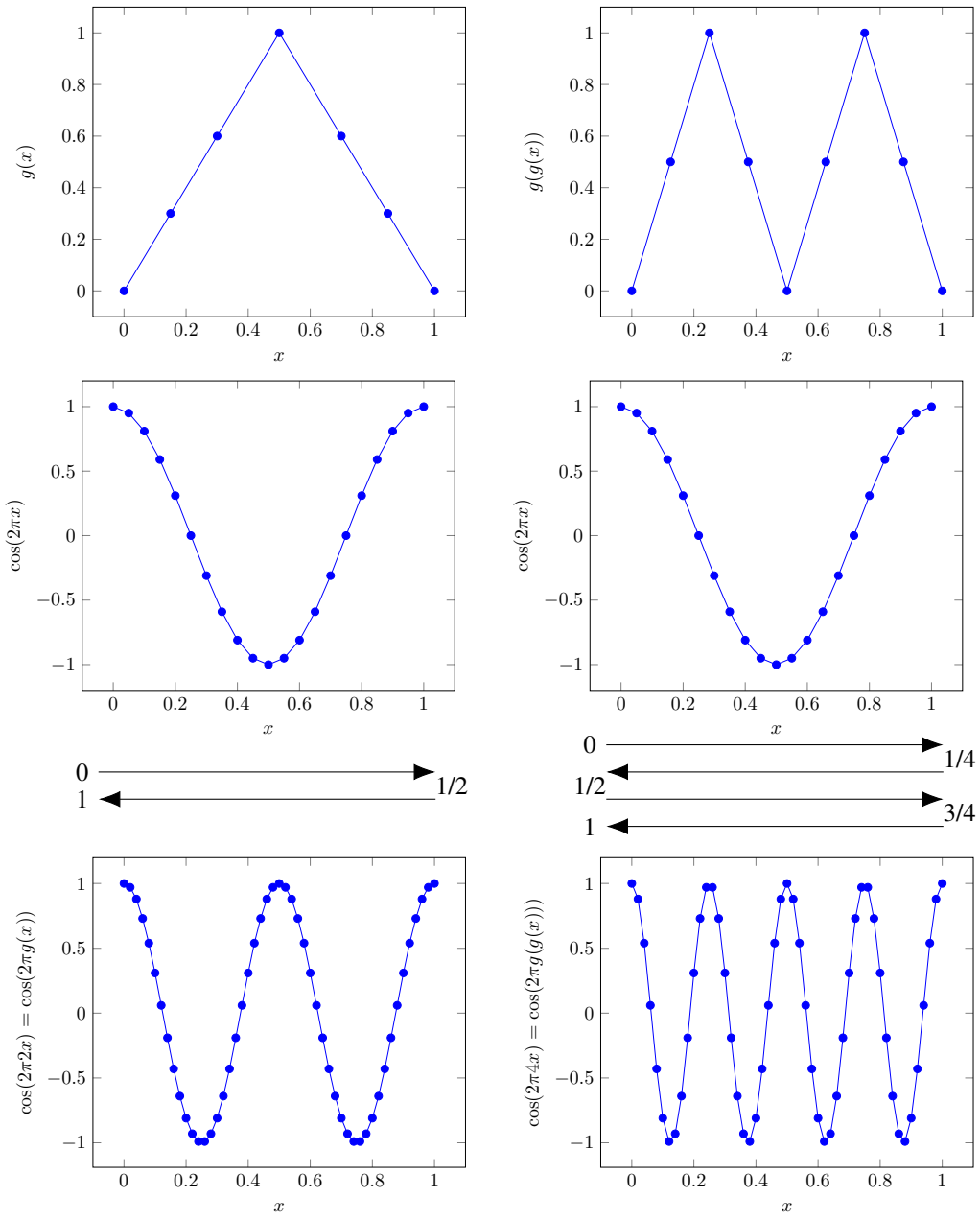


Fig. 2: Approximation of the function $\cos(2\pi ax)$ according to Theorem IV.1 using iterated "sawtooth" functions $g(\dots g(x))$, left $a = 2$, right $a = 4$.

The result just obtained extends to the approximation of $x \mapsto \sin(ax)$, formalized next, simply by noting that $\sin(x) = \cos(x - \pi/2)$.

Corollary IV.2. *There exists a constant $C > 0$ such that for every $a, D \in \mathbb{R}_+$, $b \in \mathbb{R}$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{a,b,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Psi_{a,b,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil aD + |b| \rceil))$, $\mathcal{W}(\Psi_{a,b,D,\varepsilon}) \leq 16$,*

$\mathcal{B}(\Psi_{a,b,D,\varepsilon}) \leq C$, *and*

$$\|\Psi_{a,b,D,\varepsilon} - \cos(a\,\cdot - b)\|_{L^\infty([-D,D])} \leq \varepsilon. \tag{40}$$

Proof. For given $a, D \in \mathbb{R}_+$, $b \in \mathbb{R}$, $\varepsilon \in (0, 1/2)$, consider the network $\Psi_{a,b,D,\varepsilon}(x) := \Psi_{a,D+\frac{|b|}{a},\varepsilon}\left(x - \frac{b}{a}\right)$ with $\Psi_{a,D,\varepsilon}$ as defined in the proof of Theorem IV.1, and observe that

$$\sup_{x\in[-D,D]} |\Psi_{a,b,D,\varepsilon}(x) - \cos(ax-b)| \leq \sup_{y\in\left[-(D+\frac{|b|}{a}),D+\frac{|b|}{a}\right]} |\Psi_{a,D+\frac{|b|}{a},\varepsilon}(y) - \cos(ay)| \leq \varepsilon.$$

□

We conclude by noting that both Theorem IV.1 and Corollary IV.2 provide approximation with exponential accuracy.

V. Quantifying Approximation Quality

We now proceed to developing a framework that allows to formally evaluate the approximation quality achievable by deep neural networks. The best-known results on approximation by neural networks are the universal approximation theorems of Hornik [10] and Cybenko [9], stating that continuous functions on bounded domains can be approximated arbitrarily well by a single-hidden-layer ($L = 2$ in our terminology) neural network with sigmoidal activation function. The literature on approximation-theoretic properties of networks with a single hidden layer continuing this line of work is abundant. Without any claim to completeness, we mention work on approximation error bounds in terms of the number of neurons for functions with bounded first moments [11], [25], the non-existence of localized approximations [26], a fundamental lower bound on approximation rates [27], [28], and the approximation of smooth or analytic functions [29], [30].

Approximation-theoretic results for networks with multiple hidden layers were obtained in [31], [32] for general functions, in [33] for continuous functions, and for functions together with their derivatives in [34]. In [26] it was shown that for certain approximation tasks deep networks can perform fundamentally better than single-hidden-layer networks. We also highlight two recent papers, which investigate the benefit—from an approximation-theoretic perspective—of multiple hidden layers. Specifically, in [35] it was shown that there exists a function which, although expressible through a small three-layer network, can only be represented through a very large two-layer network; here size is measured in terms of the total number of neurons in the network.

In the setting of deep convolutional neural networks first results of a nature similar to those in [35] were reported in [36]. Linking the expressivity properties of neural networks to tensor decompositions, [37], [38] established the existence of functions that can be realized by relatively small deep convolutional networks but require exponentially larger shallow convolutional networks.

We conclude by mentioning recent results bearing witness to the approximation power of deep ReLU networks in the context of PDEs. Specifically, it was shown in [21] that deep ReLU networks can approximate certain solution

families of parametric PDEs depending on a large (possibly infinite) number of parameters while overcoming the curse of dimensionality. The series of papers [39], [40], [41], [42] constructs and analyzes a deep-learning-based numerical solver for Black-Scholes PDEs that for the first time breaks the curse of dimensionality.

For survey articles on approximation-theoretic aspects of neural networks, we refer the interested reader to [43], [44]. Most closely related to the framework we develop here is the recent paper by Shaham, Cloninger, and Coifman [45], which shows that for functions that are sparse in specific wavelet frames, the best M-weight approximation rate (see Definition V.6 below) of three-layer neural networks is at least as high as the best M-term approximation rate in piecewise linear wavelet frames.

We begin the development of our framework with a review of a widely used theoretical foundation for deterministic lossy data compression [46], [47]. Our presentation essentially follows [48], [49].

A. Min-Max (Kolmogorov) Rate Distortion Theory

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and consider the function class $\mathcal{C} \subset L^2(\Omega)$. Then, for each $\ell \in \mathbb{N}$, we denote by

$$\mathfrak{E}^\ell := \left\{E : \mathcal{C} \to \{0,1\}^\ell\right\}$$

the set of *binary encoders of $\mathcal{C}$ of length ℓ*, and we let

$$\mathfrak{D}^\ell := \left\{D : \{0,1\}^\ell \to L^2(\Omega)\right\}$$

be the set of *binary decoders of length ℓ*. An encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ is said to *achieve uniform error ε over the function class $\mathcal{C}$*, if

$$\sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon.$$

A quantity of central interest is the minimal length $\ell \in \mathbb{N}$ for which there exists an encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ that achieves uniform error ε over the function class $\mathcal{C}$, along with its asymptotic behavior as made precise in the following definition.

Definition V.1. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and $\mathcal{C} \subset L^2(\Omega)$. Then, for $\varepsilon > 0$, the* minimax code length $L(\varepsilon, \mathcal{C})$ *is*

$$L(\varepsilon, \mathcal{C}) := \min\left\{\ell \in \mathbb{N} : \exists (E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell : \sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon\right\}.$$

Moreover, the optimal exponent $\gamma^*(\mathcal{C})$ *is defined as*

$$\gamma^*(\mathcal{C}) := \sup\left\{\gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\left(\varepsilon^{-1/\gamma}\right),\ \varepsilon \to 0\right\}.$$

The optimal exponent $\gamma^*(\mathcal{C})$ determines the minimum growth rate of $L(\varepsilon, \mathcal{C})$ as the error ε tends to zero and can hence be seen as quantifying the "description complexity" of the function class $\mathcal{C}$. Larger $\gamma^*(\mathcal{C})$ results in smaller growth rate and hence smaller memory requirements for storing signals $f \in \mathcal{C}$ such that reconstruction with

uniformly bounded error is possible. The quantity $\gamma^*(\mathcal{C})$ is closely related to the concept of Kolmogorov entropy [50]. Remark 5.10 in [49] makes this connection explicit.

The optimal exponent is known for several function classes, such as subsets of Besov spaces $B^s_{p,q}(\mathbb{R}^d)$ with $1 \leq p, q < \infty, s > 0$, and $q > (s + 1/2)^{-1}$, namely all functions in $B^s_{p,q}(\mathbb{R}^d)$ of bounded norm, see e.g. [51]. Specifically, for $\mathcal{C}$ a bounded subset of $B^s_{p,q}(\mathbb{R}^d)$, $\gamma^*(\mathcal{C}) = s/d$. Further results are available for β-cartoon-like functions, which have $\gamma^*(\mathcal{C}) = \beta/2$ (see [52], [53]), and for modulation spaces M_p with $1 \leq p < 2$, where $\gamma^*(\mathcal{C}) = \frac{1}{-1/2+1/p}$ (see [13]).

B. Approximation with Representation Systems

Fix $\Omega \subset \mathbb{R}^d$. Let $\mathcal{C}$ be a compact set of functions in $L^2(\Omega)$, henceforth referred to as *function class*, and consider a corresponding system $\mathcal{D} := (\varphi_i)_{i\in I} \subset L^2(\Omega)$ with I countable, termed *representation system*. We study the *best M-term approximation error* of $f \in \mathcal{C}$ in $\mathcal{D}$ defined as follows.

Definition V.2. *[46] Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a representation system $\mathcal{D} = (\varphi_i)_{i\in I} \subset L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma^{\mathcal{D}}_M(f) := \inf_{\substack{I_M \subseteq I,\\ \#I_M = M, (c_i)_{i\in I_M}}} \left\| f - \sum_{i\in I_M} c_i \varphi_i \right\|_{L^2(\Omega)}. \tag{41}$$

We call $\Gamma^{\mathcal{D}}_M(f)$ the best M-term approximation error of f in $\mathcal{D}$. *Every $f_M = \sum_{i\in I_M} c_i\varphi_i$ attaining the infimum in* (41) *is referred to as a* best M-term approximation *of f in $\mathcal{D}$. The supremal $\gamma > 0$ such that*

$$\sup_{f\in\mathcal{C}} \Gamma^{\mathcal{D}}_M(f) \in \mathcal{O}(M^{-\gamma}),\ M \to \infty,$$

will be denoted by $\gamma^*(\mathcal{C}, \mathcal{D})$. We say that the best M-term approximation rate of $\mathcal{C}$ in the representation system $\mathcal{D}$ *is $\gamma^*(\mathcal{C}, \mathcal{D})$.*

Function classes $\mathcal{C}$ widely studied in the approximation theory literature include unit balls in Lebesgue, Sobolev, or Besov spaces [47], as well as α-cartoon-like functions [54]. A wealth of structured representation systems $\mathcal{D}$ is provided by the area of applied harmonic analysis, starting with wavelets [55], followed by ridgelets [28], curvelets [56], shearlets [57], parabolic molecules [58], and most generally α-molecules [54], which include all previously named systems as special cases. Further examples are Gabor frames [14], local cosine bases [13], and wave atoms [15].

The best M-term approximation rate $\gamma^*(\mathcal{C}, \mathcal{D})$ according to Definition V.2 quantifies how difficult it is to approximate a given function class $\mathcal{C}$ in a fixed representation system $\mathcal{D}$. It is sensible to ask whether for given $\mathcal{C}$, there is a fundamental limit on $\gamma^*(\mathcal{C}, \mathcal{D})$ when one is allowed to vary over $\mathcal{D}$. As shown in [48], [49], every dense (and countable) $\mathcal{D} \subset L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$, results in $\gamma^*(\mathcal{C}, \mathcal{D}) = \infty$ for all function classes $\mathcal{C} \subset L^2(\Omega)$. However, identifying the elements in $\mathcal{D}$ participating in the best M-term approximation is practically infeasible as it entails

searching through the infinite set $\mathcal{D}$ and requires, in general, an infinite number of bits to describe the indices of the participating elements. This insight leads to the concept of "best M-term approximation subject to polynomial-depth search" as introduced by Donoho in [48]. Here, the basic idea is to restrict i) the search for the elements in $\mathcal{D}$ participating in the best M-term approximation to the first $\pi(M)$ elements of $\mathcal{D}$, with π a polynomial, and ii) the coefficients c_i in the best M-term approximation $f_M = \sum_{i \in I_M} c_i \varphi_i$ to be uniformly bounded. We formalize this under the name of effective best M-term approximation as follows.

Definition V.3. *Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a representation system $\mathcal{D} = (\varphi_i)_{i \in I} \subset L^2(\Omega)$, the supremal $\gamma > 0$ so that there exist a polynomial π and a constant $D > 0$ such that*

$$\sup_{f \in \mathcal{C}} \inf_{\substack{I_M \subset \{1,2,\dots,\pi(M)\}, \\ \#I_M = M, (c_i)_{i \in I_M}, \max_{i \in I_M} |c_i| \leq D}} \left\| f - \sum_{i \in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}), \ M \to \infty, \tag{42}$$

will be denoted by $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D})$ and referred to as effective best M-term approximation rate of $\mathcal{C}$ in the representation system $\mathcal{D}$.

We next recall a result from [48], [49] which states that $\sup_{\mathcal{D} \subset L^2(\Omega)} \gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D})$ is, indeed, finite under quite general conditions on $\mathcal{C}$; more specifically, it is upper-bounded by $\gamma^*(\mathcal{C})$ and hence limited by the "description complexity" of $\mathcal{C}$. This endows $\gamma^*(\mathcal{C})$ with operational meaning.

Theorem V.4. *[48], [49] Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$. The effective best M-term approximation rate of the function class $\mathcal{C} \subset L^2(\Omega)$ in the representation system $\mathcal{D} \subset L^2(\Omega)$ satisfies*

$$\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}) \leq \gamma^*(\mathcal{C}).$$

In light of this result the following definition is natural (see also [49]).

Definition V.5. *Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$. If the effective best M-term approximation rate of the function class $\mathcal{C} \subset L^2(\Omega)$ in the representation system $\mathcal{D} \subset L^2(\Omega)$ satisfies*

$$\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C}),$$

we say that the function class $\mathcal{C}$ is optimally representable *by $\mathcal{D}$.*

We next outline how the polynomial depth search constraint and the restriction to bounded coefficients c_i lead to rate-distortion-optimal encoder-decoder pairs. The reader is referred to [49] for a rigorous analysis. We start by noting that, thanks to the polynomial depth search constraint, the indices of the elements of $\mathcal{D}$ participating in the best M-term representation of f can be represented by a total of $M \log \pi(M) = CM \log(M)$ bits for some constant C. The corresponding coefficients c_i are quantized by rounding to integer multiples of $\lceil M^{-\alpha} \rceil$ for some constant α. As the c_i are bounded by a universal constant D, this leads to $\mathcal{O}(M^\alpha)$ quantization levels and hence to a total of $C' M \log(M)$ bits, for some constant C', needed to store the quantized coefficients. In summary,

we have a representation of f by $\mathcal{O}(M\log(M))$ bits. An encoder-decoder pair allowing to reconstruct f from a bitstring of length $\mathcal{O}(M\log(M))$ with approximation error $\varepsilon \propto M^{-\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})}$ (see Definition V.3) is described in [49]. The basic idea is to encode f by concatenating the binary representations of the indices of the participating elements of $\mathcal{D}$ and of the corresponding quantized coefficients such that the decoder can uniquely read them out from the overall bitstring. The minimax code length corresponding to this encoder-decoder pair scales according to $\varepsilon^{-1/\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})}\log(\varepsilon^{-1/\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})}) \in \mathcal{O}(\varepsilon^{-1/(\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})-\delta)})$, $\varepsilon \to 0$, for every $\delta \in (0,\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}))$. By Definition V.1 this leads to $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) - \delta \leq \gamma^*(\mathcal{C})$, for every $\delta \in (0,\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}))$, which is Theorem V.4 as δ can be chosen arbitrarily small. In particular, the encoder-decoder pair is rate-distortion-optimal if $\mathcal{C}$ is optimally representable by $\mathcal{D}$, i.e., if $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C})$.

C. Approximation with Deep Neural Networks

Inspired by the theory of best M-term approximation with representation systems, we now systematically develop the new concept of best M-weight approximation through neural networks. At the heart of our philosophy lies the interpretation of the network weights as the counterpart of the coefficients c_i in best M-term approximation. In other words, parsimony in terms of the number of participating elements in a representation system is replaced by parsimony in terms of network connectivity. Our development will parallel that for best M-term approximation in the previous section. We start by introducing the concept of best M-weight approximation rate.

Definition V.6. *Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and a function class $\mathcal{C} \subset L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma_M^{\mathcal{NN}}(f) := \inf_{\Phi \in \mathcal{NN}_{\infty,M,d,1}} \|f - \Phi\|_{L^2(\Omega)}. \tag{43}$$

We call $\Gamma_M^{\mathcal{NN}}(f)$ the best M-weight approximation error of f. *The supremal $\gamma > 0$ such that*

$$\sup_{f\in\mathcal{C}} \Gamma_M^{\mathcal{NN}}(f) \in \mathcal{O}(M^{-\gamma}),\ M \to \infty,$$

will be denoted by $\gamma^*_{\mathcal{NN}}(\mathcal{C})$. We say that the best M-weight approximation rate of $\mathcal{C}$ by neural networks *is $\gamma^*_{\mathcal{NN}}(\mathcal{C})$.*

We emphasize that the infimum in (43) is taken over all networks with fixed input dimension d, no more than M nonzero (edge and node) weights, and arbitrary depth L. In particular, this means that the infimum is taken over all possible network topologies and weight choices. The best M-weight approximation rate is fundamental as it benchmarks all algorithms that map a function f and an $\varepsilon > 0$ to a neural network approximating f with error no more than ε.

The two restrictions underlying the concept of effective best M-term approximation through representation systems, namely polynomial depth search and bounded coefficients, are next addressed in the context of approximation through deep neural networks. We start by noting that the need for the former is obviated by the tree-like-structure of neural networks. Specifically, a network Φ of connectivity M can not have its width $\mathcal{W}(\Phi)$ grow faster than proportional to M. Similarly, the network depth $\mathcal{L}(\Phi)$ can not grow faster than proportional to M either. As the

total number of nonzero weights in the network can not exceed $\mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1)$, this yields at most $\mathcal{O}(M^3)$ possibilities for the "locations" (in terms of entries in the A_ℓ and the b_ℓ) of the M nonzero weights. In fact, there are only $\mathcal{O}(M^2)$ possibilities for the locations of the M nonzero weights. To see this, first note that $N_0, N_L \leq M$ and $\sum_{k=1}^{L-1} N_k \leq M$. The total number of weights in the network is hence upper-bounded according to

$$\sum_{k=1}^{L}(N_k N_{k-1} + N_k) \leq \sum_{k=1}^{L} N_k \sum_{k=1}^{L} N_{k-1} + \sum_{k=1}^{L} N_k \leq 4M^2 + 2M = \mathcal{O}(M^2). \tag{44}$$

Encoding the locations of the M non-zero weights hence requires $\log(\binom{M^2}{M}) = \mathcal{O}(M\log(M))$ bits. This assumes, however, that the topology of the network, i.e., the number of layers L and the N_k are known. Proposition V.12 below shows that the topology can also be encoded with $\mathcal{O}(M\log(M))$ bits. In summary, we can therefore conclude that the tree-like-structure of neural networks automatically guarantees what we had to enforce through the polynomial depth search constraint in the case of best M-term approximation. Inspection of the approximation results in Section III reveals that a sublinear growth restriction on $\mathcal{L}(\Phi)$ as a function of M is natural. Specifically, the approximation results in Section III all have $\mathcal{L}(\Phi)$ proportional to a polynomial in $\log(\varepsilon^{-1})$. As we are interested in approximation error decay according to $M^{-\gamma}$, see Definition V.6, this suggests to restrict $\mathcal{L}(\Phi)$ to growth that is polynomial in $\log(M)$. Such a growth behavior, referred to as polylogarithmic in M, will also turn out crucial for allowing rate-distortion-optimal quantization. More specifically, it will be required for the quantization result in Lemma V.13 to hold in a way that is compatible with the achievability result Proposition V.12. The second restriction made in the definition of effective best M-term approximation, namely bounded coefficients, will be replaced by a more generous growth condition on the network weights; specifically, we will allow the magnitude of the weights to grow polynomially in M. This growth condition will turn out natural in the context of the approximation results we are interested in and will be seen below to allow rate-distortion-optimal quantization of the network weights. We remark, however, that Proposition A.1 allows to convert networks with weights growing polynomially in M into networks with bounded weights at the expense of increased depth, albeit still with depth polylogarithmic in M.

In summary, we will develop the concept of "best M-weight approximation subject to polylogarithmic depth and polynomial weight growth".

We start by introducing notation for neural networks with bounded weights.

Definition V.7. *Let $L, M, d \in \mathbb{N}$ and $R \in \mathbb{R}_+$. Then, we define*

$$\mathcal{NN}^R_{L,M,d,1} := \{\Phi \in \mathcal{NN}_{L,M,d,1} \colon \mathcal{B}(\Phi) \leq R\}.$$

We are now ready to formalize the notion of effective best M-weight approximation rate subject to polylogarithmic depth and polynomial weight growth.

Definition V.8. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and $\mathcal{C} \subset L^2(\Omega)$ be a function class. The supremal $\gamma > 0$ so that there is a polynomial π such that*

$$\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}_{\pi(\log(M)),M,d,1}^{\pi(M)}} \|f-\Phi\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M\to\infty, \tag{45}$$

is referred to as effective best M-weight approximation rate of $\mathcal{C}$ by neural networks *and will be denoted by* $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C})$.

We next state the equivalent of Theorem V.4 for approximation by deep neural networks. Specifically, we establish that the optimal exponent $\gamma^*(\mathcal{C})$ constitutes a fundamental bound on the effective best M-weight approximation rate of $\mathcal{C}$ as well.

Theorem V.9. *Let $d\in\mathbb{N}$, $\Omega\subset\mathbb{R}^d$ be bounded, and $\mathcal{C}\subset L^2(\Omega)$. Then, we have*

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \le \gamma^*(\mathcal{C}).$$

The key ingredients of the proof of Theorem V.9 are developed throughout this section and the formal proof will be given at the end of the section. Before getting started, we note that, in analogy to Definition V.5, what we just found suggests the following.

Definition V.10. *For $d\in\mathbb{N}$ and $\Omega\subset\mathbb{R}^d$ bounded, we say that the function class $\mathcal{C}\subset L^2(\Omega)$ is* optimally representable by neural networks *if*

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) = \gamma^*(\mathcal{C}).$$

It is remarkable that the fundamental limits of effective best M-term approximation through representation systems and effective best M-edge approximation in neural networks are determined by the same quantity, although the approximants in the two cases are vastly different. We have linear combinations of elements of a representation system with the participating functions identified subject to a polynomial-depth search constraint in the former, and concatenations of affine functions followed by non-linearities under polynomial growth constraints on the coefficients of the affine functions as well as polylogarithmic growth constraints on the number of concatenations in the latter case.

We now commence the program developing the proof of Theorem V.9. As in the arguments in the proof sketch of Theorem V.4 provided at the end of Section V-B, the main idea is to compare the code length corresponding to the approximating network to the minimax code length of the function class $\mathcal{C}$ to be approximated. To this end, we will need to represent the approximating network's nonzero weights and its topology, i.e., L, the N_k and the nonzero weights' locations as a bitstring. As the weights are real numbers and hence require, in principle, an infinite number of bits for their binary representations, we will have to suitably quantize them. In particular, the resolution of the corresponding quantizer will have to increase with decreasing ε. To formalize this idea, we start by defining the quantization employed.

Definition V.11. *Let $m \in \mathbb{N}$ and $\varepsilon \in (0,\infty)$. The network Φ is said to have (m,ε)-quantized weights if all its weights are elements of $2^{-m\lceil \log(\varepsilon^{-1})\rceil}\mathbb{Z} \cap [-\varepsilon^{-m},\varepsilon^{-m}]$.*

A key ingredient of the proof of Theorem V.9 is the following result, which establishes a fundamental lower bound on the connectivity of networks with quantized weights achieving uniform error ε over a given function class $\mathcal{C}$.

Proposition V.12. *Let $d, d' \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, $\mathcal{C} \subset L^2(\Omega)$, and let π be a polynomial. Further, let*

$$\Psi : \left(0,\frac{1}{2}\right) \times \mathcal{C} \to \mathcal{NN}_{\infty,\infty,d,d'}$$

be a map such that for every $f \in \mathcal{C}$, $\varepsilon \in (0,1/2)$, the network $\Psi(\varepsilon,f)$ has $(\lceil \pi(\log(\varepsilon^{-1}))\rceil,\varepsilon)$-quantized weights and satisfies

$$\sup_{f\in\mathcal{C}} \|f - \Psi(\varepsilon,f)\|_{L^2(\Omega)} \leq \varepsilon. \tag{46}$$

Then,

$$\sup_{f\in\mathcal{C}} \mathcal{M}(\Psi(\varepsilon,f)) \notin \mathcal{O}\left(\varepsilon^{-1/\gamma}\right),\ \varepsilon\to 0, \quad \textit{for all } \gamma > \gamma^*(\mathcal{C}). \tag{47}$$

Proof. The proof is by contradiction. Let $\gamma > \gamma^*(\mathcal{C})$ and assume that $\sup_{f\in\mathcal{C}} \mathcal{M}(\Psi(\varepsilon,f)) \in \mathcal{O}(\varepsilon^{-1/\gamma}), \varepsilon\to 0$. The contradiction will be effected by constructing encoder-decoder pairs $(E_\varepsilon, D_\varepsilon) \in \mathfrak{E}^{\ell(\varepsilon)} \times \mathfrak{D}^{\ell(\varepsilon)}$ achieving uniform error ε over $\mathcal{C}$ with

$$\ell(\varepsilon) \leq C_0 \cdot \sup_{f\in\mathcal{C}} \left(\mathcal{M}(\Psi(\varepsilon,f))\log(\mathcal{M}(\Psi(\varepsilon,f)))+1\right)(\log(\varepsilon^{-1}))^q \tag{48}$$

$$\leq C_0\left(\varepsilon^{-1/\gamma}\log(\varepsilon^{-1/\gamma})+1\right)(\log(\varepsilon^{-1}))^q \tag{49}$$

$$\leq C_1\left(\varepsilon^{-1/\gamma}(\log(\varepsilon^{-1}))^{q+1} + (\log(\varepsilon^{-1}))^q\right) \in \mathcal{O}\left(\varepsilon^{-1/\nu}\right), \quad \text{for } \varepsilon\to 0, \tag{50}$$

where $C_0, C_1, q > 0$ are constants not depending on f, ε and $\gamma > \nu > \gamma^*(\mathcal{C})$.

We proceed to the construction of the encoder-decoder pairs $(E_\varepsilon, D_\varepsilon) \in \mathfrak{E}^{\ell(\varepsilon)} \times \mathfrak{D}^{\ell(\varepsilon)}$, which will be accomplished by encoding the network topology and the quantized weights in bitstrings of length $\ell(\varepsilon)$ satisfying (48) while guaranteeing unique reconstruction (of the network). Fix $f \in \mathcal{C}$ and $\varepsilon \in (0,1/2)$. For the sake of notational simplicity, we set $\Psi := \Psi(\varepsilon,f)$, $M := \mathcal{M}(\Psi)$, and $L := \mathcal{L}(\Psi)$. Without loss of generality, we assume throughout that M is a power of 2 and larger than 1. The case $M = 0$ will be dealt with in Step 1 below. For all M that are not powers of 2 and for $M = 1$, we make use of the fact that $\mathcal{NN}_{L,M,d,d'} \subset \mathcal{NN}_{L,M',d,d'}$, where M' is the smallest power of 2 larger than M, and we encode the network like an M'-edge network. Since $M < M' \leq 2M$, this affects $\ell(\varepsilon)$ by a multiplicative constant only.

Recall that the number of nodes in layers $1,\dots,L$ is denoted by $N_1,\dots,N_L$, where $d' = N_L$, and $d = N_0$ is the dimension of the input layer (see Definition II.1). We further denote the number of nodes in layer $\ell = 1,...,L-1$ associated with edges of nonzero weight in the following layer by $\widetilde{N}_\ell$. It follows that

$$d + d' + \sum_{\ell=1}^{L-1} \widetilde{N}_\ell \leq \widetilde{M}, \tag{51}$$

where we set $\widetilde{M} := M + d + d'$. All other nodes do not contribute to the mapping $\Psi(x)$ and can hence be ignored. Moreover, we can assume that

$$L \leq \widetilde{M} \tag{52}$$

as otherwise there would be at least one layer $\ell \geq 1$ such that $A_\ell = 0$. As a consequence, the reduced network

$$x \mapsto W_L \rho(W_{L-1} \dots W_{\ell+1} \rho(0 \cdot x + b_\ell)),$$

realizes the same function as the original network Ψ but has less than L layers. This reduction can be repeated inductively until the resulting reduced network satisfies (52).

The bitstring representing Ψ is constructed according to the following steps.

Step 1: If $M = 0$, we encode the network by a single 0. Upon defining $0 \log(0) = 0$, we then note that (48) holds trivially and we terminate the encoding procedure. Else, we encode the network connectivity, M, by starting the overall bitstring with M 1's followed by a single 0. The length of this bitstring is therefore bounded by $\widetilde{M}$.

Step 2: We continue by encoding the number of layers in the network. Thanks to (52) this requires no more than $\lceil \log(\widetilde{M}) \rceil$ bits. We thus reserve the next $\lceil \log(\widetilde{M}) \rceil$ bits for the binary representation of L.

Step 3: Next, we store the dimensions d and d' of the input and the output layer, respectively, and the numbers of nodes $\widetilde{N}_\ell, \ell = 1, \dots, L-1$, associated with edges of nonzero weight. As $d, d' \leq \widetilde{M}$ and $\widetilde{N}_\ell \leq \widetilde{M}$, for $\ell = 1, \dots, L-1$, we can encode (generously) d, d', and each $\widetilde{N}_\ell$ using $\lceil \log(\widetilde{M}) \rceil$ bits. For the sake of concreteness, we first encode d followed by d' and $\widetilde{N}_1, \dots, \widetilde{N}_{L-1}$. In total, Step 3 requires a bitstring of length

$$(L+1)\lceil \log(\widetilde{M}) \rceil \leq (\widetilde{M}+1)\lceil \log(\widetilde{M}) \rceil.$$

In combination with Steps 1 and 2 this yields an overall bitstring of length at most

$$\widetilde{M}\lceil \log(\widetilde{M}) \rceil + 2\lceil \log(\widetilde{M}) \rceil + \widetilde{M}. \tag{53}$$

Step 4: We encode the topology of the graph associated with Ψ and consider only nodes that contribute to the mapping $\Psi(x)$. To this end, we enumerate the nodes in Ψ by assigning a unique index i—increasing from left to right in every layer and ranging from 1 to $\widetilde{N} := d + d' + \sum_{\ell=1}^{L-1} \widetilde{N}_\ell$—to each of these nodes. By (51) each of these indices can be encoded by a bitstring of length $\lceil \log(\widetilde{M}) \rceil$. We denote the bitstring corresponding to index i by $b(i) \in \{0,1\}^{\lceil \log(\widetilde{M}) \rceil}$ and let $n(i)$ be the number of children of the node with index i, i.e., the number of nodes in the next layer connected to the node with index i via an edge (of nonzero weight). For each node $i = 1, \dots, \widetilde{N}$, we form a bitstring of length $n(i) \cdot \lceil \log(\widetilde{M}) \rceil$ by concatenating the bitstrings $b(j)$ for all j such that there is an edge between i and j. We follow this string with an all-zeros bitstring of length $\lceil \log(\widetilde{M}) \rceil$ to signal the transition to the node with index $i+1$. The enumeration is concluded with an all-zeros bitstring of length $\lceil \log(\widetilde{M}) \rceil$ signaling that the last node has been reached. Overall, this yields a bitstring of length

$$\sum_{i=1}^{\widetilde{N}}(n(i)+1)\cdot\lceil\log(\widetilde{M})\rceil<2\widetilde{M}\lceil\log(\widetilde{M})\rceil, \tag{54}$$

where we used $\sum_{i=1}^{\widetilde{N}} n(i)<\widetilde{M}$ and (51). Combining (53) and (54) it follows that we have encoded the overall topology of the network Ψ using at most

$$\widetilde{M}+3\widetilde{M}\lceil\log(\widetilde{M})\rceil+2\lceil\log(\widetilde{M})\rceil \tag{55}$$

bits.

Step 5: We encode the weights of Ψ. By assumption, Ψ has $(\lceil\pi(\log(\varepsilon^{-1}))\rceil,\varepsilon)$-quantized weights, which means that each weight of Ψ can be represented by no more than $B_\varepsilon:=2(\pi(\log(\varepsilon^{-1}))+2)\log(\varepsilon^{-1})$ bits. For each node $i=1,\dots,\widetilde{N}$, we reserve the first B_ε bits to encode its associated node weight and, for each of its children a bitstring of length B_ε to encode the weight corresponding to the edge between that child and its parent node. Concatenating the results in ascending order of child node indices, we get a bitstring of length $(n(i)+1)B_\varepsilon$ for node i, and an overall bitstring of length

$$\sum_{i=1}^{\widetilde{N}}(n(i)+1)B_\varepsilon\leq 2\widetilde{M}B_\varepsilon \tag{56}$$

representing the weights of the graph associated with the network Ψ. With (55) this shows that the overall number of bits needed to encode the network topology and weights is no more than

$$\widetilde{M}+3\widetilde{M}\lceil\log(\widetilde{M})\rceil+2\lceil\log(\widetilde{M})\rceil+2\widetilde{M}B_\varepsilon. \tag{57}$$

The network can be recovered by sequentially reading out M,L,d,d', the $\widetilde{N}_\ell$, the topology, and the quantized weights from the overall bitstring. It is not difficult to verify that the individual steps in the encoding procedure were crafted such that this yields unique recovery. As (57) can be upper-bounded by

$$C_0M\log(M)(\log\big(\varepsilon^{-1}\big))^q \tag{58}$$

for constants $C_0,q>0$ depending on d,d', and π only, we have constructed an encoder-decoder pair $(E_\varepsilon,D_\varepsilon)\in\mathfrak{E}^{\ell(\varepsilon)}\times\mathfrak{D}^{\ell(\varepsilon)}$ with $\ell(\varepsilon)$ satisfying (48). This concludes the proof. □

The result just established applies to networks that have each weight represented by a finite number of bits scaling according to $(\log(\varepsilon^{-1}))^q$, for some $q\in\mathbb{N}$, while guaranteeing that the underlying encoder-decoder pair achieves uniform error ε over $\mathcal{C}$. We next show that such a compatibility is, indeed, possible. Specifically, this requires a careful interplay between the network's depth and connectivity scaling, and its weight growth, all as a function of ε. This delicate balancing will be seen to be met by our assumptions.

Lemma V.13. *Let $B,L,d,d',k\in\mathbb{N}$, $\Omega\subseteq[-B,B]^d$, $\varepsilon\in(0,1/2)$, and $M\leq\varepsilon^{-k}$. Further, let $\Phi\in\mathcal{NN}_{L,M,d,d'}$ with $\mathcal{B}(\Phi)\leq\varepsilon^{-k}$ and let $m\in\mathbb{N}$ be such that*

$$m\geq 3kL+\log(\max\{1,B\}).$$

Then, there exists a network $\tilde{\Phi} \in \mathcal{NN}_{L,M,d,d'}$ with (m,ε)-quantized weights satisfying

$$\sup_{x\in\Omega} \|\Phi(x) - \tilde{\Phi}(x)\|_\infty \leq \varepsilon.$$

Proof. By Definition II.1 there exist integers $N_0, N_1, \ldots, N_L \in \mathbb{N}$ and affine maps

$$W_\ell \colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell},\, x \mapsto W_\ell(x) = A_\ell x + b_\ell, \quad \ell = 1, 2, \ldots, L$$

with $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, $b_\ell \in \mathbb{R}^{N_\ell}$ such that

$$\Phi(x) = \begin{cases} W_2(\rho\,(W_1(x))), & L = 2 \\ W_L(\rho\,(W_{L-1}(\rho\,(\ldots\rho\,(W_1(x)))))), & L \geq 3. \end{cases}$$

We now consider the partial networks $\Phi^\ell \colon \Omega \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L-1\}$, given by

$$\Phi^\ell(x) = \begin{cases} \rho\,(W_1(x)), & \ell = 1 \\ \rho\,(W_2(\rho\,(W_1(x)))), & \ell = 2 \\ \rho\,(W_\ell(\rho\,(W_{\ell-1}(\ldots\rho\,(W_1(x)))))), & \ell \geq 3 \end{cases}$$

and to simplify notation we write $\Phi^L := \Phi$. Furthermore, for $\ell \in \{1, 2, \ldots, L\}$, let $\tilde{\Phi}^\ell$ be the (partial) network obtained by replacing all the entries of the A_ℓ, b_ℓ by a corresponding closest element in $2^{-m\lceil \log_2(\varepsilon^{-1})\rceil}\,\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$. The resulting networks $\tilde{\Phi}^\ell$, $\ell \in \{1, 2, \ldots, L\}$, are hence defined by matrices $\tilde{A}_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and vectors $\tilde{b}_\ell \in \mathbb{R}^{N_\ell}$ satisfying, for all $\ell \in \{1, 2, \ldots, L\}$,

$$\begin{aligned} \max_{i,j} |A_{\ell,i,j} - \tilde{A}_{\ell,i,j}| &\leq \tfrac{1}{2} 2^{-m\lceil \log_2(\varepsilon^{-1})\rceil} \leq \tfrac{1}{2}\varepsilon^m, \\ \max_{i,j} |b_{\ell,i,j} - \tilde{b}_{\ell,i,j}| &\leq \tfrac{1}{2} 2^{-m\lceil \log_2(\varepsilon^{-1})\rceil} \leq \tfrac{1}{2}\varepsilon^m. \end{aligned} \tag{59}$$

The proof will be effected by upper-bounding the error building up across layers as a result of the quantization of the edge and node weights. To this end, we define, for $\ell \in \{1, 2, \ldots, L\}$, the error in the ℓth layer as

$$e_\ell := \sup_{x\in\Omega} \|\Phi^\ell(x) - \tilde{\Phi}^\ell(x)\|_\infty.$$

We further set $C_0 := \max\{1, B\}$ and $C_\ell := \max\{1, \sup_{x\in\Omega} \|\Phi^\ell(x)\|_\infty\}$. As each entry of the vector $\Phi^\ell(x) \in \mathbb{R}^{N_\ell}$ is a weighted sum of at most $N_{\ell-1}$ components of the vector $\Phi^{\ell-1}(x) \in \mathbb{R}^{N_{\ell-1}}$ and an affine component $b_{\ell,i}$ and $\mathcal{B}(\Phi) \leq \varepsilon^{-k}$, by assumption, we have for all $\ell \in \{1, 2, \ldots, L\}$,

$$C_\ell \leq N_{\ell-1}\varepsilon^{-k} C_{\ell-1} + \varepsilon^{-k} \leq (N_{\ell-1} + 1)\,\varepsilon^{-k} C_{\ell-1},$$

which implies, for all $\ell \in \{1, 2, \ldots, L\}$, that

$$C_\ell \leq C_0\,\varepsilon^{-k\ell} \prod_{i=0}^{\ell-1} (N_i + 1). \tag{60}$$

Next, note that each component $(\tilde{\Phi}^1(x))_i, i \in \{1, 2, \ldots, N_1\}$, of the vector $\tilde{\Phi}^1(x) \in \mathbb{R}^{N_1}$ can be written as

$$(\tilde{\Phi}^1(x))_i = \rho\left(\left(\sum_{j=1}^{N_0} \tilde{A}_{1,i,j} x_j\right) + \tilde{b}_{1,i}\right),$$

which, combined with (59) and the fact that ρ is 1-Lipschitz implies

$$e_1 \leq C_0 N_0 \tfrac{\varepsilon^m}{2} + \tfrac{\varepsilon^m}{2} \leq C_0(N_0+1)\tfrac{\varepsilon^m}{2}. \tag{61}$$

Similarly, we have for $\ell \in \{2, 3, \ldots, L\}$, $i \in \{1, 2, \ldots, N_\ell\}$,

$$(\tilde{\Phi}^\ell(x))_i = \rho\left(\left(\sum_{j=1}^{N_{\ell-1}} \tilde{A}_{\ell,i,j}(\tilde{\Phi}^{\ell-1}(x))_j\right) + \tilde{b}_{\ell,i}\right),$$

which implies

$$\begin{aligned} e_\ell &= \sup_{x\in\Omega} \|\Phi^\ell(x) - \tilde{\Phi}^\ell(x)\|_\infty = \sup_{x\in\Omega, i\in\{1,\ldots,N_l\}} |(\Phi^\ell(x))_i - (\tilde{\Phi}^\ell(x))_i| \\ &= \sup_{x\in\Omega, i\in\{1,\ldots,N_\ell\}} \left| \left[\left(\sum_{j=1}^{N_{\ell-1}} A_{\ell,i,j}(\Phi^{\ell-1}(x))_j\right) + b_{\ell,i}\right] - \left[\left(\sum_{j=1}^{N_{\ell-1}} \tilde{A}_{\ell,i,j}(\tilde{\Phi}^{\ell-1}(x))_j\right) + \tilde{b}_{\ell,i}\right] \right| \\ &\leq \sup_{x\in\Omega, i\in\{1,\ldots,N_\ell\}} \left[\left(\sum_{j=1}^{N_{\ell-1}} \left|A_{\ell,i,j}(\Phi^{\ell-1}(x))_j - \tilde{A}_{\ell,i,j}(\tilde{\Phi}^{\ell-1}(x))_j\right|\right) + \left|b_{\ell,i} - \tilde{b}_{\ell,i}\right|\right]. \end{aligned} \tag{62}$$

Since $|(\Phi^{\ell-1}(x))_j - (\tilde{\Phi}^{\ell-1}(x))_j| \leq e_{\ell-1}$ and $|(\Phi^{\ell-1}(x))_j| \leq C_{\ell-1}$, both for all $x \in \Omega$ and all $j \in \{1, \ldots, N_\ell\}$, by definition, and $|A_{\ell,i,j}| \leq \varepsilon^{-k}$ by assumption, upon invoking (59), we get

$$|A_{\ell,i,j}(\Phi^{\ell-1}(x))_j - \tilde{A}_{\ell,i,j}(\tilde{\Phi}^{\ell-1}(x))_j| \leq e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2} + e_{\ell-1}\tfrac{\varepsilon^m}{2}.$$

Since $\varepsilon \in (0, 1/2)$ it therefore follows from (62), that for all $\ell \in \{2, 3, \ldots, L\}$,

$$e_\ell \leq N_{\ell-1}(e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2} + e_{\ell-1}\tfrac{\varepsilon^m}{2}) + \tfrac{\varepsilon^m}{2} \leq (N_{\ell-1}+1)(2e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2}). \tag{63}$$

We now claim that, for all $\ell \in \{2, 3, \ldots, L\}$,

$$e_\ell \leq \tfrac{1}{2}(2^\ell - 1)C_0\varepsilon^{m-(\ell-1)k}\prod_{i=0}^{\ell-1}(N_i+1), \tag{64}$$

which we prove by induction. The base case $\ell = 1$ was already established in (61). For the induction step we assume that (64) holds for a given ℓ which, in combination with (60) and (63), implies

$$\begin{aligned} e_{\ell+1} &\leq \big(N_\ell+1\big)\big(2e_\ell\varepsilon^{-k} + C_\ell\tfrac{\varepsilon^m}{2}\big) \\ &\leq (N_\ell+1)\left((2^\ell-1)C_0\varepsilon^{m-(\ell-1)k}\varepsilon^{-k}\prod_{i=0}^{\ell-1}(N_i+1) + C_0\varepsilon^{-k\ell}\tfrac{\varepsilon^m}{2}\prod_{i=0}^{\ell-1}(N_i+1)\right) \\ &= \frac{1}{2}(2^{\ell+1}-1)C_0\varepsilon^{m-\ell k}\prod_{i=0}^{\ell}(N_i+1). \end{aligned}$$

This completes the induction argument and establishes (64). Using $2^{L-1} \leq \varepsilon^{-(L-1)}$, $\prod_{i=0}^{L-1}(N_i+1) \leq M^L \leq \varepsilon^{-kL}$, $C_0 \leq \varepsilon^{-\log(\max\{1,B\})}$, and $m \geq 3kL + \log(\max\{1,B\})$ by assumption, we get

$$\begin{aligned}\sup_{x\in\Omega} \|\Phi(x) - \tilde{\Phi}(x)\|_\infty = e_L &\leq \tfrac{1}{2}(2^L-1)C_0\varepsilon^{m-(L-1)k}\prod_{i=0}^{L-1}(N_i+1)\\ &\leq \varepsilon^{m-(L-1+kL-k+\log(\max\{1,B\})+kL)}\\ &\leq \varepsilon^{m-(3kL+\log(\max\{1,B\})-1)} \leq \varepsilon.\end{aligned}$$

This completes the proof. □

Proposition V.12 not only says that the connectivity growth rate of networks with quantized weights achieving uniform approximation error ε over a function class $\mathcal{C}$ must exceed $\mathcal{O}\big(\varepsilon^{-1/\gamma^*(\mathcal{C})}\big)$, $\varepsilon \to 0$, but its proof, by virtue of constructing an encoder-decoder pair that achieves this growth rate also provides an achievability result. We next establish a matching strong converse—for networks with polynomially bounded weights—in the sense of showing that for $\gamma > \gamma^*(\mathcal{C})$, the uniform approximation error remains bounded away from zero for infinitely many $M \in \mathbb{N}$.

Proposition V.14. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, π a polynomial, and $\mathcal{C} \subset L^2(\Omega)$. Then, for all $C > 0$ and $\gamma > \gamma^*(\mathcal{C})$, we have*

$$\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}^{\pi(M)}_{\pi(\log(M)),M,d,1}} \|f-\Phi\|_{L^2(\Omega)} \geq CM^{-\gamma}, \text{ for infinitely many } M \in \mathbb{N}. \tag{65}$$

Proof. Let $\gamma > \gamma^*(\mathcal{C})$. Assume, towards a contradiction, that (65) holds only for finitely many $M \in \mathbb{N}$. Then, there exists a constant C' such that the inequality in (65) holds for no $M \in \mathbb{N}$ and hence there is a constant C so that

$$\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}^{\pi(M)}_{\pi(\log(M)),M,d,1}} \|f-\Phi\|_{L^2(\Omega)} \leq CM^{-\gamma}, \quad \text{for all } M \in \mathbb{N}.$$

Setting $M_\varepsilon := \lceil(\varepsilon/(4C))^{-1/\gamma}\rceil$, it follows that, for every $f \in \mathcal{C}$ and every $\varepsilon \in (0,1/2)$, there exists a neural network $\Phi_{\varepsilon,f} \in \mathcal{NN}^{\pi(M_\varepsilon)}_{\pi(\log(M_\varepsilon)),M_\varepsilon,d,1}$ such that

$$\|f-\Phi_{\varepsilon,f}\|_{L^2(\Omega)} \leq 2\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}^{\pi(M_\varepsilon)}_{\pi(\log(M_\varepsilon)),M_\varepsilon,d,1}} \|f-\Phi\|_{L^2(\Omega)} \leq 2CM_\varepsilon^{-\gamma} \leq \frac{\varepsilon}{2}.$$

Next, by Lemma V.13 there exists a polynomial π^* such that for every $f \in \mathcal{C}$, $\varepsilon \in (0,1/2)$, there is a network $\widetilde{\Phi}_{\varepsilon,f}$ with $(\lceil\pi^*(\log(\varepsilon^{-1}))\rceil,\varepsilon)$-quantized weights satisfying

$$\left\|\Phi_{\varepsilon,f} - \widetilde{\Phi}_{\varepsilon,f}\right\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2}.$$

Defining

$$\Psi : \left(0,\frac{1}{2}\right)\times\mathcal{C} \to \mathcal{NN}_{\infty,\infty,d,1}, \quad (\varepsilon,f) \mapsto \widetilde{\Phi}_{\varepsilon,f},$$

it follows that

$$\sup_{f\in\mathcal{C}} \|f-\Psi(\varepsilon,f)\|_{L^2(\Omega)} \leq \varepsilon \quad \text{with} \quad \mathcal{M}(\Psi(\varepsilon,f)) \leq M_\varepsilon \in \mathcal{O}(\varepsilon^{-1/\gamma}),\ \varepsilon \to 0.$$

The proof is concluded by noting that $\Psi(\varepsilon, f)$ violates Proposition V.12. □

We are now ready to proceed to the proof of Theorem V.9.

Proof of Theorem V.9. Suppose towards a contradiction that $\gamma^{*,\text{eff}}_{\mathcal{NN}}(\mathcal{C}) > \gamma^*(\mathcal{C})$. Let $\gamma \in \left(\gamma^*(\mathcal{C}), \gamma^{*,\text{eff}}_{\mathcal{NN}}(\mathcal{C})\right)$. Then, Definition V.8 implies the existence of a polynomial π and a constant $C > 0$ such that

$$\sup_{f\in\mathcal{C}} \inf_{\Phi_M \in \mathcal{NN}^{\pi(M)}_{\pi(\log(M)),M,d,1}} \|f - \Phi_M\|_{L^2(\Omega)} \leq CM^{-\gamma}, \text{ for all } M \in \mathbb{N}.$$

This constitutes a contradiction to Proposition V.14. □

We conclude this section with a discussion of the conceptual implications of the results established above. Proposition V.12 combined with Lemma V.13 establishes that neural networks achieving uniform approximation error ε while having weights that are polynomially bounded in ε^{-1} and depth growing polylogarithmically in ε^{-1} cannot exhibit connectivity growth rate smaller than $\mathcal{O}(\varepsilon^{-1/\gamma^*(\mathcal{C})}), \varepsilon \to 0$; in other words, a decay of the uniform approximation error, as a function of M, faster than $\mathcal{O}(M^{-\gamma^*(\mathcal{C})}), M \to \infty$, is not possible.

VI. Transitioning from Representation Systems to Neural Networks

We next develop a general framework for transferring results on function approximation through representation systems to results on approximation by neural networks. In particular, we prove that for a given function class $\mathcal{C}$ and an associated representation system $\mathcal{D}$ satisfying certain conditions, there exists a neural network with connectivity $\mathcal{O}(M)$ that achieves (up to a multiplicative constant) the same uniform error over $\mathcal{C}$ as a best M-term approximation in $\mathcal{D}$. This will lead to a characterization of function classes $\mathcal{C}$ that are optimally representable by neural networks in the sense of Definition V.10.

We start by defining the effective representability of representation systems through neural networks.

Definition VI.1. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ be a representation system. Then, $\mathcal{D}$ is said to be* effectively representable by neural networks*, if there exists a bivariate polynomial π such that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$ there is a neural network $\Phi_{i,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying $\mathcal{M}(\Phi_{i,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(i))$, $\mathcal{B}(\Phi_{i,\varepsilon}) \leq \pi(\varepsilon^{-1}, i)$, and*

$$\|\varphi_i - \Phi_{i,\varepsilon}\|_{L^2(\Omega)} \leq \varepsilon.$$

The next result will allow us to conclude that optimality—in the sense of Definition V.5—of a representation system $\mathcal{D}$ for a signal class $\mathcal{C}$ combined with effective representability of $\mathcal{D}$ by neural networks implies optimal representability of $\mathcal{C}$ by neural networks.

Theorem VI.2. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and consider the function class $\mathcal{C} \subset L^2(\Omega)$. Suppose that the representation system $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ is effectively representable by neural networks. Then, for all $0 < \gamma < \gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$, there exist a polynomial π and a map*

$$\Psi : \left(0, \frac{1}{2}\right) \times \mathcal{C} \to \mathcal{NN}_{\infty,\infty,d,1},$$

such that for every $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$ *the network* $\Psi(\varepsilon, f)$ *has* $(\lceil \pi(\log(\varepsilon^{-1}))\rceil, \varepsilon)$*-quantized weights while satisfying* $\|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon$, $\mathcal{L}(\Psi(\varepsilon, f)) \leq \pi(\log(\varepsilon^{-1}))$, $\mathcal{B}(\Psi(\varepsilon, f)) \leq \pi(\varepsilon^{-1})$, *and*

$$\mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma}), \varepsilon \to 0, \tag{66}$$

where the implicit constant in (66) is independent of f. *In particular, it holds that*

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}).$$

Remark VI.3. *Theorem VI.2 allows to draw the following conclusion. If* $\mathcal{D}$ *optimally represents the function class* $\mathcal{C}$ *in the sense of Definition V.5, i.e.,* $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C})$, *and if it is, in addition, effectively representable by neural networks in the sense of Definition VI.1, then, thanks to Theorem V.9, which says that* $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \leq \gamma^*(\mathcal{C})$, *we have* $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) = \gamma^*(\mathcal{C})$ *and hence* $\mathcal{C}$ *is optimally representable by neural networks in the sense of Definition V.10.*

Proof of Theorem VI.2. Let $\gamma' \in (\gamma, \gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D}))$. According to Definition V.3, there exist constants $C, D \geq 1$ and a polynomial π_1 such that for every $f \in \mathcal{C}$, $M \in \mathbb{N}$ there is a subset $I_{f,M} \subset \{1, \dots, \pi_1(M)\}$ of cardinality M and coefficients $(c_{f,i})_{i \in I_{f,M}}$ with $\max_{i \in I_{f,M}} |c_{f,i}| \leq D$ such that

$$\left\| f - \sum_{i \in I_{f,M}} c_{f,i} \varphi_i \right\|_{L^2(\Omega)} \leq \frac{C M^{-\gamma'}}{2}. \tag{67}$$

Let $A := \max\{1, |\Omega|^{1/2}\}$ and note that we can assume w.l.o.g. that $D \geq C$, which, in turn, implies $\frac{C}{4D} M^{-(\gamma'+1)} \leq 1/2$. Effective representability of $\mathcal{D}$ according to Definition VI.1 therefore ensures the existence of a bivariate polynomial π_2 such that for all $M \in \mathbb{N}$, $i \in I_{f,M}$, there is a neural network $\Phi_{i,M} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|\varphi_i - \Phi_{i,M}\|_{L^2(\Omega)} \leq \tfrac{C}{4DA} M^{-(\gamma'+1)} \tag{68}$$

with

$$\mathcal{M}(\Phi_{i,M}) \leq \pi_2\left(\log\left(\left(\tfrac{C}{4DA} M^{-(\gamma'+1)}\right)^{-1}\right), \log(i)\right) = \pi_2\left((\gamma'+1)\log(M) + \log(\tfrac{4DA}{C}), \log(i)\right), \tag{69}$$

$$\mathcal{B}(\Phi_{i,M}) \leq \pi_2\left(\left(\tfrac{C}{4DA} M^{-(\gamma'+1)}\right)^{-1}, i\right) = \pi_2\left(\tfrac{4DA}{C} M^{\gamma'+1}, i\right). \tag{70}$$

Consider now for $f \in \mathcal{C}$, $M \in \mathbb{N}$ the networks given by

$$\Psi_{f,M}(x) := \sum_{i \in I_{f,M}} c_{f,i} \Phi_{i,M}(x).$$

Thanks to $\max(I_{f,M}) \leq \pi_1(M)$, (69) implies the existence of a polynomial π_3 such that $\mathcal{L}(\Psi_{f,M}) \leq \pi_3(\log(M))$ and $\mathcal{M}(\Psi_{f,M}) \leq M \pi_3(\log(M))$, for all $f \in \mathcal{C}$, $M \in \mathbb{N}$, and, owing to 68,

$$\left\| \Psi_{f,M} - \sum_{i \in I_{f,M}} c_{f,i} \varphi_i \right\|_{L^2(\Omega)} \leq \sum_{i \in I_{f,M}} c_{f,i} \tfrac{C}{4DA} M^{-(\gamma'+1)} \leq \tfrac{C M^{-\gamma'}}{4A} \sum_{i=1}^{|I_{f,M}|} \tfrac{\max_{i \in I_{f,M}} |c_{f,i}|}{MD} \leq \tfrac{C M^{-\gamma'}}{4A}. \tag{71}$$

As the weights of the networks $\Phi_{i,M}$ are polynomially bounded in i, M and $\max(I_{f,M}) \leq \pi_1(M)$, we can conclude that the weights of the networks $\Psi_{f,M}$ are polynomially bounded in M. Lemma V.13 therefore ensures the existence of a polynomial π_4 such that for all $f \in \mathcal{C}$, $M \in \mathbb{N}$, there is a network $\widetilde{\Psi}_{f,M} \in \mathcal{NN}_{\infty,\infty,d,1}$ with $(\lceil \pi_4(\log(\varepsilon^{-1}))\rceil, \varepsilon)$-quantized weights satisfying $\mathcal{L}(\widetilde{\Psi}_{f,M}) = \mathcal{L}(\Psi_{f,M})$, $\mathcal{M}(\widetilde{\Psi}_{f,M}) = \mathcal{M}(\Psi_{f,M})$, $\mathcal{B}(\widetilde{\Psi}_{f,M}) \leq \mathcal{B}(\Psi_{f,M}) + \frac{CM^{-\gamma'}}{4A}$, and

$$\left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^\infty(\Omega)} \leq \tfrac{CM^{-\gamma'}}{4A}. \tag{72}$$

As Ω is bounded by assumption, we have

$$\left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^2(\Omega)} \leq |\Omega|^{1/2} \left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^\infty(\Omega)} \leq \tfrac{CM^{-\gamma'}}{4}, \tag{73}$$

for all $f \in \mathcal{C}$, $M \in \mathbb{N}$. Combining (73) with (67) and (71), we get, for all $f \in \mathcal{C}$, $M \in \mathbb{N}$,

$$\begin{aligned}\left\|f - \widetilde{\Psi}_{f,M}\right\|_{L^2(\Omega)} &\leq \left\|f - \sum_{i\in I_{f,M}} c_{f,i}\varphi_i\right\|_{L^2(\Omega)} + \left\|\sum_{i\in I_{f,M}} c_{f,i}\varphi_i - \Psi_{f,M}\right\|_{L^2(\Omega)} + \left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^2(\Omega)} \\ &\leq CM^{-\gamma'}.\end{aligned} \tag{74}$$

For $\varepsilon \in (0, 1/2)$ and $f \in \mathcal{C}$, we now set $M_\varepsilon := \left\lceil (C/\varepsilon)^{1/\gamma'} \right\rceil$ and

$$\Psi(\varepsilon, f) := \widetilde{\Psi}_{f,M_\varepsilon}.$$

Thus, (74) yields

$$\|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq CM_\varepsilon^{-\gamma'} \leq \varepsilon. \tag{75}$$

For the next step, we first note that, for all polynomials π and $0 \leq m < n$,

$$\mathcal{O}(\varepsilon^{-m}\pi(\log(\varepsilon^{-1}))) \subseteq \mathcal{O}(\varepsilon^{-n}),\ \varepsilon \to 0.$$

Since $1/\gamma' < 1/\gamma$ this establishes

$$\mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(M_\varepsilon \pi_3(\log(M_\varepsilon))) \subseteq \mathcal{O}(\varepsilon^{-1/\gamma}),\ \varepsilon \to 0. \tag{76}$$

Since M_ε and π_3 are independent of f, the implicit constant in (76) does not depend on f. By construction there exist constants c' and q so that the weights of the networks $\Psi(\varepsilon, f)$ can be represented with no more than $\lceil c'(\log(\varepsilon^{-1}))^q\rceil$ bits. Moreover, there exist a polynomial π_5 such that $\mathcal{L}(\Psi(\varepsilon, f)) \leq \pi_5(\log(\varepsilon^{-1}))$, for all $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, and a polynomial π_6 so that $\Psi(\varepsilon, f) \in \mathcal{NN}^{\pi_6(\mathcal{M}(\Psi(\varepsilon,f)))}_{\pi_6(\log(\mathcal{M}(\Psi(\varepsilon,f)))),\mathcal{M}(\Psi(\varepsilon,f)),d,1}$, for all $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$. Therefore, (76) implies

$$\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}^{\pi_6(M)}_{\pi_6(\log(M)),M,d,1}} \|f - \Phi\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M \to \infty.$$

Owing to Definition V.8, it therefore holds that

$$\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D}),$$

which concludes the proof. □

Remark VI.4. *We note that Theorem VI.2 continues to hold for $\Omega = \mathbb{R}^n$ if the elements of $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}}$ are compactly supported with the size of their support sets growing no more than polynomially in i. The technical elements required to show this can be found in the context of the approximation of Gabor systems in the proof of Theorem VIII.3, but are omitted here for ease of exposition.*

VII. Affine Representation Systems are Effectively Representable by Neural Networks

We now proceed to showing that a large class of representation systems, namely *affine systems*, including wavelets, ridgelets, curvelets, shearlets, and α-molecules, are effectively representable by neural networks. Thanks to Theorems VI.2 and V.9, this will then allow us to conclude that any function class that is optimally representable—in the sense of Definition V.5—by an affine system with a suitable generator function is optimally representable by neural networks in the sense of Definition V.10. By "suitable" we mean that the generator function can be approximated well by ReLU networks in a sense to be made precise below. Wavelets, for example, provide optimal representations (i.e., optimal non-linear approximation) of Besov spaces [59]. For concreteness, we consider the example of spline wavelet systems at the end of this section.

A. Invariance to Affine Transformations

Affine systems consist of translations and dilations of a given generator function. It is therefore important to understand the impact of these operations on the approximability—by neural networks—of a given function. As neural networks realize concatenations of affine functions and non-linearities, it is clear that translations and dilations can be absorbed into the first layer of the network and the transformed function should inherit its approximability from the generator function. What we will mostly be concerned with is how the weights in the network and its domain of approximation are affected.

Proposition VII.1. *Let $d \in \mathbb{N}$, $p \in [1,\infty]$, and $f \in L^p(\mathbb{R}^d)$. Assume that there exists a bivariate polynomial π_1 such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying*

$$\|f - \Phi_{D,\varepsilon}\|_{L^p([-D,D]^d)} \leq \varepsilon \tag{77}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi_1(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Then, there exists a bivariate polynomial π_2 such that for all full-rank $A \in \mathbb{R}^{d\times d}$, $e \in \mathbb{R}^d$, $E \in \mathbb{R}_+$, and $\eta \in (0,1/2)$, there is a network $\Psi_{A,e,E,\eta} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\left\| |\det(A)|^{\frac{1}{p}} f(A\cdot - e) - \Psi_{A,e,E,\eta} \right\|_{L^p([-E,E]^d)} \leq \eta$$

with $\mathcal{M}(\Psi_{A,e,E,\eta}) \leq \pi_2(\log(\eta^{-1}), \log(\lceil E\|A\|_\infty + \|e\|_\infty \rceil))$. Moreover, if the weights of the networks $\Phi_{D,\varepsilon}$ are polynomially bounded in (D, ε^{-1}), then the weights of the networks $\Psi_{A,e,E,\eta}$ are polynomially bounded in $(\|A\|_\infty, \|e\|_\infty, E, \eta^{-1})$.

Proof. By a change of variables, we have for every $\Phi \in \mathcal{NN}_{\infty,\infty,d,1}$,

$$\left\| |\det(A)|^{\frac{1}{p}} f(A \cdot - e) - |\det(A)|^{\frac{1}{p}} \Phi(A \cdot - e) \right\|_{L^p([-E,E]^d)} = \|f - \Phi\|_{L^p(A \cdot [-E,E]^d - e)}. \tag{78}$$

Furthermore, observe that

$$A \cdot [-E,E]^d - e \subset [-(dE\|A\|_\infty + \|e\|_\infty), (dE\|A\|_\infty + \|e\|_\infty)]^d. \tag{79}$$

We now set $F = \lceil dE\|A\|_\infty + \|e\|_\infty \rceil$ and $\Psi_{A,e,E,\eta} := |\det(A)|^{\frac{1}{p}} \Phi_{F,\eta}(A \cdot - e)$ and note that

$$\left\| |\det(A)|^{\frac{1}{p}} f(A \cdot - e) - \Psi_{A,e,E,\eta} \right\|_{L^p([-E,E]^d)} = \|f - \Phi_{F,\eta}\|_{L^p(A \cdot [-E,E]^d - e)} \leq \|f - \Phi_{F,\eta}\|_{L^p([-F,F]^d)} \leq \eta,$$

where we applied the same reasoning as in (78) in the first equality and used (79) in the first inequality and (77) in the second inequality. The existence of a polynomial π_2 such that $\mathcal{M}(\Psi_{A,e,E,\eta}) \leq \pi_2(\log(\eta^{-1}), \log(\lceil E\|A\|_\infty + \|e\|_\infty \rceil))$ follows by combining the definition of $\Psi_{A,e,E,\eta}$ with the assumption $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi_1(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Moreover, we see that if the weights of $\Phi_{F,\eta}$ are polynomially bounded in (F, η^{-1}), then the weights of $\Psi_{A,e,E,\eta}$ are polynomially bounded in $(\|A\|_\infty, |\det(A)|, \|e\|_\infty, E, \eta^{-1})$. Since $|\det(A)|$ is polynomially bounded in $\|A\|_\infty$, it follows that the weights of $\Psi_{A,e,E,\eta}$ are polynomially bounded in $(\|A\|_\infty, \|e\|_\infty, E, \eta^{-1})$. This yields the claim. □

The second auxiliary result we shall need concerns linear combinations of translates of a given function.

Proposition VII.2. *Let $r, d \in \mathbb{N}$, $p \in [1,\infty]$, and $f \in L^p(\mathbb{R}^d)$. Assume that there exists a bivariate polynomial π_1 such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying*

$$\|f - \Phi_{D,\varepsilon}\|_{L^p([-D,D]^d)} \leq \varepsilon \tag{80}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi_1(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Then, there exists a bivariate polynomial π_2 such that for all $c = (c_i)_{i=1}^r \subset \mathbb{R}$, $b = (b_i)_{i=1}^r \subset \mathbb{R}^d$, $E \in \mathbb{R}_+$, and $\eta \in (0, 1/2)$, there is a network $\Psi_{c,b,E,\eta} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\left\| \sum_{i=1}^r c_i f(\cdot - b_i) - \Psi_{c,b,E,\eta} \right\|_{L^p([-E,E]^d)} \leq \eta \tag{81}$$

with $\mathcal{M}(\Psi_{c,b,E,\eta}) \leq \pi_2(\log(\eta_c^{-1}), \log(E_b))$, where $E_b = \lceil E + \max_{i=1,\dots,r} \|b_i\|_\infty \rceil$ and $\eta_c = \eta / \max\{1, \sum_{i=1}^r |c_i|\}$. Moreover, if the weights of the networks $\Phi_{D,\varepsilon}$ are polynomially bounded in (D, ε^{-1}), then the weights of the networks $\Psi_{c,b,E,\eta}$ are polynomially bounded in $(\max\{1, \sum_{i=1}^r |c_i|\}, \max\{1, \max_{i=1,\dots,r} \|b_i\|_\infty\}, E, \eta^{-1})$.

Proof. It follows from assumption (80) that, for all $c = (c_i)_{i=1}^r \subset \mathbb{R}$, $b = (b_i)_{i=1}^r \subset \mathbb{R}^d$, $E \in \mathbb{R}_+$, and $\eta \in (0, 1/2)$,

$$\left\| \sum_{i=1}^r c_i f(\cdot - b_i) - \sum_{i=1}^r c_i \Phi_{E_b,\eta_c}(\cdot - b_i) \right\|_{L^p([-E,E]^d)} \leq \left(\sum_{i=1}^r |c_i| \right) \cdot \|f - \Phi_{E_b,\eta_c}\|_{L^p([-(E_b),(E_b)]^d)} \leq \eta.$$

The network $\Psi_{c,b,E,\eta} := \sum_{i=1}^r c_i \Phi_{E_b,\eta_c}(\cdot - b_i)$ thus satisfies (81). The existence of a bivariate polynomial π_2 such that $\mathcal{M}(\Psi_{c,b,E,\eta}) \leq \pi_2(\log(\eta_c^{-1}), \log(E_b))$ follows by combining $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi_1(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$, which is

by assumption, with Lemma II.7. Finally, it follows from $\Psi_{c,b,E,\eta} = \sum_{i=1}^{r} c_i \Phi_{E_b,\eta_c}(\cdot - b_i)$ that Φ_{E_b,η_c} with weights polynomially bounded in (E_b, η_c^{-1}) implies $\Psi_{c,b,E,\eta}$ with weights polynomially bounded in

$$\left(\max\left\{1, \sum_{i=1}^{r} |c_i|\right\}, \max\left\{1, \max_{i=1,\dots,r} \|b_i\|_\infty\right\}, E, \eta^{-1} \right).$$

□

B. Affine Representation Systems

We are now ready to introduce the class of representation systems announced above as *affine systems*. This class includes all representation systems based on affine transformations of a given generator function. Important special cases include wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules.

We proceed to the formal definition of affine systems.

Definition VII.3. *Let $d, r, S \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and $f \in L^\infty(\mathbb{R}^d)$ compactly supported. Let $\delta > 0$, $(c_k^s)_{k=1}^r \subset \mathbb{R}$, for $s = 0, \dots, S$, and $(b_k)_{k=1}^r \subset \mathbb{R}^d$. Further, let $A_j \in \mathbb{R}^{d\times d}, j \in \mathbb{N}$, be full-rank, with the absolute values of the eigenvalues of A_j bounded below by 1. Consider the compactly supported functions*

$$g_s := \sum_{k=1}^{r} c_k^s f(\cdot - b_k), \quad s = 0, \dots, S.$$

We define the affine system $\mathcal{D} \subset L^2(\Omega)$ *corresponding to $(g_s)_{s=0}^S$ according to*

$$\begin{aligned}\mathcal{D} := \Big\{ g_s^{j,e} := \Big(|\det(A_j)|^{\frac{1}{2}} g_s(A_j \cdot - \delta e)\Big)\big|_\Omega \colon\ s \in \{1, 2, \dots, S\},\ e \in \mathbb{Z}^d,\ j \in \mathbb{N},\ \text{and } g_s^{j,e} \neq 0 \Big\} \\ \cup \{ g_0^e := g_0(\cdot - \delta e)\big|_\Omega \colon\ e \in \mathbb{Z}^d \text{ and } g_0^e \neq 0\},\end{aligned}$$

and refer to f as the generator (function) of $\mathcal{D}$.

We define the sub-systems $\mathcal{D}_0 := \{g_0^e \in \mathcal{D} :\ e \in \mathbb{Z}^d\}$ and $\mathcal{D}_{s,j} := \{g_s^{j,e} \in \mathcal{D} :\ e \in \mathbb{Z}^d\}$, for $s \in \{1, 2, \dots, S\}$, $j \in \mathbb{N}$. Since every g_s, $s \in \{0, 1, \dots, S\}$, has compact support, $|\mathcal{D}_0|$ and $|\mathcal{D}_{s,j}|$, for all $s = 1, \dots, S$ and $j \in \mathbb{N}$, are finite. Indeed, we observe that there exists $c_{\mathrm{e}} := c_{\mathrm{e}}(\Omega, (g_s)_{s=0}^S, \delta, d) > 0$ such that for all $s \in \{1, \dots, S\}$, $j \in \mathbb{Z}$, and $e \in \mathbb{Z}^d$,

$$\begin{aligned} g_s^{j,e} \in \mathcal{D}_{s,j} &\implies \|e\|_\infty \leq c_e \|A_j\|_\infty, \text{ and} \\ g_0^e \in \mathcal{D}_0 &\implies \|e\|_\infty \leq c_e. \end{aligned} \tag{82}$$

As all subsystems $\mathcal{D}_0$ and $\mathcal{D}_{s,j}$ are finite, we can organize the representation system $\mathcal{D}$ according to

$$\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} = (\mathcal{D}_0, \mathcal{D}_{1,1}, \dots, \mathcal{D}_{S,1}, \mathcal{D}_{1,2}, \dots, \mathcal{D}_{S,2}, \dots), \tag{83}$$

where the elements within each sub-system may be ordered arbitrarily. This ordering of $\mathcal{D}$ is assumed in the remainder of the paper and will be referred to as *canonical ordering*.

Moreover, we note that if there exists $s_o \in \{1, \dots, S\}$ such that g_{s_o} is nonzero, then there is a constant $c_{\mathrm{o}} := c_{\mathrm{o}}(\Omega, (g_s)_{s=1}^S, \delta, d) > 0$ such that

$$\sum_{s=1}^{S} |\mathcal{D}_{s,j}| \geq c_{\mathrm{o}} |\det(A_j)|, \text{ for all } j \in \mathbb{N}. \tag{84}$$

The next result establishes that affine systems with generator function that can be approximated well by neural networks are effectively representable by neural networks.

Theorem VII.4. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ an affine system with generator function f. Assume that there exists a bivariate polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying*

$$\|f - \Phi_{D,\varepsilon}\|_{L^2([-D,D])} \leq \varepsilon \tag{85}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(\varepsilon^{-1}, D)$. Assume furthermore that there exist $a, c > 0$ such that

$$\sum_{k=1}^{j-1} |\det(A_k)| \geq c \|A_j\|_\infty^a, \quad \textit{for all } j \in \mathbb{N},\ j \geq 2. \tag{86}$$

Then, $\mathcal{D}$ is effectively representable by neural networks.

Proof. Let $(g_s)_{s=0}^S$ be as in Definition VII.3. If $g_s = 0$ for all $s \in \{0, \dots, S\}$, then $\mathcal{D} = \varnothing$ and the result is trivial. We can hence assume that there exists at least one $s \in \{0, \dots, S\}$ such that $g_s \neq 0$, which, in turn, implies that (84) holds.

Pick D such that $\Omega \subset [-D, D]^d$. By Definition VI.1 we need to establish the existence of a bivariate polynomial π such that for each $i \in \mathbb{N}$, $\eta \in (0, 1/2)$, there is a network $\Phi_{i,\eta} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|\varphi_i - \Phi_{i,\eta}\|_{L^2(\Omega)} \leq \eta \tag{87}$$

with $\mathcal{M}(\Phi_{i,\eta}) \leq \pi(\log(\eta^{-1}), \log(i))$ and $\mathcal{B}(\Phi_{i,\eta}) \leq \pi(\eta^{-1}, i)$. Note that the elements of $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ consist of dilations and translations of f according to

$$\varphi_i = \left(|\det(A_{j_i})|^{\frac{1}{2}} g_{s_i}(A_{j_i} \cdot - \delta e_i) \right) \big|_\Omega \text{ or } \varphi_i = g_0(\cdot - \delta e_i)\big|_\Omega, \tag{88}$$

for $s_i \in \{1, \dots, S\}$, $j_i \in \mathbb{N}$, and $e_i \in \mathbb{Z}^d$, where

$$g_s := \sum_{k=1}^{r} c_k^s f(\cdot - b_k), \quad s = 0, \dots, S,$$

for some $r, S \in \mathbb{N}$. Thanks to (85), combining Propositions VII.2 and VII.1 establishes the claim provided that the quantities

$$\|A_{j_i}\|_\infty, D, \|e_i\|_\infty, \max_{s\in\{0,\dots,S\}} \sum_{k=1}^{r} |c_k^s|, \max_{k=1,\dots,r} \|b_k\|_\infty \tag{89}$$

are polynomially bounded in i. To see that this is, indeed, the case, we start by noting that $\|e_i\|_\infty \in \mathcal{O}(\|A_{j_i}\|_\infty)$ thanks to (82). Moreover, since Ω is bounded and r is fixed, the quantities D, $\max_{s\in\{0,\dots,S\}} \sum_{k=1}^{r} |c_k^s|$, and

$\max_{k=1,\dots,r} \|b_k\|_\infty$ do not depend on i. It therefore suffices to establish that $\|A_{j_i}\|_\infty$ is polynomially bounded in i. To this end simply note that thanks to (86) and (84), for all $i \in \mathbb{N}$ with $j_i \geq 2$,

$$c\|A_{j_i}\|_\infty^a \leq \sum_{k=1}^{j_i-1} |\det(A_k)| \leq \tfrac{1}{c_0} \sum_{k=1}^{j_i-1} \sum_{s=1}^{S} |\mathcal{D}_{s,k}| \leq \tfrac{1}{c_0} i,$$

where the last inequality follows from the fact that $\varphi_i \in \mathcal{D}_{s_i,j_i}$ for some s_i. This ensures that

$$\|A_{j_i}\|_\infty \leq \|A_1\|_\infty + \left(\frac{1}{c_0 c} i\right)^{\frac{1}{a}}, \quad \text{for all } i \in \mathbb{N},$$

thereby completing the proof. □

Remark VII.5. *Theorem VII.4 is restricted to bounded Ω and compactly supported generator function f for ease of exposition. It can be extended to $\Omega = \mathbb{R}$ and generator functions f of unbounded support but sufficiently fast decay. This extension requires additional technical steps and an alternative definition of canonical ordering. For conciseness we do not provide the details here, but refer to the proofs of Theorems VIII.3 and VIII.5, which deal with the corresponding technical aspects in the context of approximation of Gabor systems by neural networks. We further remark that condition (86) is very weak; in fact, we are not aware of any affine systems in the literature that would violate it.*

We proceed to establishing a remarkable universality and optimality property of neural networks: Neural networks provide optimal approximations for all function classes that are optimally approximated by affine systems generated by functions f that can be approximated well by neural networks.

Theorem VII.6. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ an affine system with generator function f. Assume that there exists a bivariate polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying $\|f - \Phi_{D,\varepsilon}\|_{L^2([-D,D])} \leq \varepsilon$ with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(\varepsilon^{-1}, D)$, and there are constants $a, c > 0$ such that* (86) *holds. Then, we have*

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D})$$

for all function classes $\mathcal{C} \subseteq L^2(\Omega)$. In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{D}$ (in the sense of Definition V.5), then $\mathcal{C}$ is optimally representable by neural networks (in the sense of Definition V.10).

Proof. The first statement follows from Theorems VI.2 and VII.4, the second is by Theorem V.9. □

C. Spline wavelets

We next particularize the results developed in the previous two subsections to show that neural networks optimally represent all function classes that are optimally representable by spline wavelet systems. As spline wavelet systems have B-splines as generator functions, we start by defining B-splines. For simplicity of exposition, we shall restrict ourselves to the univariate case throughout.

Definition VII.7. *Let $N_1 := \chi_{[0,1]}$ and for $m \in \mathbb{N}$, define*

$$N_{m+1} := N_1 * N_m.$$

We refer to N_m as the univariate cardinal B-spline of order m.

Recognizing that B-splines are piecewise polynomial, we can apply Proposition III.3 to get the following statement on the approximation of B-splines by deep neural networks.

Lemma VII.8. *Let $m \in \mathbb{N}$. Then, there exist a constant $C > 0$ and a polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying*

$$\|\Phi_{D,\varepsilon} - N_m\|_{L^\infty([-D,D])} \leq \varepsilon$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq C(\log(\varepsilon^{-1}) + \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(D)$.

Proof. The proof is based on the following representation [60, Eq. 19]

$$N_m(x) = \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho((x-k)^m).$$

Note that the spline order m is fixed in the following and we are concerned with the approximating network's dependence on D and ε only. Proposition III.3 now ensures the existence of a constant $C_1 > 0$ and a polynomial π_1 such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\|\Psi_{D,\varepsilon}(x) - x^m\|_{L^\infty([-(D+m+1),D+m+1])} \leq \frac{\varepsilon}{2(m+2)}$$

with $\mathcal{M}(\Psi_{D,\varepsilon}) \leq C_1(\log(\varepsilon^{-1}) + \log(\lceil D \rceil))$ and $\mathcal{B}(\Psi_{D,\varepsilon}) \leq \pi_1(D)$. Next, define, for $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, the network

$$\Phi_{D,\varepsilon} := \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho(\Psi_{D,\varepsilon}(\cdot - k))$$

and note that since the component networks $\rho(\Psi_{D,\varepsilon}(\cdot - k))$ simply have one layer more than $\Psi_{D,\varepsilon}(\cdot - k)$, there exists a constant C_2 such that $\mathcal{M}(\Phi_{D,\varepsilon}) \leq C_2(\log(\varepsilon^{-1}) + \log(\lceil D \rceil))$. Moreover, thanks to

$$\frac{1}{m!} \binom{m+1}{k} = \frac{m+1}{k!(m-k+1)!} \leq 2 \tag{90}$$

it follows from Lemmata II.5 and II.7 that there exists a polynomial π_2 such that $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi_2(D)$.

Finally, since ρ is 1-Lipschitz, we have for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}\|\Phi_{D,\varepsilon} - N_m\|_{L^\infty([-D,D])} &\leq \sum_{k=0}^{m+1} \frac{1}{m!} \binom{m+1}{k} \|\rho(\Psi_{D,\varepsilon}(\cdot - k)) - \rho((\cdot - k)^m)\|_{L^\infty([-D,D])} \\ &\leq 2 \sum_{k=0}^{m+1} \|\Psi_{D,\varepsilon}(x) - x^m\|_{L^\infty([-(D+m+1),D+m+1])} \leq \varepsilon,\end{aligned}$$

where we used (90). This completes the proof. □

We are now ready to introduce spline wavelet systems. For $k, j \in \mathbb{Z}$, set

$$V_k := \mathrm{clos}_{L^2}\Big(\mathrm{span}\,\{N_m(2^k x - j) : j \in \mathbb{Z}\}\Big), \tag{91}$$

where clos_{L^2} denotes closure with respect to L^2-norm. Spline spaces V_k, $k \in \mathbb{Z}$, constitute a multiresolution analysis of $L^2(\mathbb{R})$ according to

$$\{0\} \subset \ldots V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset L^2(\mathbb{R}).$$

Moreover, with the orthogonal complements $(\ldots, W_{-1}, W_0, W_1, \ldots)$ such that $V_{k+1} = V_k \oplus W_k$, where $\oplus$ denotes the orthogonal sum, we have

$$L^2(\mathbb{R}) = V_0 \oplus \bigoplus_{k=0}^{\infty} W_k.$$

Theorem VII.9 ([61, Theorem 1])**.** *Let $m \in \mathbb{N}$. The m-th order spline*

$$\psi_m(x) = \frac{1}{2^{m-1}} \sum_{j=0}^{2m-2} (-1)^j N_{2m}(j+1) \frac{d^m}{dx^m} N_{2m}(2x - j), \tag{92}$$

with support $[0, 2m-1]$, is a basic wavelet that generates W_0 and thereby all the wavelet spaces W_k, $k \in \mathbb{Z}$. Consequently, the set

$$\mathcal{W}_m := \{\psi_{k,n}(x) = 2^{k/2}\psi_m(2^k x - n) : n \in \mathbb{Z}, k \in \mathbb{N}_0\} \cup \{\phi_n(x) = N_m(x - n) : n \in \mathbb{Z}\} \tag{93}$$

is a countable complete orthonormal wavelet basis in $L^2(\mathbb{R})$.

It remains to establish that the spline wavelet system (93) is an affine system in the sense of Definition VII.3. To this end, we first devise an alternative representation of (92). Specifically, using the identity [61, Eq. 2.2]

$$\frac{d^m}{dx^m} N_{2m}(x) = \sum_{j=0}^{m} (-1)^j \binom{m}{j} N_m(x - j),$$

we get

$$\psi_m(x) = \sum_{n=1}^{3m-1} q_n N_m(2x - n + 1), \tag{94}$$

with

$$q_n = \frac{(-1)^{n+1}}{2^{m-1}} \sum_{j=0}^{m} \binom{m}{j} N_{2m}(n - j).$$

Combining (94) and the two-scale relation [61, Eq. 1.9]

$$N_m(x) = \sum_{j=0}^{m} 2^{-m+1} \binom{m}{j} N_m(2x - j), \tag{95}$$

we next establish that the spline wavelet system (93) is an affine system in the sense of Definition VII.3 with generator $f(x) = N_m(2x)$. In particular, in the notation of Definition VII.3, let $d = 1$, $S = 1$, $r = 4m$, $\delta = 1$,

$$c_k^1 = \begin{cases} q_k, & 1 \leq k \leq 3m-1 \\ 0, & 3m \leq k \leq r \end{cases},$$

$$c_k^0 = \begin{cases} 0, & 1 \leq k \leq 3m-1 \\ 2^{-m+1}\begin{pmatrix} m \\ k-3m \end{pmatrix}, & 3m \leq k \leq r \end{cases}, \tag{96}$$

$$b_k = \begin{cases} \frac{k-1}{2}, & 1 \leq k \leq 3m-1 \\ \frac{k-3m}{2}, & 3m \leq k \leq r \end{cases},$$

and $A_j = 2^{j-1}$, for $j \in \mathbb{N}$. Then, we have $g_1 = \psi_m$, $g_0 = N_m$, and the affine system $\mathcal{D} \subset L^2(\Omega)$ corresponding to g_1, g_0 is

$$\begin{aligned} \mathcal{D} :=& \left\{ g_1^{j,e}(x) := \left(|A_j|^{\frac{1}{2}} g_1(A_j \cdot - \delta e) \right)\Big|_\Omega : \ e \in \mathbb{Z}, \ j \in \mathbb{N}, \text{ and } g_1^{j,e} \neq 0 \right\} \\ & \cup \{ g_0^e := g_0(\cdot - \delta e)\big|_\Omega : \ e \in \mathbb{Z} \text{ and } g_0^e \neq 0 \} \\ =& \mathcal{W}_m. \end{aligned} \tag{97}$$

We have therefore established the following.

Theorem VII.10. *Let $\Omega \subset \mathbb{R}$ be bounded and $\mathcal{D} = (\varphi_i)_{i \in \mathbb{N}} \subset L^2(\Omega)$ a spline wavelet system according to* (97)*. Then, all function classes $\mathcal{C} \subseteq L^2(\Omega)$ that are optimally representable by $\mathcal{D}$ (in the sense of Definition V.5), are optimally representable by neural networks (in the sense of Definition V.10).*

Proof. Note that (97) is an affine system with generator function $f(x) := N_m(2x)$ in the sense of Definition VII.3 and condition (86) holds for $a = 1$, and $c = 1/2$, as

$$\sum_{k=1}^{j-1} 2^{k-1} = \sum_{k=0}^{j-2} 2^k = 2^{j-1} - 1 \geq \tfrac{1}{2} 2^{j-1},$$

for all $j \geq 2$. Now let $\Phi_{D,\varepsilon}$ be the networks from Lemma VII.8 and consider the networks $\Psi_{D,\varepsilon}(x) := \Phi_{2D,\eta}(2x)$ with $\eta := (2D)^{-\frac{1}{2}}\varepsilon$. Then we have for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$ that

$$\begin{aligned} \|\Psi_{D,\varepsilon} - f\|_{L^2([-D,D])} &\leq (2D)^{\frac{1}{2}} \|\Psi_{D,\varepsilon} - f\|_{L^\infty([-D,D])} \\ &= (2D)^{\frac{1}{2}} \|\Phi_{2D,\eta}(2\,\cdot) - N_m(2\,\cdot)\|_{L^\infty([-D,D])} \\ &= (2D)^{\frac{1}{2}} \|\Phi_{2D,\eta} - N_m\|_{L^\infty([-2D,2D])} \leq (2D)^{\frac{1}{2}} \eta = \varepsilon. \end{aligned}$$

Lemma VII.8 further ensures the existence of a constant $C \in \mathbb{R}_+$ and a polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$ it holds that $\mathcal{M}(\Psi_{D,\varepsilon}) \leq C(\log((2D)^{-\frac{1}{2}}\varepsilon^{-1}) + \log(\lceil 2D \rceil))$ and $\mathcal{B}(\Psi_{D,\varepsilon}) \leq \pi(2D)$. Thus Theorem VII.6 establishes the claim. □

VIII. GABOR SYSTEMS

Affine representation systems discussed in the previous section have the affine group underlying as their generating structure. In this section, we consider Gabor systems [14], which are generated by the Weyl-Heisenberg group [62], and consist of time-frequency translates of a given generator function. Gabor systems play a fundamental role in time-frequency analysis [14] and in the study of partial differential equations [63]. We start with the formal definition of Gabor systems.

Definition VIII.1 (Gabor systems)**.** *Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d)$, and $x, \xi \in \mathbb{R}^d$. Define the translation operator $T_x \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ according to*

$$T_x f(t) := f(t-x)$$

and the modulation operator $M_\xi \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d, \mathbb{C})$ as

$$M_\xi f(t) := e^{2\pi i \langle \xi, t \rangle} f(t).$$

Let $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, and $g \in L^2(\mathbb{R}^d)$. The Gabor system $\mathcal{G}(g, \alpha, \beta, \Omega) \subseteq L^2(\Omega)$ is defined as

$$\mathcal{G}(g, \alpha, \beta, \Omega) := \left\{ M_\xi T_x g\big|_\Omega \colon (x, \xi) \in \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \right\}. \tag{98}$$

In order to be able to talk about representability in the sense of Definition VI.1, we need to order the elements in $\mathcal{G}(g, \alpha, \beta, \Omega)$. To this end, let $\mathcal{G}_0(g, \alpha, \beta, \Omega) := \{g\big|_\Omega\}$ and define $\mathcal{G}_n(g, \alpha, \beta, \Omega)$, $n \in \mathbb{N}$, recursively according to

$$\mathcal{G}_n(g, \alpha, \beta, \Omega) := \{M_\xi T_x g\big|_\Omega \colon (x, \xi) \in \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d, \|x\|_\infty \leq n\alpha, \|\xi\|_\infty \leq n\beta\} \setminus \bigcup_{k=0}^{n-1} \mathcal{G}_k(g, \alpha, \beta, \Omega).$$

We then organize $\mathcal{G}(g, \alpha, \beta, \Omega)$ according to

$$\mathcal{G}(g, \alpha, \beta, \Omega) = (\mathcal{G}_0(g, \alpha, \beta, \Omega), \mathcal{G}_1(g, \alpha, \beta, \Omega), \dots), \tag{99}$$

where the ordering within the sets $\mathcal{G}_n(g, \alpha, \beta, \Omega)$ is arbitrary. Note that the specifics of the overall ordering in (99) are irrelevant as long as $\mathcal{G}(g, \alpha, \beta, \Omega) = (\varphi_i)_{i \in \mathbb{N}}$ with $\varphi_i = \mathcal{M}_{\xi(i)} T_{x(i)} g\big|_\Omega$ is such that $\|x(i)\|_\infty$ and $\|\xi(i)\|_\infty$ do not grow faster than polynomial in i; this will become apparent in the proof of Theorem VIII.3.

As Gabor systems are built from time-shifted and modulated versions of a generator function, we first establish that the approximation-theoretic properties of a given function, here the generator function, are inherited by its modulated versions. Time shifts are dealt with straightforwardly as they can be incorporated into the affine transform in the first layer of the network.

Lemma VIII.2. *Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and for every $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, let $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ be a network satisfying*

$$\|f - \Phi_{D,\varepsilon}\|_{L^\infty([-D,D]^d)} \leq \varepsilon. \tag{100}$$

Then, there exists a constant $C > 0$, not depending on d and f, such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, $\xi \in \mathbb{R}^d$, there are networks $\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}, \Phi^{\mathrm{Im}}_{D,\xi,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|\mathrm{Re}(M_\xi f) - \Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}\|_{L^\infty([-D,D]^d)} + \|\mathrm{Im}(M_\xi f) - \Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}\|_{L^\infty([-D,D]^d)} \leq 3\varepsilon$$

with

$$\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}), \mathcal{L}(\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) + (\log(\lceil S_f \rceil))^2) + \mathcal{L}(\Phi_{D,\varepsilon}),$$

$$\mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}), \mathcal{M}(\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) + (\log(\lceil S_f \rceil))^2 + d) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Phi_{D,\varepsilon}),$$

where $S_f := \max\{1, \|f\|_{L^\infty(\mathbb{R}^d)}\}$. Moreover, if the weights of the networks $\Phi_{D,\varepsilon}$ are polynomially bounded in (D, ε^{-1}), then the weights of the networks $\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}, \Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}$ are polynomially bounded in $(D, \varepsilon^{-1}, \|\xi\|_\infty)$.

Proof. All statements in the proof involving ε pertain to $\varepsilon \in (0, 1/2)$ without explicitly stating this every time. We start by observing that

$$\mathrm{Re}(M_\xi f)(t) = \cos(2\pi\langle \xi, t\rangle) f(t)$$

$$\mathrm{Im}(M_\xi f)(t) = \sin(2\pi\langle \xi, t\rangle) f(t)$$

thanks to $f \in \mathbb{R}$. Note that for given $\xi \in \mathbb{R}^d$, the map $t \mapsto \langle \xi, t\rangle = \xi^T t = t_1\xi_1 + \cdots + t_d\xi_d$ is simply a linear transformation. Theorem IV.1 hence guarantees the existence of a constant C_1 such that for all $D \in \mathbb{R}_+$, $\xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{D,\xi,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\sup_{t \in [-D,D]^d} |\cos(2\pi\langle \xi, t\rangle) - \Psi_{D,\xi,\varepsilon}(t)| \leq \tfrac{\varepsilon}{6S_f} \tag{101}$$

with

$$\mathcal{L}(\Psi_{D,\xi,\varepsilon}) \leq C_1((\log(1/\varepsilon))^2 + (\log(S_f))^2 + \log(\lceil dD\|\xi\|_\infty \rceil)), \tag{102}$$

$$\mathcal{M}(\Psi_{D,\xi,\varepsilon}) \leq C_1((\log(1/\varepsilon))^2 + (\log(S_f))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) + d), \tag{103}$$

and $\mathcal{B}(\Psi_{D,\xi,\varepsilon}) \leq C_1\|\xi\|_\infty$. Moreover, Proposition III.2 guarantees the existence of a constant $C_2 > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\mu_\varepsilon \in \mathcal{NN}_{\infty,\infty,2,1}$ satisfying

$$\sup_{x,y \in [-S_f - 1/2, S_f + 1/2]} |\mu_\varepsilon(x, y) - xy| \leq \tfrac{\varepsilon}{6} \tag{104}$$

with

$$\mathcal{L}(\mu_\varepsilon), \mathcal{M}(\mu_\varepsilon) \leq C_2(\log(1/\varepsilon) + \log(\lceil S_f \rceil)) \tag{105}$$

and $\mathcal{B}(\mu_\varepsilon) \leq \max\{4, 2\lceil S_f + 1/2\rceil^2\}$. By Lemma II.7, the network $\Phi(t) := (\Psi_{D,\xi,\varepsilon}(t), \Phi_{D,\varepsilon}(t)) \in \mathcal{NN}_{\infty,\infty,d,2}$ satisfies

$$\mathcal{M}(\Phi) \leq 2\,\mathcal{M}(\Psi_{D,\xi,\varepsilon}) + 2\,\mathcal{M}(\Phi_{D,\varepsilon}) + 2\,\mathcal{L}(\Psi_{D,\xi,\varepsilon}) + 2\,\mathcal{L}(\Phi_{D,\varepsilon}), \tag{106}$$

$$\mathcal{L}(\Phi) \leq \max\{\mathcal{L}(\Psi_{D,\xi,\varepsilon}), \mathcal{L}(\Phi_{D,\varepsilon})\}, \tag{107}$$

and $\mathcal{B}(\Phi) \leq \max\{\mathcal{B}(\Psi_{D,\xi,\varepsilon}), \mathcal{B}(\Phi_{D,\varepsilon})\}$. Finally, applying Lemma II.5 to concatenate the networks Φ and μ_ε, we obtain the network

$$\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}(t) := \mu_\varepsilon(\Phi(t)) = \mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(t), \Phi_{D,\varepsilon}(t)) \in \mathcal{NN}_{\infty,\infty,d,1}$$

satisfying

$$\mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 4\mathcal{M}(\Psi_{D,\xi,\varepsilon}) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Psi_{D,\xi,\varepsilon}) + 4\mathcal{L}(\Phi_{D,\varepsilon}) + 2\mathcal{M}(\mu_\varepsilon), \tag{108}$$

and

$$\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq \max\{\mathcal{L}(\Psi_{D,\xi,\varepsilon}), \mathcal{L}(\Phi_{D,\varepsilon})\} + \mathcal{L}(\mu_\varepsilon). \tag{109}$$

Next, observe that (101) and (104) imply

$$\begin{aligned}
\|\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon} - \mathrm{Re}(M_\xi f)\|_{L^\infty([-D,D]^d)} &= \|\mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(\,\cdot\,), \Phi_{D,\varepsilon}(\,\cdot\,)) - \cos(2\pi\langle\xi,\,\cdot\,\rangle) f(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\leq \|\mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(\,\cdot\,), \Phi_{D,\varepsilon}(\,\cdot\,)) - \Psi_{D,\xi,\varepsilon}(\,\cdot\,)\Phi_{D,\varepsilon}(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\quad + \|\Psi_{D,\xi,\varepsilon}(\,\cdot\,)\Phi_{D,\varepsilon}(\,\cdot\,) - \cos(2\pi\langle\xi,\,\cdot\,\rangle) f(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\leq \|\mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(\,\cdot\,), \Phi_{D,\varepsilon}(\,\cdot\,)) - \Psi_{D,\xi,\varepsilon}(\,\cdot\,)\Phi_{D,\varepsilon}(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\quad + \|\Psi_{D,\xi,\varepsilon}(\,\cdot\,)(\Phi_{D,\varepsilon}(\,\cdot\,) - f(\,\cdot\,))\|_{L^\infty([-D,D]^d)} \\
&\quad + \|\Psi_{D,\xi,\varepsilon}(\,\cdot\,) f(\,\cdot\,) - \cos(2\pi\langle\xi,\,\cdot\,\rangle) f(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\leq \tfrac{\varepsilon}{6} + (1 + \tfrac{\varepsilon}{6S_f})\varepsilon + \tfrac{\varepsilon}{6} \leq \tfrac{3}{2}\varepsilon.
\end{aligned}$$

Combining (108), (109), (102), and (105), we can further see that there exists a constant $C > 0$ such that

$$\begin{aligned}
&\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) + (\log(\lceil S_f \rceil))^2) + \mathcal{L}(\Phi_{D,\varepsilon}), \\
&\mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) + (\log(\lceil S_f \rceil))^2 + d) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Phi_{D,\varepsilon}).
\end{aligned}$$

It finally follows from Lemmata II.5 and II.7, Proposition III.2, and $\mathcal{B}(\Psi_{D,\xi,\varepsilon}) \leq C_1\|\xi\|_\infty$ that the weights of $\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}$ are polynomially bounded in $(D, \varepsilon^{-1}, \|\xi\|_\infty)$ if the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in (D, ε^{-1}). The results for $\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}$ follow analogously simply by noting that $\sin(x) = \cos(x - \pi/2)$. This completes the proof. □

Note that Gabor systems necessarily contain complex-valued functions. The theory developed so far applies, however, to neural networks with real-valued outputs. As evident from the proof of Lemma VIII.2, this is not an issue when the generator function g is real-valued. For complex-valued generator functions one needs a version of Proposition III.2 that applies to the multiplication of complex numbers. Thanks to $(a + ib)(a' + ib') = (aa' - bb') + i(ab' + a'b)$ such a network can be constructed by realizing the real and imaginary parts of the product as a sum of real-valued multiplication networks and then proceeding as in the proof above. We omit the details as they are straightforward and would not lead to new conceptual insights. Furthermore, an extension—to the

complex-valued case—of the concept of (effective) representability according to Definition VI.1 is needed. This is effected by considering the set of neural networks with 1-dimensional complex-valued output as neural networks with 2-dimensional real-valued output, i.e., by setting

$$\mathcal{NN}^{\mathbb{C}}_{\infty,\infty,d,1} := \mathcal{NN}_{\infty,\infty,d,2},$$

with the convention that the first component represents the real part and the second the imaginary part.

We proceed to establishing conditions for effective representability of Gabor systems by neural networks.

Theorem VIII.3. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, $g \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and let $\mathcal{G}(g, \alpha, \beta, \Omega)$ be the corresponding Gabor system with ordering as defined in* (99)*. Assume that either Ω is bounded or g is compactly supported. Further, suppose that there exists a polynomial π such that for every $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying*

$$\|g - \Phi_{D,\varepsilon}\|_{L^\infty([-D,D]^d)} \leq \varepsilon \tag{110}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(\varepsilon^{-1}, D)$. Then, $\mathcal{G}(g, \alpha, \beta, \Omega)$ is effectively representable by neural networks.

Proof. First note that owing to (99), we have $\mathcal{G}(g, \alpha, \beta, \Omega) = (\varphi_i)_{i\in\mathbb{N}}$ with $\varphi_i = \mathcal{M}_{\xi(i)} T_{x(i)} g \in \mathcal{G}_{n(i)}(g, \alpha, \beta, \Omega)$, where

$$\|\xi(i)\|_\infty \leq n(i)\beta \leq i\beta \quad \text{and} \quad \|x(i)\|_\infty \leq n(i)\alpha \leq i\alpha. \tag{111}$$

We start by considering the case where Ω is bounded, i.e., there exists a constant $D > 0$ such that $\Omega \subseteq [-D, D]^d$. Combining Proposition VII.2 and Lemma VIII.2, with $\|g\|_{L^\infty(\mathbb{R}^d)}$ considered constant, we can infer the existence of a multivariate polynomial π_1 such that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{i,\varepsilon} = (\Phi^{\mathrm{Re}}_{i,\varepsilon}, \Phi^{\mathrm{Im}}_{i,\varepsilon}) \in \mathcal{NN}^{\mathbb{C}}_{\infty,\infty,d,1}$ satisfying

$$\|\mathrm{Re}(\mathcal{M}_{\xi(i)} T_{x(i)} g) - \Phi^{\mathrm{Re}}_{i,\varepsilon}\|_{L^\infty(\Omega)} + \|\mathrm{Im}(\mathcal{M}_{\xi(i)} T_{x(i)} g) - \Phi^{\mathrm{Im}}_{i,\varepsilon}\|_{L^\infty(\Omega)} \leq (2D)^{-\frac{d}{2}} \varepsilon. \tag{112}$$

with

$$\begin{aligned} \mathcal{M}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{M}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) &\leq \pi_1(\log(\varepsilon^{-1}), \log(\|\xi(i)\|_\infty), \log(\|x(i)\|_\infty)), \\ \mathcal{B}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{B}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) &\leq \pi_1(\varepsilon^{-1}, \|\xi(i)\|_\infty, \|x(i)\|_\infty). \end{aligned}$$

As $|z| \leq |\mathrm{Re}(z)| + |\mathrm{Im}(z)|$, it follows from (112) that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned} \|\varphi_i - \Phi_{i,\varepsilon}\|_{L^2(\Omega,\mathbb{C})} &\leq (2D)^{\frac{d}{2}} \|\varphi_i - \Phi_{i,\varepsilon}\|_{L^\infty(\Omega,\mathbb{C})} \\ &\leq (2D)^{\frac{d}{2}} \left(\|\mathrm{Re}(\varphi_i) - \Phi^{\mathrm{Re}}_{i,\varepsilon}\|_{L^\infty(\Omega)} + \|\mathrm{Im}(\varphi_i) - \Phi^{\mathrm{Im}}_{i,\varepsilon}\|_{L^\infty(\Omega)} \right) \leq \varepsilon. \end{aligned}$$

Moreover, (111) implies the existence of a polynomial π_2 such that $\mathcal{M}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{M}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) \leq \pi_2(\log(\varepsilon^{-1}), \log(i))$ and $\mathcal{B}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{B}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) \leq \pi_2(\varepsilon^{-1}, i)$, for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$.We can therefore conclude that $\mathcal{G}(g, \alpha, \beta, \Omega)$ is effectively

representable by neural networks.

We proceed to proving the statement for the case where g is compactly supported. To be concrete, we let $\mathrm{supp}(g) \subseteq [-E, E]^d$ with $E > 0$. This implies

$$\mathrm{supp}(M_\xi T_x g) = \mathrm{supp}(T_x g) \subseteq x + [-E, E]^d \subseteq [-(\|x\|_\infty + E), \|x\|_\infty + E]^d. \tag{113}$$

Again, combining Proposition VII.2 and Lemma VIII.2 ensures the existence of a polynomial π_3 such that for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there are networks $\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \Psi^{\mathrm{Im}}_{x,\xi,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|\mathrm{Re}(M_\xi T_x g) - \Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} + \|\mathrm{Im}(M_\xi T_x g) - \Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \leq \tfrac{\varepsilon}{2 s_x}, \tag{114}$$

with

$$\begin{aligned}
\mathcal{M}(\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{M}(\Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}) &\leq \pi_3(\log(\varepsilon^{-1}), \log(\|x\|_\infty), \log(\|\xi\|_\infty)), \\
\mathcal{B}(\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{B}(\Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}) &\leq \pi_3(\varepsilon^{-1}, \|x\|_\infty, \|\xi\|_\infty),
\end{aligned}$$

and where we set $S_x := [-(\|x\|_\infty + E + 1), \|x\|_\infty + E + 1]^d$ and $s_x := |S_x|^{1/2}$ to simplify notation. As we want to establish effective representability for general, possibly unbounded, domains Ω, the estimate in (114) is insufficient. In particular, we have no control over the behavior of the networks $\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}$ outside the set S_x. We can, however, construct networks which exhibit the same scaling behavior in terms of $\mathcal{M}$ and $\mathcal{B}$, are strictly supported in S_x, and realize the same output for all inputs in $[-(\|x\|_\infty + E), \|x\|_\infty + E]^d$. To this end let, for $y \in \mathbb{R}_+$, the network $\alpha_y \in \mathcal{NN}_{2,8,1,1}$ be given by

$$\alpha_y(t) := \rho(t - (-y - 1)) - \rho(t - (-y)) - \rho(t - y) + \rho(t - (y + 1)), \quad t \in \mathbb{R}.$$

Note that $\alpha_y(t) = 1$ for $t \in [-y, y]$, $\alpha_y(t) = 0$ for $t \notin [-y - 1, y + 1]$, and $\alpha_y(t) \in (0, 1)$ else. Next, consider, for $x \in \mathbb{R}^d$, the network given by

$$\chi_x(t) := \rho\left(\left[\sum_{i=1}^d \alpha_{\|x\|_\infty + E}(t_i)\right] - (d - 1)\right), \quad t = (t_1, t_2, \ldots, t_d) \in \mathbb{R}^d,$$

and note that

$$\begin{aligned}
\chi_x(t) &= 1, \quad \forall t \in [-(\|x\|_\infty + E), \|x\|_\infty + E]^d \\
\chi_x(t) &= 0, \quad \forall t \notin [-(\|x\|_\infty + E + 1), \|x\|_\infty + E + 1]^d \\
0 \leq \chi_x(t) &\leq 1, \quad \forall t \in \mathbb{R}^d.
\end{aligned}$$

As the dimension d and E are considered fixed here, there exists a constant C_1 such that, for all $x \in \mathbb{R}^d$, we have $\mathcal{M}(\chi_x) \leq C_1$ and $\mathcal{B}(\chi_x) \leq C_1 \max\{1, \|x\|_\infty\}$. Now, let $B := \max\{1, \|g\|_{L^\infty(\mathbb{R})}\}$. Next, by Proposition III.2 there exists a constant C_2 such that, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\mu_{x,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\sup_{y,z \in [-2B, 2B]} |\mu_{x,\varepsilon}(y, z) - yz| \leq \tfrac{\varepsilon}{4 s_x}, \tag{115}$$

and, for all $y \in \mathbb{R}$,

$$\mu_{x,\varepsilon}(0,y) = \mu_{x,\varepsilon}(y,0) = 0, \tag{116}$$

with $\mathcal{M}(\mu_{x,\varepsilon}) \leq C_2(\log(\varepsilon^{-1}) + \log(s_x))$ and $\mathcal{B}(\mu_{x,\varepsilon}) \leq C_2$. Again, with E fixed, there exists a constant C_3 such that $\mathcal{M}(\mu_{x,\varepsilon}) \leq C_3(\log(\varepsilon^{-1}) + \log(\|x\|_\infty + 1))$, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$. Let now

$$\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon} := \mu_{x,\varepsilon}(\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \chi_x) \quad \text{and} \quad \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon} := \mu_{x,\varepsilon}(\Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}, \chi_x). \tag{117}$$

Using Lemma II.5 along with (114), (115), and (116) establishes the existence of a polynomial π_4 such that for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$,

$$\|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \leq \tfrac{\varepsilon}{s_x} \tag{118}$$

and

$$\begin{aligned} \mathcal{M}(\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{M}(\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}) &\leq \pi_4(\log(\varepsilon^{-1}), \log(\|x\|_\infty), \log(\|\xi\|_\infty)), \\ \mathcal{B}(\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{B}(\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}) &\leq \pi_4(\varepsilon^{-1}, \|x\|_\infty, \|\xi\|_\infty). \end{aligned} \tag{119}$$

As $M_\xi T_x g$, $\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}$, and $\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}$ are supported in S_x for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, using (116), we get

$$\begin{aligned} &\|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^2(\mathbb{R}^d)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^2(\mathbb{R}^d)} \\ &= \|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^2(S_x)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^2(S_x)} \\ &\leq s_x\|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} + s_x\|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \leq \varepsilon. \end{aligned} \tag{120}$$

Consider now, for $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, the complex-valued network $\Gamma_{i,\varepsilon} \in \mathcal{NN}^{\mathbb{C}}_{\infty,\infty,d,1}$ given by

$$\Gamma_{i,\varepsilon} := (\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}, \Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon})$$

and note that, for $f \in L^2(\Omega, \mathbb{C})$,

$$\begin{aligned} \|f\|_{L^2(\Omega,\mathbb{C})} &= \left(\int_\Omega |f(t)|^2 \mathrm{d}t\right)^{\frac{1}{2}} = \left(\int_\Omega |\mathrm{Re}(f(t))|^2 + |\mathrm{Im}(f(t))|^2 \mathrm{d}t\right)^{\frac{1}{2}} = \left(\|\mathrm{Re}(f)\|^2_{L^2(\Omega)} + \|\mathrm{Im}(f)\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \\ &\leq \|\mathrm{Re}(f)\|_{L^2(\Omega)} + \|\mathrm{Im}(f)\|_{L^2(\Omega)}. \end{aligned}$$

Hence, (120) implies that, for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$,

$$\|\varphi_i - \Gamma_{i,\varepsilon}\|_{L^2(\mathbb{R}^d,\mathbb{C})} = \|M_{\xi(i)} T_{x(i)} g - (\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}, \Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon})\|_{L^2(\mathbb{R}^d,\mathbb{C})} \leq \varepsilon.$$

Finally, using (111) in (119), it follows that there exists a polynomial π_5 such that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, we have $\mathcal{M}(\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}), \mathcal{M}(\Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon}) \leq \pi_5(\log(\varepsilon^{-1}), \log(i))$ and $\mathcal{B}(\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}), \mathcal{B}(\Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon}) \leq \pi_5(\varepsilon^{-1}, i)$, which establishes effective representability of $\mathcal{G}(g, \alpha, \beta, \Omega)$ by neural networks. □

We proceed to establishing the central result of this section. To this end, we first recall that according to Theorem VII.6 neural networks provide optimal approximations for all function classes that are optimally approximated by

affine systems (generated by functions f that can be approximated well by neural networks). While this universality property is significant as it applies to all affine systems, it is perhaps not completely surprising as affine systems are generated by affine transformations and neural networks consist of concatenations of affine transformations and non-linearities. Gabor systems, on the other hand, are generated by the Weyl-Heisenberg group, as opposed to the affine group, and hence exhibit a fundamentally different mathematical structure. The next result shows that neural networks also provide optimal approximations for all function classes that are optimally approximated by Gabor systems (generated by functions g that can be approximated well by neural networks).

Theorem VIII.4. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, $g \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and let $\mathcal{G}(g, \alpha, \beta, \Omega)$ be the corresponding Gabor system with ordering as defined in* (99)*. Assume that either Ω is bounded or g is compactly supported. Further, suppose that there exists a polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying*

$$\|g - \Phi_{D,\varepsilon}\|_{L^\infty([-D,D]^d)} \leq \varepsilon \tag{121}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(\varepsilon^{-1}, D)$. Then, for all function classes $\mathcal{C} \subseteq L^2(\Omega)$,

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{G}(g, \alpha, \beta, \Omega)).$$

In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{G}(g, \alpha, \beta, \Omega)$ (in the sense of Definition V.5), then $\mathcal{C}$ is optimally representable by neural networks (in the sense of Definition V.10).

Proof. The first statement follows from Theorems VI.2 and VIII.3, the second is by Theorem V.9. □

We complete the program in this section by showing that the Gaussian function satisfies the conditions on the generator g in Theorem VIII.3. Gaussian functions are widely used generator functions for Gabor systems owing to their excellent time-frequency localization and their frame-theoretic optimality properties [14]. We hasten to add that the result below can be extended to sufficiently smooth generator functions g.

Lemma VIII.5. *Let $d \in \mathbb{N}$ and let $g \in L^2(\mathbb{R}^d)$ be given by*

$$g(x) := e^{-\|x\|_2^2}.$$

Then, there exist a constant $C > 0$ and a polynomial π such that for every $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|\Phi_\varepsilon - g\|_{L^\infty(\mathbb{R}^d)} \leq \varepsilon$$

with $\mathcal{M}(\Phi_\varepsilon) \leq \pi(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_\varepsilon) \leq \pi(\varepsilon^{-1})$.

Proof. Observe that g can be written as the composition of the functions $g_1 \colon \mathbb{R}^d \to \mathbb{R}_+$ and $g_2 \colon \mathbb{R}_+ \to \mathbb{R}$ given by

$$g_1(x) := \|x\|_2^2 = \sum_{i=1}^{d} x_i^2,$$

$$g_2(y) := e^{-y}.$$

By Proposition III.2 and Lemma II.7 there exist a constant $C_1 > 0$ and a polynomial π_1 (note that we consider d fixed) such that for every $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi^1_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\sup_{x\in[-D,D]^d} |\Psi^1_{D,\varepsilon} - \|x\|_2^2| \leq \tfrac{\varepsilon}{2} \tag{122}$$

with

$$\begin{aligned} \mathcal{M}(\Psi^1_{D,\varepsilon}) &\leq C_1(\log(\varepsilon^{-1}) + \log(\lceil D\rceil)), \\ \mathcal{B}(\Psi^1_{D,\varepsilon}) &\leq \pi_1(\lceil D\rceil). \end{aligned} \tag{123}$$

Moreover, as $|\frac{d^n}{dy^n} e^{-y}| = |e^{-y}| \leq 1$ for all $n \in \mathbb{N}$, $y \geq 0$, Lemma III.6 and Remark III.7 imply the existence of a constant $C_2 > 0$ and a polynomial π_2 such that for every $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi^2_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\sup_{y\in[0,dD^2]} |\Psi^2_{D,\varepsilon} - e^{-y}| \leq \tfrac{\varepsilon}{2} \tag{124}$$

with

$$\begin{aligned} \mathcal{M}(\Psi^2_{D,\varepsilon}) &\leq C_2\lceil D^2\rceil(\log(\varepsilon^{-1}))^2, \\ \mathcal{B}(\Psi^2_{D,\varepsilon}) &\leq \max\left\{\tfrac{dD^2}{2}, \max\{\tfrac{4}{dD^2}, \lceil\tfrac{dD^2}{2}\rceil\}\pi_2(\varepsilon^{-1})\right\}. \end{aligned} \tag{125}$$

By Lemma II.5 there exists for every $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying $\Phi_{D,\varepsilon}(x) = \Psi^2_{D,\varepsilon}(\Psi^1_{D,\varepsilon}(x))$, for $x \in \mathbb{R}^d$, with

$$\begin{aligned} \mathcal{M}(\Phi_{D,\varepsilon}) &\leq 2(\mathcal{M}(\Psi^1_{D,\varepsilon}) + \mathcal{M}(\Psi^2_{D,\varepsilon})), \\ \mathcal{B}(\Phi_{D,\varepsilon}) &= \max\{\mathcal{B}(\Psi^1_{D,\varepsilon}), \mathcal{B}(\Psi^2_{D,\varepsilon})\}. \end{aligned} \tag{126}$$

Setting $D_\varepsilon := \log(\varepsilon^{-1})$, it follows from (123), (125), and (126) that there exists a polynomial π_3 such that for all $\varepsilon \in (0, 1/2)$, $\mathcal{M}(\Phi_{D_\varepsilon,\varepsilon}) \leq \pi_3(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_{D_\varepsilon,\varepsilon}) \leq \pi_3(\varepsilon^{-1})$. As $|e^{-y}| \leq 1$ for all $y \geq 0$, combining (122) and (124) yields for all $\varepsilon \in (0, 1/2)$, $x \in [-D_\varepsilon, D_\varepsilon]^d$,

$$\begin{aligned} |g(x) - \Phi_{D_\varepsilon,\varepsilon}(x)| &= |e^{-\|x\|_2^2} - \Psi^2_{D_\varepsilon,\varepsilon}(\Psi^1_{D_\varepsilon,\varepsilon}(x))| \\ &\leq |e^{-\|x\|_2^2} - e^{-\Psi^1_{D_\varepsilon,\varepsilon}(x)}| + |e^{-\Psi^1_{D_\varepsilon,\varepsilon}(x)} - \Psi^2_{D_\varepsilon,\varepsilon}(\Psi^1_{D_\varepsilon,\varepsilon}(x))| \\ &\leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon. \end{aligned} \tag{127}$$

Based on (127), we can now use the same approach as in the proof of Theorem VIII.3 to construct networks Φ_ε which are supported on $[-D_\varepsilon, D_\varepsilon]^d$, where they approximate g to within error ε, and obey $\mathcal{M}(\Phi_\varepsilon) \leq \pi(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_\varepsilon) \leq \pi(\varepsilon^{-1})$. Together with $|g(x)| \leq \varepsilon$, for all $|x| \in \mathbb{R}^d \backslash [-D_\varepsilon, D_\varepsilon]^d$, this completes the proof. □

Remark VIII.6. *Note that Lemma VIII.5 establishes an approximation result that is even stronger than what is required in Theorem VIII.3. Specifically, we construct a neural network which achieves ε-approximation of the Gaussian function on all of $\mathbb{R}^d$ while exhibiting growth rates on $\mathcal{M}$ and $\mathcal{B}$ that are independent of the domain of approximation and satisfy the required dependency on the approximation error ε. This construction can be used to strengthen Theorem VIII.3 to apply to domain $\Omega = \mathbb{R}^d$ and generator functions of unbounded support, but sufficiently rapid decay.*

Remark VIII.7. *The real-valued counterpart of Gabor systems is known as local cosine bases [13], [14] and consists of cosine-modulated versions of a given generator function. The techniques developed in this section can be used to show that neural networks also provide optimal approximations for all function classes that are optimally approximated by local cosine bases (generated by functions that can be approximated well by neural networks). Gabor systems and local cosine bases provide optimal non-linear approximation of modulation spaces [64], [13]. The mathematical structure underlying modulation spaces is that of the Weyl-Heisenberg group [65] generated by the time-shift and the frequency-shift operator.*

IX. Oscillatory Textures and the Weierstrass Function

As mentioned in the introduction, there are functions that are known to be hard to approximate. Specifically, for the class of oscillatory textures as considered below and for the Weierstrass function, there are no known methods that achieve exponential accuracy, i.e., an approximation error that decays exponentially in the number of parameters employed in the approximant. We establish below that deep ReLU networks fill this gap.

Let us start by defining one-dimensional "oscillatory textures" according to [15].

Definition IX.1. *Let the sets $\mathcal{F}_{D,a}$, $D, a \in \mathbb{R}_+$, be given by*

$$\mathcal{F}_{D,a} = \{\cos(ag)h \colon g, h \in \mathcal{S}_D\}\,. \tag{128}$$

The efficient approximation of functions in $\mathcal{F}_{D,a}$ with a large represents a notoriously difficult problem due to the combination of the rapidly oscillating cosine term and the warping g. The best approximation results available in the literature [15] are based on wave-atom dictionaries* and yield low-order polynomial approximation rates. In what follows we show that finite-width deep networks drastically improve these results to exponential approximation rates.

Proposition IX.2. *There exist a constant $C > 0$ and a polynomial π such that for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, and $\varepsilon \in (0, 1/2)$, there is a network $\Gamma_{f,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying*

*To be precise, the results of [15] are concerned with the two-dimensional case, whereas we focus on the one-dimensional case. Note, however, that all our results can be readily extended to the multi-dimensional case.

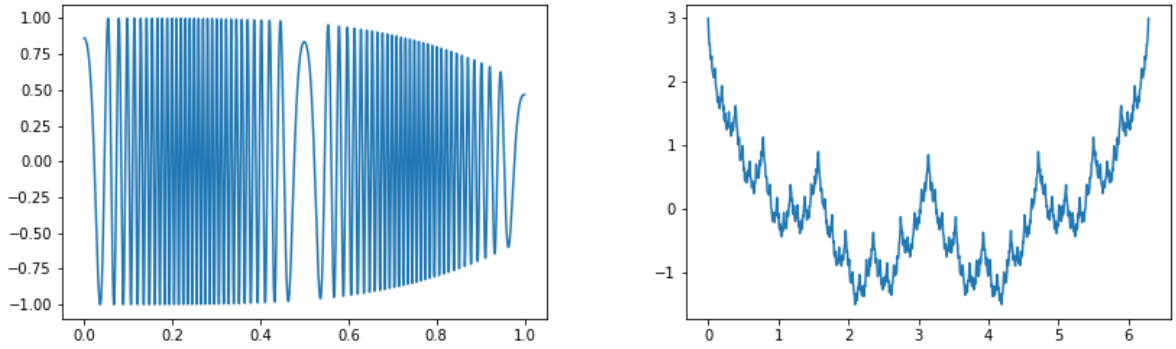


Fig. 3: Left: A function in $\mathcal{F}_{1,100}$. Right: The function $W_{\frac{1}{\sqrt{2}},2}$.

$$\|f - \Gamma_{f,\varepsilon}\|_{L^\infty([-D,D])} \leq \varepsilon \tag{129}$$

with $\mathcal{L}(\Gamma_{f,\varepsilon}) \leq C\lceil D\rceil(\log(\varepsilon^{-1}) + \log(\lceil a\rceil))^2$, $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 23$, *and* $\mathcal{B}(\Gamma_{f,\varepsilon}) \leq \max\{1/D, \lceil D\rceil\}\pi((\varepsilon/\lceil a\rceil)^{-1})$.

Proof. For all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, let $g_f, h_f \in \mathcal{S}_D$ be functions such that $f = \cos(ag_f)h_f$. Note that Lemma III.6 guarantees the existence of a constant $C_1 > 0$ and a polynomial π_1 such that for all $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there are networks $\Psi_{g_f,\varepsilon}, \Psi_{h_f,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\|\Psi_{g_f,\varepsilon} - g_f\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{12\lceil a\rceil}, \quad \|\Psi_{h_f,\varepsilon} - h_f\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{12\lceil a\rceil} \tag{130}$$

with $\mathcal{W}(\Psi_{g_f,\varepsilon}), \mathcal{W}(\Psi_{h_f,\varepsilon}) \leq 23$, $\mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon}) \leq C_1\lceil D\rceil(\log((\frac{\varepsilon}{12\lceil a\rceil})^{-1}))^2$, and $\mathcal{B}(\Psi_{g_f,\varepsilon}), \mathcal{B}(\Psi_{h_f,\varepsilon}) \leq \max\{1/D, \lceil D\rceil\}\pi_1((\frac{\varepsilon}{12\lceil a\rceil})^{-1})$. Theorem IV.1 now ensures the existence of a constant $C_2 > 0$ such that for all $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_{a,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\|\Phi_{a,D,\varepsilon} - \cos(a\,\cdot)\|_{L^\infty([-3/2,3/2])} \leq \tfrac{\varepsilon}{3} \tag{131}$$

with $\mathcal{W}(\Phi_{a,D,\varepsilon}) \leq 16$, $\mathcal{L}(\Phi_{a,D,\varepsilon}) \leq C_2((\log(1/\varepsilon))^2 + \log(\lceil 3a/2\rceil))$, and $\mathcal{B}(\Phi_{a,D,\varepsilon}) \leq C_2$. Further, thanks to Proposition III.2, there exists a constant $C_3 > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\mu_\varepsilon \in \mathcal{NN}_{\infty,\infty,2,1}$ satisfying

$$\sup_{x,y\in[-3/2,3/2]} |\mu_\varepsilon(x, y) - xy| \leq \tfrac{\varepsilon}{3} \tag{132}$$

with $\mathcal{W}(\mu_\varepsilon) \leq 12$, $\mathcal{L}(\mu_\varepsilon) \leq C_3 \log(\varepsilon^{-1})$, and $\mathcal{B}(\mu_\varepsilon) \leq 8$. By Lemma II.5 there exists a network Ψ^1 satisfying $\Psi^1(x) = \Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x))$, for all $x \in \mathbb{R}$, with $\mathcal{L}(\Psi^1) = \mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon})$ and $\mathcal{W}(\Psi^1) \leq 23$. Furthermore, by Lemma II.7 there exists a network Ψ^2 satisfying $\Psi^2(x) = (\Psi^1(x), \Psi_{h_f,\varepsilon}(x)) = (\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x))$ with $\mathcal{L}(\Psi^2) = \max\{\mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon})\}$, and $\mathcal{W}(\Psi^2) \leq 23$. Next, for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, $\varepsilon \in (0, 1/2)$, we define the network

$$\Gamma_{f,\varepsilon} := \mu_\varepsilon(\Psi^2) = \mu_\varepsilon(\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}), \Psi_{h_f,\varepsilon}). \tag{133}$$

Next, by (130), (131), and $\sup_{x\in\mathbb{R}} |\frac{d}{dx}\cos(ax)| = a$, we have, for all $x \in [-D, D]$,

$$\begin{aligned}|\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)) - \cos(ag_f(x))| &\leq |\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)) - \cos(a\Psi_{g_f,\varepsilon}(x))| \\ &\quad + |\cos(a\Psi_{g_f,\varepsilon}(x)) - \cos(ag_f(x))| \\ &\leq \tfrac{\varepsilon}{3} + a\tfrac{\varepsilon}{12\lceil a\rceil} \leq \tfrac{5\varepsilon}{12}.\end{aligned}$$

Combining this with (130), (132), and $\|\cos\|_{L^\infty([-D,D])}, \|f\|_{L^\infty([-D,D])} \leq 1$ yields for all $x \in [-D, D]$,

$$\begin{aligned}|\Gamma_{f,\varepsilon}(x) - f(x)| &= |\mu_\varepsilon(\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x)) - \cos(ag_f(x))h_f(x)| \\ &\leq |\mu_\varepsilon(\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x)) - \Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x))\Psi_{h_f,\varepsilon}(x)| \\ &\quad + |\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x))\Psi_{h_f,\varepsilon}(x) - \cos(ag_f(x))\Psi_{h_f,\varepsilon}(x)| \\ &\quad + |\cos(ag_f(x))\Psi_{h_f,\varepsilon}(x) - \cos(ag_f(x))h_f(x)| \\ &\leq \tfrac{\varepsilon}{3} + \tfrac{5\varepsilon}{12}\left(1 + \tfrac{\varepsilon}{12\lceil a\rceil}\right) + \tfrac{\varepsilon}{12\lceil a\rceil} \leq \varepsilon.\end{aligned}$$

By Lemma II.5 there exist a constant C_4 and a polynomial π_2 such that for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, $\varepsilon \in (0, 1/2)$, it holds that $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 23$,

$$\mathcal{L}(\Gamma_{f,\varepsilon}) \leq \mathcal{L}(\mu_\varepsilon) + \max\{\mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon})\} \leq C_4\lceil D\rceil((\log(\varepsilon^{-1}) + \log(\lceil a\rceil))^2,$$

and

$$\mathcal{B}(\Gamma_{f,\varepsilon}) \leq \max\{\mathcal{B}(\mu_\varepsilon), \mathcal{B}(\Phi_{a,D,\varepsilon}), \mathcal{B}(\Psi_{g_f,\varepsilon}), \mathcal{B}(\Psi_{h_f,\varepsilon})\} \leq \max\{1/D, \lceil D\rceil\}\pi_2((\varepsilon/\lceil a\rceil)^{-1}).$$

This completes the proof. □

Finally, we show how the Weierstrass function—a fractal function, which is continuous everywhere but differentiable nowhere—can be approximated with exponential accuracy by deep ReLU networks. Specifically, we consider

$$W_{p,a}(x) = \sum_{k=0}^{\infty} p^k \cos(a^k \pi x), \quad \text{for } p \in (0, 1/2),\ a \in \mathbb{R}_+, \text{ with } ap \geq 1. \tag{134}$$

Let $\alpha = -\frac{\log(p)}{\log(a)}$. It is well known [66] that $W_{p,a}$ possesses Hölder smoothness α which may be made arbitrarily small by suitable choice of a, see Figure 3 right. While classical approximation methods therefore achieve polynomial approximation rates only, it turns out that finite-width deep networks yield exponential approximation rates. This is formalized as follows.

Proposition IX.3. *There exists a constant $C > 0$ such that, for all $\varepsilon, p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, there is a network $\Psi_{p,a,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying*

$$\|\Psi_{p,a,D,\varepsilon} - W_{p,a}\|_{L^\infty([-D,D])} \leq \varepsilon \tag{135}$$

with

$$\mathcal{L}(\Psi_{p,a,D,\varepsilon}) \leq C((\log(1/\varepsilon))^3 + (\log(1/\varepsilon))^2 \log(\lceil a \rceil) + \log(1/\varepsilon)\log(\lceil D \rceil)),$$

$\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 20$*, and* $\mathcal{B}(\Psi_{p,a,D,\varepsilon}) \leq C$.

Proof. For every $N \in \mathbb{N}$, $p \in (0,1/2)$, $a \in \mathbb{R}_+$, $x \in \mathbb{R}$, let $S_{N,p,a}(x) = \sum_{k=0}^{N} p^k \cos(a^k \pi x)$ and note that

$$|S_{N,p,a}(x) - W_{p,a}(x)| \leq \sum_{k=N+1}^{\infty} |p^k \cos(a^k \pi x)| \leq \sum_{k=N+1}^{\infty} p^k = \tfrac{1}{1-p} - \tfrac{1-p^{N+1}}{1-p} \leq 2^{-N}. \tag{136}$$

Let $N_\varepsilon := \lceil \log(2/\varepsilon) \rceil$, $\varepsilon \in (0,1/2)$. Next note that Theorem IV.1 ensures the existence of a constant $C_1 > 0$ such that for all $D, a \in \mathbb{R}_+$, $k \in \mathbb{N}_0$, $\varepsilon \in (0,1/2)$, there is a network $\phi_{a^k,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\|\phi_{a^k,D,\varepsilon} - \cos(a^k \pi \cdot)\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{4} \tag{137}$$

with $\mathcal{W}(\phi_{a^k,D,\varepsilon}) \leq 16$, $\mathcal{L}(\phi_{a^k,D,\varepsilon}) \leq C_1((\log(\varepsilon^{-1}))^2 + \log(\lceil a^k \pi D \rceil))$, and $\mathcal{B}(\phi_{a^k,D,\varepsilon}) \leq C_1$. Thanks to $x = \rho(x) - \rho(-x)$, there exists a neural network $\tau \in \mathcal{NN}_{2,4,1,1}$ satisfying $\tau(x) = x$, for all $x \in \mathbb{R}$. Applying Lemma II.5 to τ and $\phi_{a^k,D,\varepsilon}, k \geq 0$, we obtain that for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $k \in \mathbb{N}_0$, $\varepsilon \in (0,1/2)$, there exist networks

$$\psi^{p,a,0}_{D,\varepsilon}(x) = \begin{pmatrix} x \\ p^0 \phi_{a^0,D,\varepsilon}(x) \\ 0 \end{pmatrix} \quad \text{and} \quad \psi^{p,a,k}_{D,\varepsilon}(x_1,x_2,x_3) = \begin{pmatrix} x_1 \\ p^k \phi_{a^k,D,\varepsilon}(x_2) \\ x_3 \end{pmatrix}, \; k > 0. \tag{138}$$

Now let $A \in \mathbb{R}^{3\times 3}$ be such that $A(y_1,y_2,y_3)^T = (y_1, y_1, y_2+y_3)^T$, for all $y \in \mathbb{R}^3$. Applying now, for $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, Lemma II.5, we obtain that there exist a network $\Psi_{p,a,D,\varepsilon}$ given by

$$\Psi_{p,a,D,\varepsilon}(x) := \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \psi^{p,a,N_\varepsilon}_{D,\varepsilon}(A\psi^{p,a,N_\varepsilon-1}_{D,\varepsilon}(\ldots(A\psi^{p,a,0}_{D,\varepsilon}(x)))). \tag{139}$$

With (137) we get, for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, $x \in [-D,D]$,

$$\begin{aligned} |\Psi_{p,a,D,\varepsilon}(x) - S_{N_\varepsilon,p,a}(x)| &= \left| \sum_{k=0}^{N_\varepsilon} p^k \phi_{a^k,D,\varepsilon}(x) - \sum_{k=0}^{N_\varepsilon} p^k \cos(a^k \pi x) \right| \\ &\leq \sum_{k=0}^{N_\varepsilon} p^k |\phi_{a^k,D,\varepsilon}(x) - \cos(a^k \pi x)| \leq \tfrac{\varepsilon}{4} \sum_{k=0}^{N_\varepsilon} 2^{-k} \leq \tfrac{\varepsilon}{2}. \end{aligned}$$

Combining this with (136) establishes that for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, $x \in [-D,D]$,

$$|\Psi_{p,a,D,\varepsilon}(x) - W_{p,a}(x)| \leq 2^{-\lceil \log(\frac{2}{\varepsilon}) \rceil} + \tfrac{\varepsilon}{2} \leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.$$

The proof is now completed by noting that there exists a constant C_2 such that for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, we have $\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 20$,

$$\begin{aligned} \mathcal{L}(\Psi_{p,a,D,\varepsilon}) &\leq \sum_{k=0}^{N_\varepsilon} \mathcal{L}(\phi_{a^k,D,\varepsilon}) \leq (N_\varepsilon + 1) C_1((\log(\varepsilon^{-1}))^2 + \log(\lceil a^{N_\varepsilon} \pi D \rceil)) \\ &\leq C_2((\log(\varepsilon^{-1}))^3 + (\log(\varepsilon^{-1}))^2 \log(\lceil a \rceil) + \log(\varepsilon^{-1}) \log(\lceil D \rceil)), \end{aligned}$$

and

$$\mathcal{B}(\Psi_{p,a,D,\varepsilon}) = \max_{k\in\{0,\dots,N_\varepsilon\}} \mathcal{B}(\phi_{a^k,D,\varepsilon}) \leq C_1.$$

□

We finally note that the restriction $p \in (0, 1/2)$ was made for simplicity of exposition and can be relaxed to $p \in (0, r)$ with $r \in (0, 1)$. For $p \to 1$ the convergence of the sum defining $W_{p,a}$ can become arbitrarily slow, leading to issues that are not specific to the approximation of $W_{p,a}$ by neural networks.

X. Impossibility results for finite-depth networks

This section makes a formal case for deep networks by establishing that, for non-constant periodic functions, finite-width deep networks require asymptotically—in the function's "highest frequency"—smaller connectivity than finite-depth wide networks. This statement is then extended to sufficiently smooth non-periodic functions, thereby formalizing the benefit of deep networks over shallow networks for the approximation of a broad class of functions.

We start with preparatory material taken from [19].

Definition X.1 ([19])**.** *Let $k \in \mathbb{N}$. A function $f : \mathbb{R} \to \mathbb{R}$ is called k-sawtooth if it is piecewise linear with no more than k pieces, i.e., its domain $\mathbb{R}$ can be partitioned into k intervals such that f is linear on each interval.*

Lemma X.2 ([19])**.** *Every $\Phi \in \mathcal{NN}_{\infty,\infty,1,1}$ is $(2\mathcal{W}(\Phi))^{\mathcal{L}(\Phi)}$-sawtooth.*

Definition X.3. *For a u-periodic function $f \in C(\mathbb{R})$, we define*

$$\xi(f) := \sup_{\delta\in[0,u)} \inf_{c,d\in\mathbb{R}} \|f(x) - (cx + d)\|_{L^\infty([\delta,\delta+u])}.$$

The quantity $\xi(f)$ measures the error incurred by the best linear approximation of f on any segment of length equal to the period of f; $\xi(f)$ can hence be interpreted as quantifying the non-linearity of f. The next result states that finite-depth networks with width scaling poly-logarithmically in the "highest frequency" of the periodic function to be approximated can not achieve arbitrarily small approximation error.

Proposition X.4. *Let $f \in C(\mathbb{R})$ be a non-constant u-periodic function, $L \in \mathbb{N}$, and π a polynomial. Then, there exists an $a \in \mathbb{N}$ such that for every network $\Phi \in \mathcal{NN}_{L,\infty,1,1}$ with $\mathcal{W}(\Phi) \leq \pi(\log(a))$,*

$$\|f(a\,\cdot) - \Phi\|_{L^\infty([0,u])} \geq \xi(f) > 0.$$

Proof. First note that there exists an even $a \in \mathbb{N}$ such that $a/2 > (2\pi(\log(a)))^L$. Lemma X.2 now implies that every network $\Phi \in \mathcal{NN}_{L,\infty,1,1}$ with $\mathcal{W}(\Phi) \leq \pi(\log(a))$ is $(2\pi(\log(a)))^L$-sawtooth and therefore consists of at most $a/2$ different linear pieces. Hence, there exists an interval $[u_1, u_2] \subseteq [0, u]$ with $u_2 - u_1 \geq (2u/a)$ on which Φ is linear. Since $u_2 - u_1 \geq (2u/a)$ the interval supports two full periods of $f(a\,\cdot)$ and we can therefore conclude that

$$\|f(a\,\cdot\,)-\Phi\|_{L^\infty([0,u])} \geq \|f(a\,\cdot\,)-\Phi\|_{L^\infty([u_1,u_2])} \geq \inf_{c,d\in\mathbb{R}} \|f(x)-(cx+d)\|_{L^\infty([0,2u])}$$

$$\geq \sup_{\delta\in[0,u)} \inf_{c,d\in\mathbb{R}} \|f(x)-(cx+d)\|_{L^\infty([\delta,u+\delta])} = \xi(f).$$

Finally note that $\xi(f)>0$ as $\xi(f)=0$ for u-periodic $f\in C(\mathbb{R})$ would entail that f is constant thereby contradicting the assumption. □

Application of Proposition X.4 to $f(x)=\cos(x)$ shows that finite-depth networks, owing to $\xi(\cos)>0$, require faster than poly-logarithmic growth of connectivity in a to approximate $x\mapsto\cos(ax)$ with arbitrarily small error, whereas finite-width networks, thanks to Theorem IV.1, can accomplish this with poly-logarithmic growth. The next result, taken from [67], allows us to extend this conclusion to non-periodic functions that are sufficiently smooth.

Theorem X.5 ([67])**.** *Let $f\in C^3([a,b])$ and consider a piecewise linear approximation of f on $[a,b]$ that is accurate to within ε in the $L^\infty([a,b])$-norm. The minimal number of linear pieces required to accomplish this scales according to*

$$s(\varepsilon)\sim\frac{c}{\sqrt{\varepsilon}},\ \varepsilon\to 0,\ \ \textit{where}\ \ c=\frac{1}{4}\int_a^b\sqrt{|f''(x)|}dx.$$

Combining this with Lemma X.2 yields the following result on depth-width tradeoff for three-times continuously differentiable functions.

Theorem X.6. *Let $f\in C^3([a,b])$ with $\int_a^b\sqrt{|f''(x)|}dx>0$, $L\in\mathbb{N}$, and π a polynomial. Then, there exists $\varepsilon>0$ such that for every network $\Phi\in\mathcal{NN}_{L,\infty,1,1}$ with $\mathcal{W}(\Phi)\leq\pi(\log(\varepsilon^{-1}))$,*

$$\|f-\Phi\|_{L^\infty([a,b])}>\varepsilon.$$

Proof. The proof will be effected by contradiction. Assume that for every $\varepsilon>0$ there exists a network $\Phi_\varepsilon\in\mathcal{NN}_{L,\infty,1,1}$ with $\mathcal{W}(\Phi_\varepsilon)\leq\pi(\log(\varepsilon^{-1}))$ and $\|f-\Phi_\varepsilon\|_{L^\infty([a,b])}\leq\varepsilon$. By Lemma X.2 and Definition X.1 every ReLU network realizes a piecewise linear function. Application of Theorem X.5 hence allows us to conclude that there exists a constant C such that, for all $\varepsilon>0$, the network Φ_ε must have at least $C\varepsilon^{-\frac{1}{2}}$ different linear pieces. This, however, leads to a contradiction as, by Lemma X.2, Φ_ε is $(2\pi(\log(\varepsilon^{-1})))^L$-sawtooth and $\tilde{\pi}(\log(\varepsilon^{-1}))\in o(\varepsilon^{-1/2})$, $\varepsilon\to 0$, for every polynomial $\tilde{\pi}$. □

This shows that any function that is at least three times continuously differentiable cannot be approximated by finite-depth networks with connectivity scaling poly-logarithmically in the inverse of the approximation error. Our results in Sections III and IV establish that, in contrast, this *is* possible for various interesting types of smooth functions such as polynomials and sinusoidal functions. Further results on the limitations of finite-depth networks akin to the statement in Theorem X.6 can be found in [17].

APPENDIX

The following result shows how to trade off the size of the weights in deep ReLU networks for depth. Here $\|.\|_0$ denotes the number of non-zero elements in a vector/matrix.

Proposition A.1. *Let $L, M, d, N_L \in \mathbb{N}$, $\Phi \in \mathcal{NN}_{L,M,d,N_L}$ with weights bounded (in absolute value) by B. Then, there exists a network Ψ satisfying*

1) $\Psi(x) = \Phi(x)$, *for all* $x \in \mathbb{R}^d$,
2) $\mathcal{W}(\Psi) = 2\mathcal{W}(\Phi)(\mathcal{W}(\Phi) + 1)$,
3) $\mathcal{M}(\Psi) \leq 2M(\lfloor \log(B) \rfloor + 2) + 1$,
4) $\mathcal{L}(\Psi) \leq (\lfloor \log(B) \rfloor + 3)L$, *and*
5) *the weights of Ψ are bounded (in absolute value) by* 2.

Proof. We start with the observation that for all $a \in \mathbb{R}$, there exists a network $\nu_a \in \mathcal{NN}_{\lfloor \log(|a|) \rfloor + 2, 2(\lfloor \log(|a|) \rfloor + 2), 1, 1}$ with $\mathcal{W}(\nu_a) = 2$, weights bounded (in absolute value) by 2 and satisfying $\nu_a(x) = ax$, for all $x \in \mathbb{R}$. For $|a| \leq 2$ this network is given by

$$\nu_a(x) = \begin{pmatrix} a & -a \end{pmatrix} \rho \left(\begin{pmatrix} 1 \\ -1 \end{pmatrix} x \right).$$

and for $|a| > 2$, we have

$$\nu_a(x) = \tfrac{a}{2^{\lfloor \log(|a|) \rfloor}} \begin{pmatrix} 1 & -1 \end{pmatrix} \rho \left(\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \rho \left(\ldots \rho \left(\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \rho \left(\begin{pmatrix} 1 \\ -1 \end{pmatrix} x \right) \right) \right) \right).$$

This implies that for every matrix $A = (a_{i,j}) \in \mathbb{R}^{N \times N'}$ with $\|A\|_0 = H$ and $\max_{i,j} |a_{i,j}| \leq |a|$, there exists a network $\alpha_A \in \mathcal{NN}_{\lfloor \log(|a|) \rfloor + 2, 2H(\lfloor \log(|a|) \rfloor + 2), N', NN'}$ with $\mathcal{W}(\alpha_A) = 2NN'$, weights bounded (in absolute value) by 2 and satisfying, for all $x \in \mathbb{R}^{N'}$,

$$\begin{aligned} \alpha_A(x) &= (\nu_{a_{1,1}}(x_1), \nu_{a_{1,2}}(x_2), \ldots, \nu_{a_{1,N'}}(x_{N'}), \ldots, \nu_{a_{N,1}}(x_1), \nu_{a_{N,2}}(x_2), \ldots, \nu_{a_{N,N'}}(x_{N'}))^T \\ &= (a_{1,1}x_1, a_{1,2}x_2, \ldots, a_{1,N'}x_{N'}, \ldots, a_{N,1}x_1, a_{N,2}x_2, \ldots, a_{N,N'}x_{N'})^T. \end{aligned}$$

Similarly, observe that for all $b = (b_i) \in \mathbb{R}^N$ with $\|b\|_0 = H$ and $\max_i |b_i| \leq |a|$, there exists a network $\beta_b \in \mathcal{NN}_{\lfloor \log(|a|) \rfloor + 2, 2H(\lfloor \log(|a|) \rfloor + 2), N', N}$ with $\mathcal{W}(\beta_b) = 2N$, weights bounded (in absolute value) by 2 and satisfying $\beta_b(x) = b$, for all $x \in \mathbb{R}^{N'}$. Noting that $(Ax)_i = \sum_{j=1}^{N'} a_{i,j}x_j$, we can therefore conclude that for every $A = (a_{i,j}) \in \mathbb{R}^{N \times N'}$ and $b = (b_i) \in \mathbb{R}^N$ with $\|A\|_0 + \|b\|_0 = H$ and $\max_{i,j} |a_{i,j}|, \max_i |b_i| \leq |a|$, there exist networks $\beta_{A,b} \in \mathcal{NN}_{\lfloor \log(|a|) \rfloor + 3, 2H(\lfloor \log(|a|) \rfloor + 2) + 1, N', N}$ and $\beta'_{A,b} \in \mathcal{NN}_{\lfloor \log(|a|) \rfloor + 2, 2H(\lfloor \log(|a|) \rfloor + 2), N', N}$ with $\mathcal{W}(\beta_{A,b}), \mathcal{W}(\beta'_{A,b}) = 2N(N' + 1)$, weights bounded (in absolute value) by 2 and satisfying, for all $x \in \mathbb{R}^{N'}$,

$$\beta_{A,b}(x) = \rho(Ax + b), \quad \beta'_{A,b}(x) = Ax + b.$$

We now apply this idea to construct a network equivalent to Φ in terms of the function realized but with weights bounded (in absolute value) by 2. First, recall that

$$\Phi(x) = W_L\left(\rho\left(W_{L-1}\left(\ldots\left(W_1(x)\right)\right)\right)\right),$$

where $W_\ell(x) = A_\ell x + b_\ell$ with $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, $b_\ell \in \mathbb{R}^{N_\ell}$, $\ell = 1, 2, \ldots, L$. Let now, for $\ell \in \{1, 2, \ldots, L\}$, the network γ_ℓ be given by

$$\gamma_\ell = \begin{cases} \rho(W_\ell) & : \ell < L,\ \max_{i,j} |a_{i,j}^{(\ell)}|, \max_i |b_i^{(\ell)}| \leq 2 \\ \beta_{A_\ell, b_\ell} & : \ell < L,\ \max_{i,j} |a_{i,j}^{(\ell)}|, \max_i |b_i^{(\ell)}| > 2 \\ W_L & : \ell = L,\ \max_{i,j} |a_{i,j}^{(\ell)}|, \max_i |b_i^{(\ell)}| \leq 2 \\ \beta'_{A_L, b_L} & : \ell = L,\ \max_{i,j} |a_{i,j}^{(\ell)}|, \max_i |b_i^{(\ell)}| > 2, \end{cases} \tag{140}$$

and set

$$\Psi(x) = \gamma_L(\gamma_{L-1}(\ldots\gamma_2(\gamma_1(x)))).$$

By construction $\Psi(x) = \Phi(x)$, for all $x \in \mathbb{R}^d$, $\mathcal{M}(\Psi) \leq 2M(\lfloor \log(B) \rfloor + 2) + 1$, $\mathcal{W}(\Psi) = 2\mathcal{W}(\Phi)(\mathcal{W}(\Phi) + 1)$, and the weights of Ψ are bounded (in absolute value) by 2. Further, note that the depth of each of the subnetworks $\beta_{A_\ell, b_\ell}, \ell \in \{1, 2, \ldots, L\}$, is bounded by $\lfloor \log(B) \rfloor + 3$, which implies $\mathcal{L}(\Psi) \leq (\lfloor \log(B) \rfloor + 3)L$. □

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[2] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Müller, E. Säckinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," *International Conference on Artificial Neural Networks*, pp. 53–60, 1995.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016. [Online]. Available: http://www.nature.com/nature/journal/v529/n7587/abs/nature16961.html#supplementary-information

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14539

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986. [Online]. Available: http://dx.doi.org/10.1038/323533a0

[8] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[9] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989. [Online]. Available: http://dx.doi.org/10.1007/BF02551274

[10] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251 – 257, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/089360809190009T

[11] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[12] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM Journal on Mathematics of Data Science*, 2019, to appear.

[13] K. Gröchenig and S. Samarah, "Nonlinear approximation with local Fourier bases," *Constructive Approximation*, vol. 16, no. 3, pp. 317–331, Jul 2000.

[14] K. Gröchenig, *Foundations of time-frequency analysis*. Springer Science & Business Media, 2013.

[15] L. Demanet and L. Ying, "Wave atoms and sparsity of oscillatory patterns," *Appl. Comput. Harmon. Anal.*, vol. 23, no. 3, pp. 368–387, 2007.

[16] C. Fefferman, "Reconstructing a neural net from its output," *Revista Matemática Iberoamericana*, vol. 10, no. 3, pp. 507–555, 1994.

[17] P. Petersen and F. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep ReLU neural networks," *Neural Networks*, vol. 108, pp. 296–330, Sep. 2018.

[18] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural Networks*, vol. 94, pp. 103–114, 2017.

[19] M. Telgarsky, "Representation benefits of deep feedforward networks," *arXiv:1509.08101*, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[21] C. Schwab and J. Zech, "Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ," *Analysis and Applications (Singapore)*, 2018.

[22] M. H. Stone, "The generalized Weierstrass approximation theorem," *Mathematics Magazine*, vol. 21, pp. 167–184, 1948.

[23] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" *International Conference on Learning Representations*, 2017.

[24] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, 1998.

[25] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, vol. 14, no. 1, pp. 115–133, 1994. [Online]. Available: http://dx.doi.org/10.1007/BF00993164

[26] C. K. Chui, X. Li, and H. N. Mhaskar, "Neural networks for localized approximation," *Math. Comp.*, vol. 63, no. 208, pp. 607–623, 1994. [Online]. Available: http://dx.doi.org/10.2307/2153285

[27] R. DeVore, K. Oskolkov, and P. Petrushev, "Approximation by feed-forward neural networks," *Ann. Numer. Math.*, vol. 4, pp. 261–287, 1996.

[28] E. J. Candès, "Ridgelets: Theory and Applications," 1998, Ph.D. thesis, Stanford University.

[29] H. N. Mhaskar, "Neural networks for optimal approximation of smooth and analytic functions," *Neural Comput.*, vol. 8, no. 1, pp. 164–177, 1996.

[30] H. Mhaskar and C. Micchelli, "Degree of approximation by neural and translation networks with a single hidden layer," *Adv. Appl. Math.*, vol. 16, no. 2, pp. 151–183, 1995.

[31] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[32] H. N. Mhaskar, "Approximation properties of a multilayered feedforward artificial neural network," *Advances in Computational Mathematics*, vol. 1, no. 1, pp. 61–80, Feb 1993. [Online]. Available: https://doi.org/10.1007/BF02070821

[33] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, 1989. [Online]. Available: //www.sciencedirect.com/science/article/pii/0893608089900038

[34] T. Nguyen-Thien and T. Tran-Cong, "Approximation of functions and their derivatives: A neural network implementation with applications," *Appl. Math. Model.*, vol. 23, no. 9, pp. 687–704, 1999. [Online]. Available: //www.sciencedirect.com/science/article/pii/S0307904X99000062

[35] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, 2016, pp. 907–940.

[36] H. N. Mhaskar and T. Poggio, "Deep vs. shallow networks: An approximation theory perspective," *Analysis and Applications*, vol. 14, no. 6, pp. 829–848, 2016. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0219530516400042

[37] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Proceedings of the 29th Conference on Learning Theory*, vol. 49, 2016, pp. 698–728.

[38] N. Cohen and A. Shashua, "Convolutional rectifier networks as generalized tensor decompositions," in *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 955–963.

[39] P. Grohs, F. Hornung, A. Jentzen, and P. von Wurstemberger, "A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations," *arXiv e-prints*, p. arXiv:1809.02362, Sep. 2018.

[40] J. Berner, P. Grohs, and A. Jentzen, "Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations," *arXiv e-prints*, p. arXiv:1809.03062, Sep. 2018.

[41] C. Beck, S. Becker, P. Grohs, N. Jaafari, and A. Jentzen, "Solving stochastic differential equations and Kolmogorov equations by means of deep learning," *arXiv e-prints*, p. arXiv:1806.00421, Jun. 2018.

[42] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab, "DNN expression rate analysis of high-dimensional PDEs: Application to option pricing," *arXiv preprint arXiv:1809.07669*, 2018.

[43] S. Ellacott, "Aspects of the numerical analysis of neural networks," *Acta Numer.*, vol. 3, pp. 145–202, 1994.

[44] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numer.*, vol. 8, pp. 143–195, 1999.

[45] U. Shaham, A. Cloninger, and R. R. Coifman, "Provable approximation properties for deep neural networks," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 3, pp. 537–557, May 2018. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1509.html#ShahamCC15

[46] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. Springer, 1993.

[47] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[48] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Appl. Comput. Harmon. Anal.*, vol. 1, no. 1, pp. 100 – 115, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1063520383710080

[49] P. Grohs, "Optimally sparse data representations," in *Harmonic and Applied Analysis*. Springer, 2015, pp. 199–248.

[50] E. Ott, *Chaos in Dynamical Systems*. Cambridge Univ. Press, 2002.

[51] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, "Tree approximation and optimal encoding," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 2, pp. 192–226, 2001.

[52] D. L. Donoho, "Sparse components of images and optimal atomic decompositions," *Constr. Approx.*, vol. 17, no. 3, pp. 353–382, 2001. [Online]. Available: http://dx.doi.org/10.1007/s003650010032

[53] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer, "Cartoon approximation with α-curvelets," *J. Fourier Anal. Appl.*, vol. 22, no. 6, pp. 1235–1293, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00041-015-9446-6

[54] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer, "α-molecules," *Appl. Comput. Harmon. Anal.*, vol. 41, no. 1, pp. 297–336, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.acha.2015.10.009

[55] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[56] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities," *Comm. Pure Appl. Math.*, vol. 57, pp. 219–266, 2002.

[57] K. Guo, G. Kutyniok, and D. Labate, "Sparse multidimensional representations using anisotropic dilation and shear operators," in *Wavelets and Splines (Athens, GA, 2005)*. Nashboro Press, Nashville, TN, 2006, pp. 189–201.

[58] P. Grohs and G. Kutyniok, "Parabolic molecules," *Found. Comput. Math.*, vol. 14, pp. 299–337, 2014.

[59] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[60] M. Unser, “Ten good reasons for using spline wavelets,” *Wavelet Applications in Signal and Image Processing V*, vol. 3169, pp. 422–431, 1997.

[61] C. K. Chui and J.-Z. Wang, “On compactly supported spline wavelets and a duality principle,” *Transactions of the American Mathematical Society*, 1992.

[62] G. B. Folland, *Harmonic Analysis in Phase Space. (AM-122)*. Princeton University Press, 1989. [Online]. Available: http://www.jstor.org/stable/j.ctt1b9rzs2

[63] C. L. Fefferman, “The uncertainty principle,” *Bull. Amer. Math. Soc. (N.S.)*, vol. 9, no. 2, pp. 129–206, 1983. [Online]. Available: https://doi.org/10.1090/S0273-0979-1983-15154-6

[64] H. Feichtinger, “On a new Segal algebra,” vol. 92, pp. 269–289, 01 1981.

[65] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2013.

[66] A. Zygmund, *Trigonometric series*. Cambridge University Press, 2002.

[67] C. Frenzen, T. Sasao, and J. T. Butler, “On the number of segments needed in a piecewise linear approximation,” *Journal of Computational and Applied Mathematics*, vol. 234, no. 2, pp. 437 – 446, 2010.