Deep Neural Network Approximation Theory

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

Abstract

This paper develops fundamental limits of deep neural network learning by characterizing what is possible if no constraints are imposed on the learning algorithm and on the amount of training data. Concretely, we consider Kolmogorov-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 establishes that deep networks are Kolmogorov-optimal approximants for markedly different function classes, such as unit balls in Besov spaces and modulation spaces. In addition, deep networks provide exponential approximation accuracy—i.e., the approximation error decays exponentially in the number of nonzero weights in the network—of the multiplication operation, polynomials, sinusoidal functions, and certain smooth functions. Moreover, this holds true even for one-dimensional oscillatory textures and the Weierstrass function—a fractal function, neither of which has previously known methods achieving exponential approximation accuracy. 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 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 that 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 are imposed on the learning algorithm and on the amount of training data [8].

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

D. Perekrestenko and H. Bölcskei are with the Chair for Mathematical Information Science, ETH Zurich, Switzerland (e-mail: pdmytro@mins.ee.ethz.ch, hboelcskei@ethz.ch).

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 was supported through the FWF projects P 30148 and I 3403 as well as the WWTF project ICT19-041.

The theory of function approximation through neural networks has a long history dating back to the work by McCulloch and Pitts [9] and the seminal paper by Kolmogorov [10], who showed, when interpreted in neural network parlance, that any continuous function of n variables can be represented exactly through a 2-layer neural network of width $2n+1$. However, the nonlinearities in Kolmogorov's neural network are highly nonsmooth and the outer nonlinearities, i.e., those in the output layer, depend on the function to be represented. In modern neural network theory, one is usually interested in networks with nonlinearities that are independent of the function to be realized and exhibit, in addition, certain smoothness properties. Significant progress in understanding the approximation capabilities of such networks has been made in [11], [12], where it was shown 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. 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-Donoho rate-distortion theory for neural networks [13] 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 theory is based on a framework for quantifying the fundamental limits of nonlinear approximation through dictionaries as introduced by Donoho [14], [15].

The purpose of this paper is to provide a comprehensive, principled, and self-contained introduction to Kolmogorov-Donoho 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 research in the field. Part of this paper is based on [13], 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 Kolmogorov-Donoho optimal approximants for vastly different function classes such as unit balls in Besov spaces [16] and modulation spaces [17]. 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 nonzero weights in the network—for vastly different functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures [18] and the Weierstrass function—a fractal function, neither of which has known methods achieving exponential approximation accuracy.

While we consider networks based on the ReLU[1] activation function throughout, certain parts of our theory carry over to strongly sigmoidal activation functions of order $k \geq 2$ as defined in [13]. For the sake of conciseness, we refrain from providing these extensions.

Outline of the paper. In Section II, we introduce notation, formally define neural networks, and record basic elements needed in the neural network constructions throughout the paper. Section III presents an algebra of function approximation by neural networks. In Section IV, we develop the Kolmogorov-Donoho rate-distortion framework that will allow us to characterize the fundamental limits of deep neural network learning of function classes. This theory is based on the concept of metric entropy, which is introduced and reviewed starting from first principles. Section V then puts the Kolmogorov-Donoho framework to work in the context of nonlinear function approximation with dictionaries. This discussion serves as a basis for the development of the concept of best M-weight approximation in neural networks presented in Section VI. We proceed, in Section VII, with the development of a method—termed the transference principle—for transferring results on function approximation through dictionaries to results on approximation by neural networks. The purpose of Section VIII is to demonstrate that function classes that are optimally approximated by affine dictionaries (e.g., wavelets), are optimally approximated by neural networks as well. In Section IX, we show that this optimality transfer extends to function classes that are optimally approximated by Weyl-Heisenberg dictionaries. Section X demonstrates that neural networks can improve the best-known approximation rates for two example functions, namely oscillatory textures and the Weierstrass function, from polynomial to exponential. The final Section XI makes a formal case for depth in neural network approximation by establishing a provable benefit of deep networks over shallow networks in the approximation of sufficiently smooth functions. The Appendices collect ancillary technical results.

Notation. For a function $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ and a set $\Omega \subseteq \mathbb{R}^d$, we define $\|f\|_{L^\infty(\Omega)} := \sup\{|f(x)| : x \in \Omega\}$. $L^p(\mathbb{R}^d)$ and $L^p(\mathbb{R}^d, \mathbb{C})$ denote the space of real-valued, respectively complex-valued, L^p-functions. When dealing with the approximation error for simple functions such as, e.g., $(x, y) \mapsto xy$, we will for brevity of exposition and with slight abuse of notation, make the arguments inside the norm explicit according to $\|f(x, y) - xy\|_{L^p(\Omega)}$. 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 $\mathbb{I}_n$. $\log$ stands for the logarithm to base 2. For a set $X \in \mathbb{R}^d$, we write $|X|$ for its Lebesgue measure. Constants like C are understood to be allowed to take on different values in different uses.

[1]ReLU stands for the Rectified Linear Unit nonlinearity defined as $x \mapsto \max\{0, x\}$.

II. SETUP AND BASIC RELU CALCULUS

This section defines neural networks, introduces the basic setup as well as further notation, and lists basic elements needed in the neural network constructions considered throughout, namely compositions and linear combinations of neural networks. 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 $L \in \mathbb{N}$ and $N_0, N_1, \dots, N_L \in \mathbb{N}$. A* ReLU neural network Φ *is a map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by*

$$\Phi = \begin{cases} W_1, & L = 1 \\ W_2 \circ \rho \circ W_1, & L = 2 \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1, & L \geq 3 \end{cases}, \tag{1}$$

where, for $\ell \in \{1, 2, \dots, L\}$, $W_\ell \colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(x) := A_\ell x + b_\ell$ are the associated affine transformations with matrices $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and (bias) vectors $b_\ell \in \mathbb{R}^{N_\ell}$, and the ReLU activation function $\rho \colon \mathbb{R} \to \mathbb{R}$, $\rho(x) := \max(0, x)$ acts component-wise, i.e., $\rho(x_1, \dots, x_N) := (\rho(x_1), \dots, \rho(x_N))$. We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$. Moreover, we define the following quantities related to the notion of size of the ReLU network Φ:

- *the* connectivity $\mathcal{M}(\Phi)$ *is the total number of nonzero entries in the matrices A_ℓ, $\ell \in \{1, 2, \dots, L\}$, and the vectors b_ℓ, $\ell \in \{1, 2, \dots, L\}$,*
- depth $\mathcal{L}(\Phi) := L$,
- width $\mathcal{W}(\Phi) := \max_{\ell=0,\dots,L} N_\ell$,
- weight magnitude $\mathcal{B}(\Phi) := \max_{\ell=1,\dots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$.

Remark II.2. *Note that for a given function $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$, which can be expressed according to* (1)*, the underlying affine transformations W_ℓ are highly nonunique in general [19], [20]. The question of uniqueness in this context is of independent interest and was addressed recently in [21], [22]. Whenever we talk about a given ReLU network Φ, we will either explicitly or implicitly associate Φ with a given set of affine transformations W_ℓ.*

N_0 is the dimension of the input layer *indexed as the 0-th layer, $N_1, \dots, N_{L-1}$ are the* dimensions of the $L-1$ hidden layers*, and N_L is the* dimension of the output layer. *Our definition of depth $\mathcal{L}(\Phi)$ counts the number of affine transformations involved in the representation (1). Single-hidden-layer neural networks hence have depth 2 in this terminology. Finally, we consider standard affine transformations as neural networks of depth 1 for technical purposes.*

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*

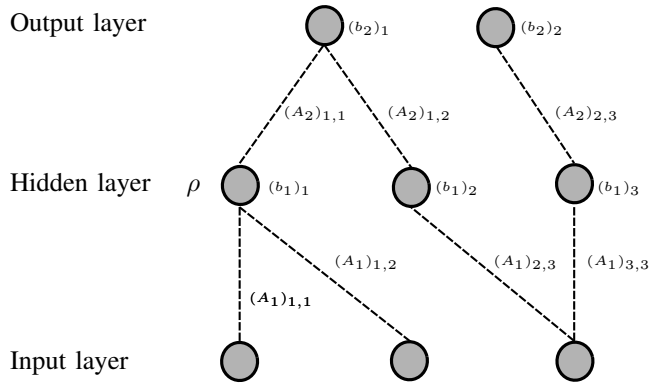
assignments are schematized in Figure 1. The real numbers $(A_\ell)_{i,j}$ and $(b_\ell)_i$ are referred to as the network's edge weights and node weights, respectively.

Throughout the paper, we assume that every node in the input layer and in layers $1, \dots, L-1$ has at least one outgoing edge and every node in the output layer L has at least one incoming edge. These nondegeneracy assumptions are basic as nodes that do not satisfy them can be removed without changing the functional relationship realized by the network.

Finally, we note that the connectivity satisfies

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

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 mostly consider the case $\Phi : \mathbb{R}^d \to \mathbb{R}$, i.e., $N_L = 1$, but emphasize that our results readily generalize to $N_L > 1$.

The neural network constructions provided in the paper frequently make use of basic elements introduced next, namely compositions and linear combinations of networks [23].

Lemma II.3. *Let $d_1, d_2, d_3 \in \mathbb{N}$, $\Phi_1 \in \mathcal{N}_{d_1,d_2}$, and $\Phi_2 \in \mathcal{N}_{d_2,d_3}$. Then, there exists a network $\Psi \in \mathcal{N}_{d_1,d_3}$ with $\mathcal{L}(\Psi) = \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_2)$, $\mathcal{M}(\Psi) \leq 2\mathcal{M}(\Phi_1) + 2\mathcal{M}(\Phi_2)$, $\mathcal{W}(\Psi) \leq \max\{2d_2, \mathcal{W}(\Phi_1), \mathcal{W}(\Phi_2)\}$, $\mathcal{B}(\Psi) = \max\{\mathcal{B}(\Phi_1), \mathcal{B}(\Phi_2)\}$, and satisfying*

$$\Psi(x) = (\Phi_2 \circ \Phi_1)(x) = \Phi_2(\Phi_1(x)), \quad \textit{for all } x \in \mathbb{R}^{d_1}.$$

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

$$\Phi_1 = W_{L_1}^1 \circ \rho \circ W_{L_1-1}^1 \circ \cdots \circ \rho \circ W_1^1 \quad \text{and} \quad \Phi_2 = W_{L_2}^2 \circ \rho \circ \cdots \circ W_2^2 \circ \rho \circ W_1^2.$$

Next, let $N_{L_1-1}^1$ denote the width of layer $L_1 - 1$ in Φ_1 and let N_1^2 denote the width of layer 1 in Φ_2. We define the affine transformations $\widetilde{W}_{L_1}^1 \colon \mathbb{R}^{N_{L_1-1}^1} \mapsto \mathbb{R}^{2d_2}$ and $\widetilde{W}_1^2 \colon \mathbb{R}^{2d_2} \mapsto \mathbb{R}^{N_1^2}$ according to

$$\widetilde{W}_{L_1}^1(x) := \begin{pmatrix} \mathbb{I}_{d_2} \\ -\mathbb{I}_{d_2} \end{pmatrix} W_{L_1}^1(x) \quad \text{and } \widetilde{W}_1^2(y) := W_1^2 \left(\begin{pmatrix} \mathbb{I}_{d_2} & -\mathbb{I}_{d_2} \end{pmatrix} y \right).$$

The proof is finalized by noting that the network

$$\Psi := W_{L_2}^2 \circ \rho \circ \cdots \circ W_2^2 \circ \rho \circ \widetilde{W}_1^2 \circ \rho \circ \widetilde{W}_{L_1}^1 \circ \rho \circ W_{L_1-1}^1 \circ \cdots \circ \rho \circ W_1^1$$

satisfies the claimed properties. □

Unless explicitly stated otherwise, the composition of two neural networks will be understood in the sense of Lemma II.3.

In order to formalize the concept of a linear combination of networks with possibly different depths, we need the following two technical lemmas which show how to augment network depth while retaining the network's input-output relation and how to parallelize networks.

Lemma II.4. *Let $d_1, d_2, K \in \mathbb{N}$, and $\Phi \in \mathcal{N}_{d_1,d_2}$ with $\mathcal{L}(\Phi) < K$. Then, there exists a network $\Psi \in \mathcal{N}_{d_1,d_2}$ with $\mathcal{L}(\Psi) = K$, $\mathcal{M}(\Psi) \leq \mathcal{M}(\Phi) + d_2\mathcal{W}(\Phi) + 2d_2(K - \mathcal{L}(\Phi))$, $\mathcal{W}(\Psi) = \max\{2d_2, \mathcal{W}(\Phi)\}$, $\mathcal{B}(\Psi) = \max\{1, \mathcal{B}(\Phi)\}$, and satisfying $\Psi(x) = \Phi(x)$ for all $x \in \mathbb{R}^{d_1}$.*

Proof. Let $\widetilde{W}_j(x) := \operatorname{diag}(\mathbb{I}_{d_2}, \mathbb{I}_{d_2})\, x$, for $j \in \{\mathcal{L}(\Phi) + 1, \ldots, K - 1\}$, $\widetilde{W}_K(x) := \begin{pmatrix} \mathbb{I}_{d_2} & -\mathbb{I}_{d_2} \end{pmatrix} x$, and note that with

$$\Phi = W_{\mathcal{L}(\Phi)} \circ \rho \circ W_{\mathcal{L}(\Phi)-1} \circ \rho \circ \cdots \circ \rho \circ W_1,$$

the network

$$\Psi := \widetilde{W}_K \circ \rho \circ \widetilde{W}_{K-1} \circ \rho \circ \cdots \circ \rho \circ \widetilde{W}_{\mathcal{L}(\Phi)+1} \circ \rho \circ \begin{pmatrix} W_{\mathcal{L}(\Phi)} \\ -W_{\mathcal{L}(\Phi)} \end{pmatrix} \circ \rho \circ W_{\mathcal{L}(\Phi)-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

satisfies the claimed properties. □

For the sake of simplicity of exposition, we state the following two lemmas only for networks of the same depth, the extension to the general case follows by straightforward application of Lemma II.4. The first of these two lemmas formalizes the notion of neural network parallelization, concretely of combining neural networks implementing the functions f and g into a neural network realizing the mapping $x \mapsto (f(x), g(x))$.

Lemma II.5. *Let $n, L \in \mathbb{N}$ and, for $i \in \{1, 2, \ldots, n\}$, let $d_i, d_i' \in \mathbb{N}$ and $\Phi_i \in \mathcal{N}_{d_i, d_i'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{\sum_{i=1}^n d_i, \sum_{i=1}^n d_i'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) = \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) = \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i \mathcal{B}(\Phi_i)$, and satisfying*

$$\Psi(x) = (\Phi_1(x_1), \Phi_2(x_2), \ldots, \Phi_n(x_n)) \in \mathbb{R}^{\sum_{i=1}^n d_i'},$$

for $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{\sum_{i=1}^n d_i}$ with $x_i \in \mathbb{R}^{d_i}$, $i \in \mathbb{N}$.

Proof. We write the networks Φ_i as

$$\Phi_i = W_L^i \circ \rho \circ W_{L-1}^i \circ \rho \circ \cdots \circ \rho \circ W_1^i,$$

with $W_\ell^i(x) = A_\ell^i x + b_\ell^i$. Furthermore, we denote the layer dimensions of Φ_i by $N_0^i, \ldots, N_L^i$ and set $N_\ell := \sum_{i=1}^n N_\ell^i$, for $\ell \in \{0, 1, \ldots, L\}$. Next, define, for $\ell \in \{1, 2, \ldots, L\}$, the block-diagonal matrices $A_\ell := \operatorname{diag}(A_\ell^1, A_\ell^2, \ldots, A_\ell^n)$, the vectors $b_\ell = (b_\ell^1, b_\ell^2, \ldots, b_\ell^n)$, and the affine transformations $W_\ell(x) := A_\ell x + b_\ell$. The proof is concluded by noting that

$$\Psi := W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

satisfies the claimed properties. □

We are now ready to formalize the concept of a linear combination of neural networks.

Lemma II.6. *Let $n, L, d' \in \mathbb{N}$ and, for $i \in \{1, 2, \ldots, n\}$, let $d_i \in \mathbb{N}$, $a_i \in \mathbb{R}$, and $\Phi_i \in \mathcal{N}_{d_i, d'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{\sum_{i=1}^n d_i, d'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) \leq \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) \leq \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i \{|a_i| \mathcal{B}(\Phi_i)\}$, and satisfying*

$$\Psi(x) = \sum_{i=1}^n a_i \Phi_i(x_i) \in \mathbb{R}^{d'},$$

for $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{\sum_{i=1}^n d_i}$ with $x_i \in \mathbb{R}^{d_i}$, $i \in \{1, 2, \ldots, n\}$.

Proof. The proof follows by taking the construction in Lemma II.5, replacing A_L by $(a_1 A_L^1, a_2 A_L^2, \ldots, a_n A_L^n)$, b_L by $\sum_{i=1}^n a_i b_L^i$, and noting that the resulting network satisfies the claimed properties. □

III. Approximation of multiplication, polynomials, smooth functions, and sinusoidals

This section constitutes the first part of the paper dealing with the approximation of basic function "templates" through neural networks. Specifically, we shall develop an algebra of neural network approximation by starting with the squaring function, building thereon to approximate the multiplication function, proceeding to polynomials and general smooth functions, and ending with sinusoidal functions.

The basic element of the neural network algebra we develop is based on an approach by Yarotsky [24] and by Schmidt-Hieber [25], both of whom, in turn, employed the "sawtooth" construction from [26].

We start by reviewing the sawtooth construction underlying our program. Consider the hat function $g : \mathbb{R} \to [0,1]$,

$$g(x) = 2\rho(x) - 4\rho(x-\tfrac{1}{2}) + 2\rho(x-1) = \begin{cases} 2x, & \text{if } 0 \le x < \frac{1}{2} \\ 2(1-x), & \text{if } \frac{1}{2} \le x \le 1 \\ 0, & \text{else} \end{cases},$$

let $g_0(x) = x, g_1(x) = g(x)$, and define the s-th order sawtooth function g_s as the s-fold composition of g with itself, i.e.,

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

We note that g can be realized by a 2-layer network $\Phi_g \in \mathcal{N}_{1,1}$ according to $\Phi_g := W_2 \circ \rho \circ W_1 = g$ with

$$W_1(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} x \; - \; \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}, \qquad W_2(x) = \begin{pmatrix} 2 & -4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The s-th order sawtooth function g_s can hence be realized by a network $\Phi_g^s \in \mathcal{N}_{1,1}$ according to

$$\Phi_g^s := W_2 \circ \rho \circ \underbrace{W_g \circ \rho \circ \cdots \circ W_g \circ \rho}_{s-1} \circ W_1 = g_s \tag{3}$$

with

$$W_g(x) = \begin{pmatrix} 2 & -4 & 2 \\ 2 & -4 & 2 \\ 2 & -4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}.$$

The following restatement of [26, Lemma 2.4] summarizes the self-similarity and symmetry properties of $g_s(x)$ we will frequently make use of.

Lemma III.1. *For $s \in \mathbb{N}$, $k \in \{0, 1, \ldots, 2^{s-1} - 1\}$, it holds that $g(2^{s-1} \cdot -k)$ is supported in $\left[\frac{k}{2^{s-1}}, \frac{k+1}{2^{s-1}}\right]$,*

$$g_s(x) = \sum_{k=0}^{2^{s-1}-1} g(2^{s-1}x - k), \quad \text{for } x \in [0,1],$$

and

$$g_s\left(\tfrac{k}{2^{s-1}} + x\right) = g_s\left(\tfrac{k+1}{2^{s-1}} - x\right), \quad \text{for } x \in \left[0, \tfrac{1}{2^{s-1}}\right].$$

We are now ready to proceed with the statement of the basic building block of our neural network algebra, namely the approximation of the squaring function through deep ReLU networks.

Proposition III.2. *There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_\varepsilon) \le C \log(\varepsilon^{-1})$, $\mathcal{W}(\Phi_\varepsilon) = 3$, $\mathcal{B}(\Phi_\varepsilon) = 1$, $\Phi_\varepsilon(0) = 0$, satisfying*

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

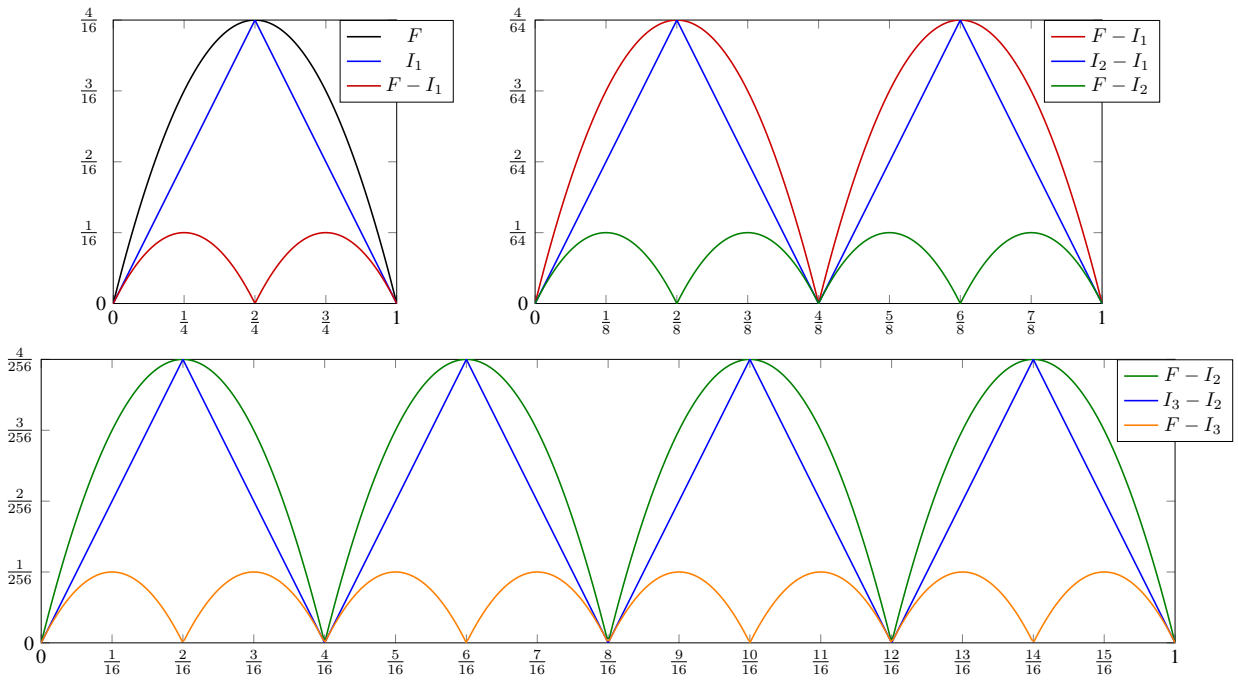


Fig. 2: First three steps of approximating $F(x) = x - x^2$ by an equispaced linear interpolation I_m at $2^m + 1$ points.

Proof. The proof builds on two rather elementary observations. The first one concerns the linear interpolation $I_m \colon [0,1] \to \mathbb{R}$, $m \in \mathbb{N}$, of the function $F(x) := x - x^2$ at the points $\frac{j}{2^m}$, $j \in \{0, 1, \dots, 2^m\}$, and in particular the self-similarity of the refinement step $I_m \to I_{m+1}$. For every $m \in \mathbb{N}$, the residual $F - I_m$ is identical on each interval between two points of interpolation (see Figure 2). Concretely, let $f_m \colon [0, 2^{-m}] \to [0, 2^{-2m-2}]$ be defined as $f_m(x) = 2^{-m}x - x^2$ and consider its linear interpolation $h_m \colon [0, 2^{-m}] \to [0, 2^{-2m-2}]$ at the midpoint and the endpoints of the interval $[0, 2^{-m}]$ given by

$$h_m(x) := \begin{cases} 2^{-m-1}x, & x \in [0, 2^{-m-1}] \\ -2^{-m-1}x + 2^{-2m-1}, & x \in [2^{-m-1}, 2^{-m}] \end{cases}.$$

Direct calculation shows that

$$f_m(x) - h_m(x) = \begin{cases} f_{m+1}(x), & x \in [0, 2^{-m-1}] \\ f_{m+1}(x - 2^{-m-1}), & x \in [2^{-m-1}, 2^{-m}] \end{cases}.$$

As $F = f_0$ and $I_1 = h_0$ this implies that, for all $m \in \mathbb{N}$,

$$F(x) - I_m(x) = f_m(x - \tfrac{j}{2^m}), \text{ for } x \in [\tfrac{j}{2^m}, \tfrac{j+1}{2^m}], \ j \in \{0, 1, \dots, 2^m - 1\}$$

and $I_m = \sum_{k=0}^{m-1} H_k$, where $H_k \colon [0,1] \to \mathbb{R}$ is given by

$$H_k(x) = h_k(x - \tfrac{j}{2^k}), \text{ for } x \in [\tfrac{j}{2^k}, \tfrac{j+1}{2^k}], \ j \in \{0, 1, \dots, 2^k - 1\}.$$

Thus, we have

$$\sup_{x\in[0,1]} |x^2 - (x - I_m(x))| = \sup_{x\in[0,1]} |F(x) - I_m(x)| = \sup_{x\in[0,2^{-m}]} |f_m(x)| = 2^{-2m-2}. \tag{4}$$

The second observation we build on is a manifestation of the sawtooth construction described above and leads to economic realizations of the H_k through k-layer networks with two neurons in each layer; a third neuron is used to realize the approximation $x - I_m(x)$ to x^2. Concretely, let $s_k(x) := 2^{-1}\rho(x) - \rho(x - 2^{-2k-1})$, and note that, for $x \in [0,1]$, $H_0 = s_0$, we get $H_k = s_k \circ H_{k-1}$. We can thus construct a network realizing $x - I_m(x)$, for $x \in [0,1]$, as follows. Let $A_1 := (1,1,1)^T \in \mathbb{R}^{3\times 1}$, $b_1 := (0, -2^{-1}, 0)^T \in \mathbb{R}^3$,

$$A_\ell := \begin{pmatrix} 2^{-1} & -1 & 0 \\ 2^{-1} & -1 & 0 \\ -2^{-1} & 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad b_\ell := \begin{pmatrix} 0 \\ -2^{-2\ell+1} \\ 0 \end{pmatrix} \in \mathbb{R}^3, \quad \text{for } \ell \in \{2, \ldots, m\},$$

and $A_{m+1} := (-2^{-1}, 1, 1) \in \mathbb{R}^{1\times 3}$, $b_{m+1} = 0$. Setting $W_\ell(x) := A_\ell x + b_\ell$, $\ell \in \{1, 2, \ldots, m+1\}$, and

$$\tilde{\Phi}_m := W_{m+1} \circ \rho \circ W_m \circ \rho \circ \cdots \circ \rho \circ W_1,$$

a direct calculation yields $\tilde{\Phi}_m(x) = x - \sum_{k=0}^{m-1} H_k(x)$, for $x \in [0,1]$. The proof is completed upon noting that the networks $\Phi_\varepsilon := \tilde{\Phi}_{\lceil \log(\varepsilon^{-1})/2 \rceil}$ satisfy the claimed properties. □

The symmetry properties of $g_s(x)$ according to Lemma III.1 lead to the interpolation error in the proof of Proposition III.2 being identical in each interval, with the maximum error taken on at the centers of the respective intervals. More importantly, however, the approximating neural networks realize linear interpolation at a number of points that grows exponentially in network depth. This is a manifestation of the fact that the number of linear regions in the sawtooth construction (3) grows exponentially with depth, which, owing to Lemma XI.1, is optimal. We emphasize that the theory developed in this paper hinges critically on this optimality property, which, however, is brittle in the sense that networks with weights obtained through training will, as observed in [27], in general, not exhibit exponential growth of the number of linear regions with network depth. An interesting approach to neural network training which manages to partially circumvent this problem was proposed recently in [28]. Understanding how the number of linear regions grows in general trained networks and quantifying the impact of this—possibly subexponential—growth behavior on the approximation-theoretic fundamental limits of neural networks constitutes a major open problem.

We proceed to the construction of networks that approximate the multiplication function over the interval $[-D, D]$. This will be effected by using the result on the approximation of x^2 just established combined with the polarization identity $xy = \frac{1}{4}((x+y)^2 - (x-y)^2)$, the fact that $\rho(x) + \rho(-x) = |x|$, and a scaling argument exploiting that the ReLU function is positive homogeneous, i.e., $\rho(\lambda x) = \lambda\rho(x)$, for all $\lambda \geq 0$, $x \in \mathbb{R}$.

Proposition III.3. *There exists a constant $C > 0$ such that, for all $D \in \mathbb{R}_+$ and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{N}_{2,1}$ with $\mathcal{L}(\Phi_{D,\varepsilon}) \leq C(\log(\lceil D \rceil) + \log(\varepsilon^{-1}))$, $\mathcal{W}(\Phi_{D,\varepsilon}) \leq 5$, $\mathcal{B}(\Phi_{D,\varepsilon}) = 1$, satisfying $\Phi_{D,\varepsilon}(0, x) = \Phi_{D,\varepsilon}(x, 0) = 0$, for all $x \in \mathbb{R}$, and*

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

Proof. We first note that, w.l.o.g., we can assume $D \geq 1$ in the following, as for $D < 1$, we can simply employ the network constructed for $D = 1$ to guarantee the claimed properties. The proof builds on the polarization identity and essentially constructs two squaring networks according to Proposition III.2 which share the neuron responsible for summing up the H_k, preceded by a layer mapping (x, y) to $(|x + y|/(2D), |x - y|/(2D))$ and followed by layers realizing the multiplication by D^2 through weights bounded by 1. Specifically, consider the network $\tilde{\Psi}_m$ with associated matrices A_ℓ and vectors b_ℓ given by

$$A_1 := \frac{1}{2D}\begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} \in \mathbb{R}^{4\times 2}, \quad b_1 := 0 \in \mathbb{R}^4, \quad A_2 := \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{5\times 4}, \quad b_2 := \begin{pmatrix} 0 \\ -2^{-1} \\ 0 \\ 0 \\ -2^{-1} \end{pmatrix}$$

$$A_\ell := \begin{pmatrix} 2^{-1} & -1 & 0 & 0 & 0 \\ 2^{-1} & -1 & 0 & 0 & 0 \\ -2^{-1} & 1 & 1 & 2^{-1} & -1 \\ 0 & 0 & 0 & 2^{-1} & -1 \\ 0 & 0 & 0 & 2^{-1} & -1 \end{pmatrix} \in \mathbb{R}^{5\times 5}, \quad b_\ell := \begin{pmatrix} 0 \\ -2^{-2\ell+3} \\ 0 \\ 0 \\ -2^{-2\ell+3} \end{pmatrix}, \quad \text{for } \ell \in \{3, \dots, m+1\},$$

and $A_{m+2} := (-2^{-1}, 1, 1, 2^{-1}, -1) \in \mathbb{R}^{1\times 5}$, $b_{m+2} := 0$. A direct calculation yields

$$\begin{aligned} \tilde{\Psi}_m(x, y) &= \left(\tfrac{|x+y|}{2D} - \sum_{k=0}^{m-1} H_k\big(\tfrac{|x+y|}{2D}\big)\right) - \left(\tfrac{|x-y|}{2D} - \sum_{k=0}^{m-1} H_k\big(\tfrac{|x-y|}{2D}\big)\right) \\ &= \tilde{\Phi}_m\left(\tfrac{|x+y|}{2D}\right) - \tilde{\Phi}_m\left(\tfrac{|x-y|}{2D}\right), \end{aligned} \tag{6}$$

with H_k and $\tilde{\Phi}_m$ as defined in the proof of Proposition III.2. With (4) this implies

$$\begin{aligned} \sup_{(x,y)\in[-D,D]^2} \left|\tilde{\Psi}_m(x, y) - \tfrac{xy}{D^2}\right| &= \sup_{(x,y)\in[-D,D]^2} \left|\left(\tilde{\Phi}_m\left(\tfrac{|x+y|}{2D}\right) - \tilde{\Phi}_m\left(\tfrac{|x-y|}{2D}\right)\right) - \left(\left(\tfrac{|x+y|}{2D}\right)^2 - \left(\tfrac{|x-y|}{2D}\right)^2\right)\right| \\ &\leq 2 \sup_{z\in[0,1]} |\tilde{\Phi}_m(z) - z^2| \leq 2^{-2m-1}. \end{aligned} \tag{7}$$

Next, let $\Psi_D(x) = D^2 x$ be the scalar multiplication network according to Lemma A.1 and take $\Phi_{D,\varepsilon} := \Psi_D \circ \tilde{\Psi}_{m(D,\varepsilon)}$, where $m(D, \varepsilon) := \lceil 2^{-1}(1 + \log(D^2\varepsilon^{-1}))\rceil$. Then, the error estimate (5) follows directly from (7) and Lemma II.3 establishes the desired bounds on depth, width, and weight magnitude. Finally, $\Phi_{D,\varepsilon}(0, x) = \Phi_{D,\varepsilon}(x, 0) = 0$, for all $x \in \mathbb{R}$, follows directly from (6). □

Remark III.4. *Note that the multiplication network just constructed has weights bounded by* 1 *irrespectively of the size* D *of the domain. This is accomplished by trading network depth for weight magnitude according to Lemma A.1.*

We proceed to the approximation of polynomials, effected by networks that realize linear combinations of monomials, which, in turn, are built by composing multiplication networks. Before presenting the specifics of this construction, we hasten to add that a similar approach was considered previously in [24] and [25]. While there are slight differences in formulation, the main distinction between our construction and those in [24] and [25] resides in their purpose. Specifically, the goal in [24] and [25] is to establish, by way of local Taylor-series approximation, that d-variate, k-times (weakly) differentiable functions can be approximated in L^∞-norm to within error ε with networks of connectivity scaling according to $\varepsilon^{-d/k}\log(\varepsilon^{-1})$. Here, on the other hand, we will be interested in functions that allow approximation with networks of connectivity scaling polylogarithmically in ε^{-1} (i.e., as a polynomial in $\log(\varepsilon^{-1})$). Moreover, for ease of exposition, we will employ finite-width networks. Polylogarithmic connectivity scaling will turn out to be crucial (see Sections VI-IX) in establishing Kolmogorov-Donoho rate-distortion optimality of neural networks in the approximation of a variety of prominent function classes. Finally, we would like to mention related recent work [29], [30], [31] on the approximation of Sobolev-class functions in certain Sobolev norms enabled by neural network approximations of the multiplication operation and of polynomials.

Proposition III.5. *There exists a constant* $C>0$ *such that for all* $m\in\mathbb{N}$, $a=(a_i)_{i=0}^m\in\mathbb{R}^{m+1}$, $D\in\mathbb{R}_+$, *and* $\varepsilon\in(0,1/2)$, *there is a network* $\Phi_{a,D,\varepsilon}\in\mathcal{N}_{1,1}$ *with* $\mathcal{L}(\Phi_{a,D,\varepsilon})\leq Cm(\log(\varepsilon^{-1})+m\log(\lceil D\rceil)+\log(m)+\log(\lceil\|a\|_\infty\rceil))$, $\mathcal{W}(\Phi_{a,D,\varepsilon})\leq 9$, $\mathcal{B}(\Phi_{a,D,\varepsilon})\leq 1$, *and satisfying*

$$\|\Phi_{a,D,\varepsilon}(x)-\sum_{i=0}^m a_i x^i\|_{L^\infty([-D,D])}\leq\varepsilon.$$

Proof. As in the proof of Proposition III.3 and for the same reason, it suffices to consider the case $D\geq 1$. For $m=1$, we simply have an affine transformation and the statement follows directly from Corollary A.2. 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 that uses the network realizing x^k not only as a building block in the network implementing x^{k+1} but also to approximate the partial sum $\sum_{i=0}^k a_i x^i$ in parallel.

We start by setting $B_k=B_k(D,\eta):=\lceil D\rceil^k+\eta\sum_{s=0}^{k-2}\lceil D\rceil^s$, $k\in\mathbb{N},\eta\in\mathbb{R}_+$ and take $\Phi_{B_k,\eta}$ to be the multiplication network from Proposition III.3. Next, we recursively define the functions

$$f_{k,D,\eta}(x)=\Phi_{B_{k-1},\eta}(x,f_{k-1,D,\eta}(x)),\quad k\geq 2,$$

with $f_{0,D,\eta}(x)=1$ and $f_{1,D,\eta}(x)=x$. For notational simplicity, we use the abbreviation $f_k=f_{k,D,\eta}$ in the following. First, we verify that the $f_{k,D,\eta}$ approximate monomials sufficiently well. Specifically, we prove by induction that

$$\|f_k(x)-x^k\|_{L^\infty([-D,D])}\leq\eta\sum_{s=0}^{k-2}\lceil D\rceil^s,\tag{8}$$

for all $k \geq 2$. The base case $k = 2$, i.e.,

$$\|f_2(x) - x^2\|_{L^\infty([-D,D])} = \|\Phi_{B_1,\eta}(x,x) - x^2\|_{L^\infty([-D,D])} \leq \eta,$$

follows directly from Proposition III.3 upon noting that $D \leq B_1 = \lceil D \rceil$ (we take the sum in the definition of B_k to equal zero when the upper limit of summation is negative). We proceed to establish the induction step $(k-1) \to k$ with the induction assumption given by

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

As

$$\|f_{k-1}\|_{L^\infty([-D,D])} \leq \|x^{k-1}\|_{L^\infty([-D,D])} + \|f_{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \leq B_{k-1},$$

application of Proposition III.3 yields

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

which completes the induction.

We now construct the network $\Phi_{a,D,\varepsilon}$ approximating the polynomial $\sum_{i=0}^m a_i x^i$. To this end, note that there exists a constant C' such that for all $m \geq 2$, $a = (a_i)_{i=0}^m \in \mathbb{R}^{m+1}$, and $i \in \{1, \ldots, m-1\}$, there is a network $\Psi^i_{a,D,\eta} \in \mathcal{N}_{3,3}$ with $\mathcal{L}(\Psi^i_{a,D,\eta}) \leq C'(\log(\eta^{-1}) + \log(\lceil B_i \rceil) + \log(\|a\|_\infty))$, $\mathcal{W}(\Psi^i_{a,D,\eta}) \leq 9$, $\mathcal{B}(\Psi^i_{a,D,\eta}) \leq 1$, and satisfying

$$\Psi^i_{a,D,\eta}(x,s,y) = (x, s + a_i y, \Phi_{B_i,\eta}(x,y)).$$

To see that this is, indeed, the case, consider the following chain of mappings

$$(x,s,y) \xrightarrow{(I)} (x,s,y,y) \xrightarrow{(II)} (x, s + a_i y, y) \xrightarrow{(III)} (x, s + a_i y, x, y) \xrightarrow{(IV)} (x, s + a_i y, \Phi_{B_i,\eta}(x,y)).$$

Observe that the mapping (I) is an affine transformation with coefficients in $\{0,1\}$, which we can simply consider to be a depth-1 network. The mapping (II) is obtained by using Corollary A.2 in order to implement the affine transformation $(s,y) \mapsto s + a_i y$ with weights bounded by 1, followed by application of Lemmas II.4 and II.5 to put this network in parallel with two networks realizing the identity mapping according to $x = \rho(x) - \rho(-x)$. Mapping (III) is obtained along the same lines by putting the result of mapping (II) in parallel with another network realizing the identity mapping. Finally, mapping (IV) is realized by putting the network $\Phi_{B_i,\eta}$ in parallel with two identity networks. Composing these four networks according to Lemma II.3 yields, for $i \in \{1, \ldots, m-1\}$, a network $\Psi^i_{a,D,\eta}$ with the claimed properties. Next, we employ Corollary A.2 to get networks $\Psi^0_{a,D,\eta}$ which implement $x \mapsto (x, a_0, x)$

as well as networks $\Psi^m_{a,D,\eta}$ realizing $(x,s,y)\mapsto s+a_m y$. Let now $\eta=\eta(a,D,\varepsilon):=(\|a\|_\infty(m-1)^2\lceil D\rceil^{m-2})^{-1}\varepsilon$ and define

$$\Phi_{a,D,\varepsilon}:=\Psi^m_{a,D,\eta}\circ\Psi^{m-1}_{a,D,\eta}\circ\cdots\circ\Psi^1_{a,D,\eta}\circ\Psi^0_{a,D,\eta}.$$

A direct calculation yields

$$\Phi_{a,D,\varepsilon}=\sum_{i=0}^m a_i f_{i,D,\eta}.$$

Hence (8) implies

$$\begin{aligned}\Big\|\Phi_{a,D,\varepsilon}(x)-\sum_{i=0}^m a_i x^i\Big\|_{L^\infty([-D,D])}&\le\sum_{i=0}^m|a_i|\|f_{i,D,\eta}(x)-x^i\|_{L^\infty([-D,D])}\le\sum_{i=2}^m|a_i|\Big(\eta\sum_{s=0}^{i-2}\lceil D\rceil^s\Big)\\&\le\|a\|_\infty\eta\sum_{k=0}^{m-2}(m-1-k)\lceil D\rceil^k\le\|a\|_\infty(m-1)^2\lceil D\rceil^{m-2}\eta=\varepsilon.\end{aligned}$$

Lemma II.3 now establishes that $\mathcal{W}(\Phi_{a,D,\varepsilon})\le 9$, $\mathcal{B}(\Phi_{a,D,\varepsilon})\le 1$, and

$$\begin{aligned}\mathcal{L}(\Phi_{a,D,\varepsilon})&\le\sum_{i=0}^m\mathcal{L}(\Psi^i_{a,D,\eta})\\&\le 2(\log(\lceil\|a\|_\infty\rceil)+5)+\sum_{i=1}^{m-1}C'(\log(\eta^{-1})+\log(\lceil B_{i-1}\rceil)+\log(\lceil\|a\|_\infty\rceil))\\&\le Cm(\log(\varepsilon^{-1})+m\log(\lceil D\rceil)+\log(m)+\log(\lceil\|a\|_\infty\rceil))\end{aligned}$$

for a suitably chosen absolute constant C. This completes 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.6 ([32])**.** *Let $[a,b]\subseteq\mathbb{R}$ and $f\in C([a,b])$. Then, for every $\varepsilon>0$, there exists a polynomial π such that*

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

Proposition III.5 hence allows us 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 9. This amounts to a variant of the universal approximation theorem [11], [12] for finite-width deep ReLU networks. A quantitative statement in terms of making the approximating network's width, depth, and weight bounds explicit can be obtained for (very) smooth functions by applying Proposition III.5 to Lagrangian interpolation with Chebyshev points.

Lemma III.7. *Consider the set*

$$\mathcal{S}_{[-1,1]} := \left\{ f \in C^\infty([-1,1],\mathbb{R}) \colon \|f^{(n)}(x)\|_{L^\infty([-1,1])} \leq n!, \text{ for all } n \in \mathbb{N}_0 \right\}.$$

There exists a constant $C > 0$ such that for all $f \in \mathcal{S}_{[-1,1]}$ and $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{f,\varepsilon} \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Psi_{f,\varepsilon}) \leq C(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Psi_{f,\varepsilon}) \leq 9$, $\mathcal{B}(\Psi_{f,\varepsilon}) \leq 1$, and satisfying

$$\|\Psi_{f,\varepsilon} - f\|_{L^\infty([-1,1])} \leq \varepsilon.$$

Proof. A fundamental result on Lagrangian interpolation with Chebyshev points (see e.g. [33, Lemma 3]) guarantees, for all $f \in \mathcal{S}_{[-1,1]}$, $m \in \mathbb{N}$, the existence of a polynomial $P_{f,m}$ of degree m such that

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

Note that $P_{f,m}$ can be expressed in the Chebyshev basis (see e.g. [34, Section 3.4.1]) according to $P_{f,m} = \sum_{j=0}^m c_{f,m,j} T_j(x)$ with $|c_{f,m,j}| \leq 2$ and the Chebyshev polynomials defined through the two-term recursion $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, $k \geq 2$, with $T_0(x) = 1$ and $T_1(x) = x$. We can moreover use this recursion to conclude that the coefficients of the T_k in the monomial basis are upper-bounded by 3^k. Consequently, we can express $P_{f,m}$ according to $P_{f,m} = \sum_{j=0}^m a_{f,m,j} x^j$ with

$$A_{f,m} := \max_{j=0,\ldots,m} |a_{f,m,j}| \leq 2(m+1)3^m.$$

Application of Proposition III.5 to $P_{f,m}$ in the monomial basis, with $m = \lceil \log(2/\varepsilon) \rceil$ and approximation error $\varepsilon/2$, completes the proof upon noting that

$$C' m(\log(2/\varepsilon) + \log(m) + \log(|A_{f,m}|)) \leq C(\log(\varepsilon^{-1}))^2$$

for some absolute constant C. □

An extension of Lemma III.7 to approximation over general intervals is provided in Lemma A.6. While Lemma III.7 shows that a specific class of C^∞-functions, namely those whose derivatives are suitably bounded, can be approximated by neural networks with connectivity growing polylogarithmically in ε^{-1}, it turns out that this is not possible for general (Sobolev-class) k-times differentiable functions [24, Thm. 4].

We are now ready to proceed to the approximation of sinusoidal functions. Before stating the corresponding result, we comment on the basic idea enabling the approximation of oscillatory functions through deep neural networks. In essence, we exploit the optimality of the sawtooth construction (3) in terms of achieving exponential—in network depth—growth in the number of linear regions. As indicated in Figure 3, the composition of the cosine function (realized according to Lemma III.7) with the sawtooth function, combined with the symmetry properties of the cosine function and the sawtooth function, yields oscillatory behavior that increases exponentially with network depth.

Theorem III.8. *There exists a constant $C > 0$ such that for every $a, D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Psi_{a,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil aD \rceil))$, $\mathcal{W}(\Psi_{a,D,\varepsilon}) \leq 9$, $\mathcal{B}(\Psi_{a,D,\varepsilon}) \leq 1$, and satisfying*

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

Proof. Note that $f(x) := (6/\pi^3)\cos(\pi x)$ is in $\mathcal{S}_{[-1,1]}$. Thus, by Lemma III.7, there exists a constant $C > 0$ such that for every $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_\varepsilon) \leq C(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Phi_\varepsilon) \leq 9$, $\mathcal{B}(\Phi_\varepsilon) \leq 1$, and satisfying

$$\|\Phi_\varepsilon - f\|_{L^\infty([-1,1])} \leq \tfrac{6}{\pi^3}\varepsilon. \tag{9}$$

We now 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 $x \mapsto \cos(\pi x)$ is 2-periodic and even. Let $g_s \colon [0, 1] \to [0, 1]$, $s \in \mathbb{N}$, be the s-th order sawtooth functions as defined in (2) and note that, due to the periodicity and the symmetry of the cosine function (see Figure 3 for illustration), we have for all $s \in \mathbb{N}_0$, $x \in [-1, 1]$,

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

For $a > \pi$, we define $s = s(a) := \lceil \log(a) - \log(\pi) \rceil$ and $\alpha = \alpha(a) := (\pi 2^s)^{-1} a \in (1/2, 1]$, and note that

$$\cos(ax) = \cos(\pi 2^s \alpha x) = \cos(\pi g_s(\alpha|x|)), \quad x \in [-1, 1].$$

As $g_s(\alpha|x|) \in [0, 1]$, it follows from (9) that

$$\|\tfrac{\pi^3}{6}\Phi_\varepsilon(g_s(\alpha|x|)) - \cos(ax)\|_{L^\infty([-1,1])} = \tfrac{\pi^3}{6}\|\Phi_\varepsilon(g_s(\alpha|x|)) - f(g_s(\alpha|x|))\|_{L^\infty([-1,1])} \leq \varepsilon. \tag{10}$$

In order to realize $\Phi_\varepsilon(g_s(\alpha|x|))$ as a neural network, we start from the networks Φ_g^s defined in (3) and apply Proposition A.3 to convert them into networks $\Psi_g^s(x) = g_s(x)$, for $x \in [0, 1]$, with $\mathcal{B}(\Psi_g^s) \leq 1$, $\mathcal{L}(\Psi_g^s) = 7(s + 1)$, and $\mathcal{W}(\Psi_g^s) = 3$. Furthermore, let $\Psi(x) := \alpha\rho(x) - \alpha\rho(-x) = \alpha|x|$ and take $\Phi^{\text{mult}}_{\pi^3/6}$ to be the scalar multiplication network from Lemma A.1. Noting that $\Psi_{a,\varepsilon} := \Phi^{\text{mult}}_{\pi^3/6} \circ \Phi_\varepsilon \circ \Psi_g^s \circ \Psi = \Phi_\varepsilon(g_s(\alpha|x|))$ and concluding from Lemma II.3 that $\mathcal{L}(\Psi_{a,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil a \rceil))$, $\mathcal{W}(\Psi_{a,\varepsilon}) \leq 9$, and $\mathcal{B}(\Psi_{a,\varepsilon}) \leq 1$, together with (10), establishes the desired result for $a > \pi$ and for approximation over the interval $[-1, 1]$. For $a \in (0, \pi)$, we can simply take $\Psi_{a,\varepsilon} := \Phi^{\text{mult}}_{\pi^3/6} \circ \Phi_\varepsilon$ as $x \mapsto (6/\pi^3)\cos(ax)$ is in $\mathcal{S}_{[-1,1]}$ in this case.

Finally, we consider the approximation of $x \mapsto \cos(ax)$ on intervals $[-D, D]$, for arbitrary $D \geq 1$. To this end, we define the networks $\Psi_{a,D,\varepsilon}(x) := \Psi_{aD,\varepsilon}(\frac{x}{D})$ and observe that

$$\begin{aligned} \sup_{x \in [-D,D]} |\Psi_{a,D,\varepsilon}(x) - \cos(ax)| &= \sup_{y \in [-1,1]} |\Psi_{a,D,\varepsilon}(Dy) - \cos(aDy)| \\ &= \sup_{y \in [-1,1]} |\Psi_{aD,\varepsilon}(y) - \cos(aDy)| \leq \varepsilon. \end{aligned} \tag{11}$$

This concludes the proof. □

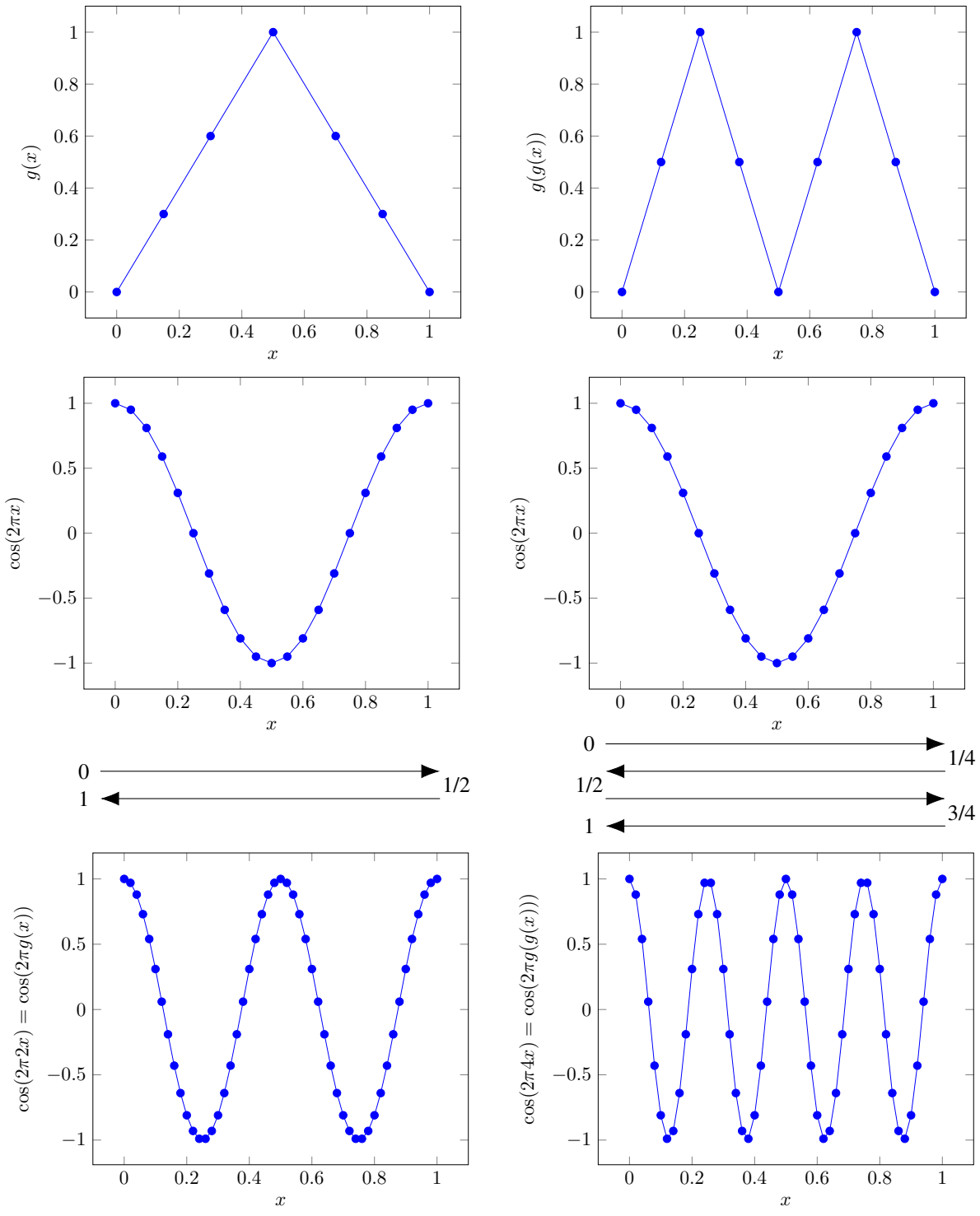


Fig. 3: Approximation of the function $\cos(2\pi ax)$ according to Theorem III.8 using "sawtooth" functions $g_s(x)$ as per (2), left $a = 2$, right $a = 4$.

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 III.9. *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 $\Psi_{a,b,D,\varepsilon} \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Psi_{a,b,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil aD + |b| \rceil))$, $\mathcal{W}(\Psi_{a,b,D,\varepsilon}) \leq 9$, $\mathcal{B}(\Psi_{a,b,D,\varepsilon}) \leq 1$, and satisfying*

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

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 III.8, and observe that, owing to (11),

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

□

Remark III.10. *The results in this section all have approximating networks of finite width and depth scaling polylogarithmically in ε^{-1}. 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 ε^{-1}. 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 Function Classes and Metric Entropy

So far we considered the explicit construction of deep neural networks for the approximation of a wide range of functions, namely polynomials, smooth functions, and sinusoidal functions, in all cases with exponential accuracy, i.e., with an approximation error that decays exponentially in network connectivity. We now proceed to lay the foundation for the development of a framework that allows us to characterize the fundamental limits of deep neural network approximation of entire function classes. But first, we provide a review of relevant literature.

The best-known results on approximation by neural networks are the universal approximation theorems of Hornik [12] and Cybenko [11], 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 Fourier transforms of bounded first moments [35], [36], the nonexistence of

localized approximations [37], a fundamental lower bound on approximation rates [38], [39], and the approximation of smooth or analytic functions [40], [41].

Approximation-theoretic results for networks with multiple hidden layers were obtained in [42], [43] for general functions, in [44] for continuous functions, and for functions together with their derivatives in [45]. In [37] 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 [46] 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 [46] were reported in [47]. Linking the expressivity properties of neural networks to tensor decompositions, [48], [49] 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 [29] that deep ReLU networks can approximate very effectively certain solution families of parametric PDEs depending on a large (possibly infinite) number of parameters. The series of papers [50], [51], [52], [53] constructs and analyzes a deep-learning-based numerical solver for Black-Scholes PDEs.

For survey articles on approximation-theoretic aspects of neural networks, we refer the interested reader to [54] and [55] as well as the very recent [56]. Most closely related to the framework we develop here is the paper by Shaham, Cloninger, and Coifman [57], which shows that for functions that are sparse in specific wavelet frames, the best M-weight approximation rate (see Definition VI.1 below) of three-layer neural networks is at least as large 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 [58], [59]. Our presentation essentially follows [14], [60].

A. Kolmogorov-Donoho Rate Distortion Theory

Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and consider a set of functions $\mathcal{C} \subseteq L^2(\Omega)$, which we will frequently refer to as *function class*. 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

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

Note that here we quantified the approximation error in $L^2(\Omega)$-norm, whereas in the previous section we used the $L^\infty(\Omega)$-norm. While results in terms of $L^\infty(\Omega)$-norm are stronger, we shall employ the $L^2(\Omega)$-norm in order to parallel the Kolmogorov-Donoho framework for nonlinear approximation through dictionaries [14], [15]. We furthermore note that for sets Ω of finite Lebesgue measure $|\Omega|$, the two norms are related through $\|f\|_{L^2(\Omega)} \leq |\Omega|^{1/2}\|f\|_{L^\infty(\Omega)}$. Finally, whenever we talk about compactness and related topological notions, we shall always mean w.r.t. the topology induced by the $L^2(\Omega)$-norm.

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

$$L(\varepsilon, \mathcal{C}) := \min\left\{\ell \in \mathbb{N} : \exists (E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell : \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon\right\}. \tag{12}$$

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

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

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 functions $f \in \mathcal{C}$ such that reconstruction with uniformly bounded error is possible.

Remark IV.2. *The optimal exponent $\gamma^*(\mathcal{C})$ can equivalently be thought of as quantifying the asymptotic behavior of the minimal achievable error for the function class $\mathcal{C}$ with a given code length. Specifically, we have*

$$\gamma^*(\mathcal{C}) = \sup\left\{\gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\big(\varepsilon^{-1/\gamma}\big),\ \varepsilon \to 0\right\} = \sup\left\{\gamma \in \mathbb{R} : \varepsilon(L) \in \mathcal{O}\big(L^{-\gamma}\big),\ L \to \infty\right\}, \tag{13}$$

where

$$\varepsilon(L) := \inf_{(E,D)\in\mathfrak{E}^L\times\mathfrak{D}^L} \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)}.$$

The quantity $\gamma^*(\mathcal{C})$ is closely related to the concept of Kolmogorov-Tikhomirov epsilon entropy a.k.a. metric entropy [61]. We next make this connection explicit.

B. Metric entropy

Most of the discussion in this subsection, which is almost exclusively of review nature, follows very closely [62, Chapter 5]. Consider the metric space $(\mathcal{X}, \rho)$ with $\mathcal{X}$ a nonempty set and $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a distance function. A natural measure for the size of a compact subset $\mathcal{C}$ of $\mathcal{X}$ is given by the number of balls of a fixed radius ε required to cover $\mathcal{C}$, a quantity known as the covering number (for covering radius ε).

Definition IV.3. *[62] Let $(\mathcal{X}, \rho)$ be a metric space. An ε-covering of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \ldots, x_N\} \subseteq \mathcal{C}$ such that for each $x \in \mathcal{C}$, there exists an $i \in \{1, \ldots, N\}$ so that $\rho(x, x_i) \leq \varepsilon$. The ε-covering number $N(\varepsilon; \mathcal{C}, \rho)$ is the cardinality of the smallest ε-covering.*

An ε-covering is a collection of balls of radius ε that cover the set $\mathcal{C}$, i.e.,

$$\mathcal{C} \subseteq \bigcup_{i=1}^{N(\varepsilon; \mathcal{C}, \rho)} B(x_i, \varepsilon),$$

where $B(x_i, \varepsilon)$ is a ball—in the metric ρ—of radius ε centered at x_i. The covering number is nonincreasing in ε, i.e., $N(\varepsilon; \mathcal{C}, \rho) \geq N(\varepsilon'; \mathcal{C}, \rho)$, for all $\varepsilon \leq \varepsilon'$. When the set $\mathcal{C}$ is not finite, the covering number goes to infinity as ε goes to zero. We shall be interested in the corresponding rate of growth, more specifically in the quantity $\log N(\varepsilon; \mathcal{C}, \rho)$ known as the metric entropy of $\mathcal{C}$ with respect to ρ. Recall that $\log$ is to the base 2, hence the unit of metric entropy is "bits". The operational significance of metric entropy follows from the question: What is the minimum number of bits needed to represent any element $x \in \mathcal{C}$ with error—quantified in terms of the distance measure ρ—of at most ε? By what was just developed, the answer to this question is $\lceil \log N(\varepsilon; \mathcal{C}, \rho) \rceil$. Specifically, for a given $x \in \mathcal{X}$, the corresponding encoder $E(x)$ simply identifies the closest ball center x_i and encodes the index i using $\lceil \log N(\varepsilon; \mathcal{C}, \rho) \rceil$ bits. The corresponding decoder D delivers the ball center x_i, which guarantees that the resulting error satisfies $\|D(E(x)) - x\| \leq \varepsilon$.

We proceed with a simple example ([62, Example 5.2]) computing an upper bound on the metric entropy of the interval $\mathcal{C} = [-1, 1]$ in $\mathbb{R}$ with respect to the metric $\rho(x, x') = |x - x'|$. To this end, we divide $\mathcal{C}$ into intervals of length 2ε by setting $x_i = -1 + 2(i-1)\varepsilon$, for $i \in [1, L]$, where $L = \lfloor \frac{1}{\varepsilon} \rfloor + 1$. This guarantees that, for every point $x \in [-1, 1]$, there is an $i \in [1, L]$ such that $|x - x_i| \leq \varepsilon$, which, in turn, establishes

$$N(\varepsilon; \mathcal{C}, \rho) \leq \left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1 \leq \frac{1}{\varepsilon} + 1$$

and hence yields an upper bound on metric entropy according to[2]

$$\log N(\varepsilon; \mathcal{C}, \rho) \leq \log \left(\frac{1}{\varepsilon} + 1 \right) \asymp \log(\varepsilon^{-1}), \quad \text{as } \varepsilon \to 0. \tag{14}$$

[2]The notation $f(\varepsilon) \asymp g(\varepsilon)$, as $\varepsilon \to 0$, means that there are constants $c, C, \varepsilon_0 > 0$ such that $cf(\varepsilon) \leq g(\varepsilon) \leq Cf(\varepsilon)$, for all $\varepsilon \leq \varepsilon_0$. For ease of exposition, we shall usually omit the qualifier $\varepsilon \to 0$.

This result can be generalized to the d-dimensional unit cube to yield $\log(N(\varepsilon;\mathcal{C},\rho)) \leq d\log(1/\varepsilon+1) \asymp d\log(\varepsilon^{-1})$. In order to show that the upper bound (14) correctly reflects metric entropy scaling for $\mathcal{C}=[-1,1]$ with respect to $\rho(x,x')=|x-x'|$, we would need a lower bound on $N(\varepsilon;\mathcal{C},\rho)$ that exhibits the same scaling (in ε) behavior. A systematic approach to establishing lower bounds on metric entropy is through the concept of packing, which will be introduced next.

We start with the definition of the packing number of a compact set $\mathcal{C}$ in a metric space $(\mathcal{X},\rho)$.

Definition IV.4. *[62, Definition 5.4] Let $(\mathcal{X},\rho)$ be a metric space. An ε-packing of a compact set $\mathcal{C}\subseteq\mathcal{X}$ with respect to the metric ρ is a set $\{x_1,\ldots,x_N\}\subseteq\mathcal{C}$ such that $\rho(x_i,x_j)>\varepsilon$, for all distinct i,j. The ε-packing number $M(\varepsilon;\mathcal{X},\rho)$ is the cardinality of the largest ε-packing.*

An ε-packing is a collection of nonintersecting balls of radius $\varepsilon/2$ and centered at elements in $\mathcal{X}$. Although different, the covering number and the packing number provide essentially the same measure of size of a set as formalized next.

Lemma IV.5. *[62, Lemma 5.5] Let $(\mathcal{X},\rho)$ be a metric space and $\mathcal{C}$ a compact set in $\mathcal{X}$. For all $\varepsilon>0$, the packing and the covering number are related according to*

$$M(2\varepsilon;\mathcal{C},\rho)\leq N(\varepsilon;\mathcal{C},\rho)\leq M(\varepsilon;\mathcal{C},\rho).$$

Proof. [62], [63] First, choose a minimal ε-covering and a maximal 2ε-packing of $\mathcal{C}$. Since no two centers of the 2ε-packing can lie in the same ball of the ε-covering, it follows that $M(2\varepsilon;\mathcal{C},\rho)\leq N(\varepsilon;\mathcal{C},\rho)$. To establish $N(\varepsilon;\mathcal{C},\rho)\leq M(\varepsilon;\mathcal{C},\rho)$, we note that, given a maximal packing $M(\varepsilon;\mathcal{C},\rho)$, for any $x\in\mathcal{C}$, we have the center of at least one of the balls in the packing within distance less than ε. If this were not the case, we could add another ball to the packing thereby violating its maximality. This maximal packing hence also provides an ε-covering and since $N(\varepsilon;\mathcal{C},\rho)$ is a minimal covering, we must have $N(\varepsilon;\mathcal{C},\rho)\leq M(\varepsilon;\mathcal{C},\rho)$. □

We now return to the example in which we computed an upper bound on the metric entropy of $\mathcal{C}=[-1,1]$ with respect to $\rho(x,x')=|x-x'|$ and show how Lemma IV.5 can be employed to establish the scaling behavior of metric entropy. To this end, we simply note that the points $x_i=-1+2(i-1)\varepsilon$, $i\in[1,L]$, are separated according to $|x_i-x_j|=2\varepsilon>\varepsilon$, for all $i\neq j$, which implies that $M(\varepsilon;\mathcal{C},|\cdot|)\geq L=\lfloor 1/\varepsilon\rfloor+1\geq\frac{1}{\varepsilon}$. Combining this with the upper bound (14) and Lemma IV.5, we obtain $\log N(\varepsilon;\mathcal{C},|\cdot|)\asymp\log(\varepsilon^{-1})$. Likewise, it can be established that $\log N(\varepsilon;\mathcal{C},\|\cdot\|)\asymp d\log(\varepsilon^{-1})$ for the d-dimensional unit cube. This illustrates how an explicit construction of a packing set can be used to determine the scaling behavior of metric entropy.

We next formalize the notion that metric entropy is determined by the volume of the corresponding covering balls. Specifically, the following result establishes a relationship between a certain volume ratio and metric entropy.

Lemma IV.6. *[62, Lemma 5.7] Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^d$, and let $\mathcal{B}$ and $\mathcal{B}'$ be their corresponding unit balls, i.e., $\mathcal{B} = \{x \in \mathbb{R}^d | \|x\| \leq 1\}$ and $\mathcal{B}' = \{x \in \mathbb{R}^d | \|x\|' \leq 1\}$. Then, the ε-covering number of $\mathcal{B}$ in the $\|\cdot\|'$-norm satisfies*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{vol(\mathcal{B})}{vol(\mathcal{B}')} \leq N(\varepsilon; \mathcal{B}, \|\cdot\|') \leq \frac{vol(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}')}{vol(\mathcal{B}')}. \tag{15}$$

Proof. [62] Let $\{x_1, \ldots, x_{N(\varepsilon;\mathcal{B},\|\cdot\|')}\}$ be an ε-covering of $\mathcal{B}$ in $\|\cdot\|'$-norm. Then, we have

$$\mathcal{B} \subseteq \bigcup_{j=1}^{N(\varepsilon;\mathcal{B},\|\cdot\|')} \{x_j + \varepsilon\mathcal{B}'\},$$

which implies $vol(\mathcal{B}) \leq N(\varepsilon; \mathcal{B}, \|\cdot\|')\, \varepsilon^d\, vol(\mathcal{B}')$, thus establishing the lower bound in (15). The upper bound is obtained by starting with a maximal ε-packing $\{x_1, \ldots, x_{M(\varepsilon;\mathcal{B},\|\cdot\|')}\}$ of $\mathcal{B}$ in the $\|\cdot\|'$-norm. The balls $\{x_j + \frac{\varepsilon}{2}\mathcal{B}', j = 1, \ldots, M(\varepsilon; \mathcal{B}, \|\cdot\|')\}$ are all disjoint and contained within $\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}'$. We can therefore conclude that

$$\sum_{j=1}^{M(\varepsilon;\mathcal{B},\|\cdot\|')} vol\Big(x_j + \frac{\varepsilon}{2}\mathcal{B}'\Big) \leq vol\Big(\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}'\Big),$$

and hence

$$M(\varepsilon; \mathcal{B}, \|\cdot\|')\, vol\Big(\frac{\varepsilon}{2}\mathcal{B}'\Big) \leq vol\Big(\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}'\Big).$$

Finally, we have $vol(\frac{\varepsilon}{2}\mathcal{B}') = (\frac{\varepsilon}{2})^d vol(\mathcal{B}')$ and $vol(\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}') = (\frac{\varepsilon}{2})^d vol(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}')$, which, together with $M(\varepsilon; \mathcal{B}, \|\cdot\|') \geq N(\varepsilon; \mathcal{B}, \|\cdot\|')$ due to Lemma IV.5, yields the upper bound in (15). □

This result now allows us to establish the scaling of the metric entropy of unit balls in terms of their own norm, thus yielding a measure of the massiveness of unit balls in d-dimensional spaces. Specifically, we set $\mathcal{B}' = \mathcal{B}$ in Lemma IV.6 and get

$$vol\left(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}'\right) = vol\left(\left(\frac{2}{\varepsilon} + 1\right)\mathcal{B}\right) = \left(\frac{2}{\varepsilon} + 1\right)^d vol(\mathcal{B}),$$

which when used in (15) yields $N(\varepsilon; \mathcal{B}, \|\cdot\|) \asymp \varepsilon^{-d}$ and hence results in metric entropy scaling according to $\log(N(\varepsilon; \mathcal{B}, \|\cdot\|)) \asymp d \log(\varepsilon^{-1})$. Particularizing this result to the unit ball $\mathcal{B}_\infty^d = [-1, 1]^d$ and the metric $\|\cdot\|_\infty$, we recover the result of our direct analysis in the example above.

So far we have been concerned with the metric entropy of subsets of $\mathbb{R}^d$. We now proceed to analyzing the metric entropy of function classes, which will eventually allow us to establish the desired connection between the optimal exponent $\gamma^*(\mathcal{C})$ and metric entropy. We begin with the simple one-parameter function class considered in [62, Example 5.9] and follow closely the exposition in [62]. For a fixed θ, define the real-valued function $f_\theta(x) = 1 - e^{-\theta x}$, and consider the class

$$\mathcal{P} = \{f_\theta : [0, 1] \to \mathbb{R} \,|\, \theta \in [0, 1]\}.$$

The set $\mathcal{P}$ constitutes a metric space under the sup-norm given by $\|f-g\|_{L^\infty([0,1])} = \sup_{x\in[0,1]} |f(x)-g(x)|$. We show that the covering number of $\mathcal{P}$ satisfies

$$1 + \left\lfloor \frac{1-1/e}{2\varepsilon} \right\rfloor \leq N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \leq \frac{1}{2\varepsilon} + 2,$$

which leads to the scaling behavior $N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \asymp \varepsilon^{-1}$ and hence to metric entropy scaling according to $\log(N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])})) \asymp \log(\varepsilon^{-1})$. We start by establishing the upper bound. For given $\varepsilon \in [0,1]$, set $T = \lfloor \frac{1}{2\varepsilon} \rfloor$, and define the points $\theta_i = 2\varepsilon i$, for $i = 0, 1, \ldots, T$. By also adding the point $\theta_{T+1} = 1$, we obtain a collection of $T+2$ points $\{\theta_0, \theta_1, \ldots, \theta_{T+1}\}$ in $[0,1]$. We show that the associated functions $\{f_{\theta_0}, f_{\theta_1}, \ldots, f_{\theta_{T+1}}\}$ form an ε-covering for $\mathcal{P}$. Indeed, for any $f_\theta \in \mathcal{P}$, we can find some θ_i in the covering such that $|\theta - \theta_i| \leq \varepsilon$. We then have

$$\|f_\theta - f_{\theta_i}\|_{L^\infty([0,1])} = \max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| \leq |\theta - \theta_i|,$$

where we used, for $\theta < \theta_i$,

$$\begin{aligned} \max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| &= \max_{x\in[0,1]} (e^{-\theta x} - e^{-\theta_i x}) = \max_{x\in[0,1]} e^{-\theta x}(1 - e^{-(\theta_i - \theta)x}) \leq \max_{x\in[0,1]} (1 - e^{-(\theta_i-\theta)x}) \\ &\leq \max_{x\in[0,1]} (\theta_i - \theta)x \leq \theta_i - \theta = |\theta - \theta_i|, \end{aligned}$$

as a consequence of $1 - e^{-x} \leq x$, for $x \in [0,1]$, which is easily verified by noting that the function $g(x) = 1 - e^{-x} - x$ satisfies $g(0) = 0$ and $g'(x) \leq 0$, for $x \in [0,1]$. The case $\theta > \theta_i$ follows similarly. In summary, we have shown that $N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \leq T + 2 \leq \frac{1}{2\varepsilon} + 2$.

In order to derive the lower bound, we first bound the packing number from below and then use Lemma IV.5. We start by constructing an explicit packing as follows. Set $\theta_0 = 0$ and define $\theta_i = -\log(1 - \varepsilon i)$, for all i such that $\theta_i \leq 1$. The largest index T such that this holds is given by $T = \lfloor \frac{1-1/e}{\varepsilon} \rfloor$. Moreover, note that for all i, j with $i \neq j$, we have $\|f_{\theta_i} - f_{\theta_j}\|_{L^\infty([0,1])} \geq |f_{\theta_i}(1) - f_{\theta_j}(1)| = |\varepsilon(i-j)| \geq \varepsilon$. We can therefore conclude that $M(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \geq \lfloor \frac{1-1/e}{\varepsilon} \rfloor + 1$, and hence, due to the lower bound in Lemma IV.5,

$$N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \geq M(2\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \geq \left\lfloor \frac{1-1/e}{2\varepsilon} \right\rfloor + 1,$$

as claimed. We have thus established that the function class $\mathcal{P}$ has metric entropy scaling according to

$$\log(N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])})) \asymp \log(1/\varepsilon), \text{ as } \varepsilon \to 0.$$

This rate is typical for one-parameter function classes.

We now turn our attention to richer function classes and start by considering Lipschitz functions on the d-dimensional unit cube, meaning real-valued functions on $[0,1]^d$ such that

$$|f(x) - f(y)| \leq L\|x - y\|_\infty, \qquad \text{for all} \quad x, y \in [0,1]^d.$$

This class, denoted as $\mathcal{F}_L([0,1]^d)$, has metric entropy scaling [64], [62]

$$\log N(\varepsilon; \mathcal{F}_L, \|\cdot\|_{L^\infty([0,1]^d)}) \asymp (L/\varepsilon)^d. \tag{16}$$

Contrasting the exponential dependence of metric entropy in (16) on the ambient dimension d to the linear dependence we identified earlier for simpler sets such as unit balls in $\mathbb{R}^d$, where we had

$$\log N(\varepsilon; \mathcal{B}, \|\cdot\|_\infty) \asymp d \log(\varepsilon^{-1}),$$

shows that $\mathcal{F}_L([0,1]^d)$ is significantly more massive.

We are now ready to relate the optimal exponent $\gamma^*(\mathcal{C})$ in Definition IV.1 to metric entropy scaling. All the examples of metric entropy scaling we have seen exhibit a behavior that fits the law $\log(N(\varepsilon; \mathcal{C}, \|\cdot\|)) \asymp \varepsilon^{-1/\gamma}$ or $\log(N(\varepsilon; \mathcal{C}, \|\cdot\|)) \asymp \varepsilon^{-1/\gamma} \log(\varepsilon^{-1})^\beta$. The optimal exponent is hence a crude measure of growth insensitive to $\log$-factors or similar factors that are dominated by the growth of $\varepsilon^{-1/\gamma}$.

While we restrict ourselves to the approximation of functions on Euclidean domains, the framework described in this section can be extended to functions on manifolds (see e.g. [65]). As such, an interesting direction for future research would be the extension of the deep neural network approximation theory developed in this paper to functions on manifolds. First results on the neural network approximation of functions on manifolds have been reported in [57], [13], [66]. For further reading on the general subject of function approximation on manifolds, we recommend [67] and references therein.

V. Approximation with Dictionaries

We now show how Kolmogorov-Donoho rate-distortion theory can be put to work in the context of optimal approximation with dictionaries. Again, this subsection is of review nature. We start with a brief discussion of basics on optimal approximation in Hilbert spaces. Specifically, we shall consider two types of approximation, namely linear and nonlinear.

Let $\mathcal{H}$ be a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|_{\mathcal{H}}$ and let e_k, $k = 1, 2, \ldots$, be an orthonormal basis for $\mathcal{H}$. For linear approximation, we use the linear space $\mathcal{H}_M := \text{span}\{e_k : 1 \leq k \leq M\}$ to approximate a given element $f \in \mathcal{H}$. We measure the approximation error by

$$E_M(f) := \inf_{g \in \mathcal{H}_M} \|f - g\|_{\mathcal{H}}.$$

In nonlinear approximation, we consider best M-term approximation, which replaces $\mathcal{H}_M$ by the set Σ_M consisting of all elements $g \in \mathcal{H}$ that can be expressed as

$$g = \sum_{k \in \Lambda} c_k e_k,$$

where $\Lambda \subseteq \mathbb{N}$ is a set of indices with $|\Lambda| \leq M$. Note that, in contrast to $\mathcal{H}_M$, the set Σ_M is not a linear space as a linear combination of two elements in Σ_M will, in general, need $2M$ terms in its representation by the e_k. Analogous to E_M, we define the error of best M-term approximation

$$\Gamma_M(f) := \inf_{g \in \Sigma_M} \|f - g\|_{\mathcal{H}}.$$

The key difference between linear and nonlinear approximation resides in the fact that in nonlinear approximation, we can choose the M elements e_k participating in the approximation of f freely from the entire orthonormal basis whereas in linear approximation we are constrained to the first M elements. A classical example for linear approximation is the approximation of periodic functions by the Fourier series elements corresponding to the M lowest frequencies (assuming natural ordering of the dictionary). This approach clearly leads to poor approximation if the function under consideration consists of high-frequency components. In contrast, in nonlinear approximation we would seek the M frequencies that yield the smallest approximation error. In summary, it is clear that (nonlinear) best M-term approximation can achieve smaller approximation error than linear M-term approximation.

We shall consider nonlinear approximation in arbitrary, possibly redundant, dictionaries, i.e., in frames [68], and will exclusively be interested in the case $\mathcal{H} = L^2(\Omega)$, in particular the approximation error will be measured in terms of $L^2(\Omega)$-norm. Specifically, let $\mathcal{C}$ be a set of functions in $L^2(\Omega)$ and consider a countable family of functions $\mathcal{D} := (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$, termed *dictionary*.

We consider the *best M-term approximation error* of $f \in \mathcal{C}$ in $\mathcal{D}$ defined as follows.

Definition V.1. *[58] Given $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, a function class $\mathcal{C} \subseteq L^2(\Omega)$, and a dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma_M^{\mathcal{D}}(f) := \inf_{\substack{I_{f,M}\subseteq\mathbb{N},\\ |I_{f,M}|=M,(c_i)_{i\in I_{f,M}}}} \left\| f - \sum_{i\in I_{f,M}} c_i\varphi_i \right\|_{L^2(\Omega)}. \tag{17}$$

We call $\Gamma_M^{\mathcal{D}}(f)$ the best M-term approximation error of f in $\mathcal{D}$. *Every $f_M = \sum_{i\in I_{f,M}} c_i\varphi_i$ attaining the infimum in* (17) *is referred to as a* best M-term approximation *of f in $\mathcal{D}$. The supremal $\gamma > 0$ such that*

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

will be denoted by $\gamma^*(\mathcal{C},\mathcal{D})$. We say that the best M-term approximation rate of $\mathcal{C}$ in the dictionary $\mathcal{D}$ *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 [59], as well as α-cartoon-like functions [69]. A wealth of structured dictionaries $\mathcal{D}$ is provided by the area of applied harmonic analysis, starting with wavelets [70], followed by ridgelets [39], curvelets [71], shearlets [72], parabolic molecules [73], and most generally α-molecules [69], which include all previously named dictionaries as special cases. Further examples are Gabor frames [17], Wilson bases [74], and wave atoms [18].

The best M-term approximation rate $\gamma^*(\mathcal{C},\mathcal{D})$ according to Definition V.1 quantifies how difficult it is to approximate a given function class $\mathcal{C}$ in a fixed dictionary $\mathcal{D}$. 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 $\mathcal{D}$. To answer this question, we first note that for every dense (and countable) $\mathcal{D}$, for any given $f \in \mathcal{C}$, by density of $\mathcal{D}$, there exists a single dictionary element that approximates f to within arbitrary accuracy thereby effectively realizing a 1-term approximation for arbitrary approximation error ε. Formally, this can be expressed through $\gamma^*(\mathcal{C},\mathcal{D}) = \infty$. Identifying this single dictionary

element or, more generally, the M elements participating in the best M-term approximation is in general, however, practically infeasible as it entails searching through the infinite set $\mathcal{D}$ and requires 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 [15]. Here, the basic idea is to restrict 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. We formalize this under the name of effective best M-term approximation as follows.

Definition V.2. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\mathcal{C} \subseteq L^2(\Omega)$ be compact, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$. We define for $M \in \mathbb{N}$ and π a polynomial*

$$\varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M) := \sup_{f\in\mathcal{C}} \inf_{\substack{I_{f,M}\subseteq\{1,2,\dots,\pi(M)\},\\ |I_{f,M}|=M,\,|c_i|\leq\pi(M)}} \left\| f - \sum_{i\in I_{f,M}} c_i\varphi_i \right\|_{L^2(\Omega)} \tag{18}$$

and

$$\gamma^{*,\mathit{eff}}(\mathcal{C},\mathcal{D}) := \sup\{\gamma \geq 0\colon \exists \text{ polynomial } \pi \text{ s.t. } \varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M) \in \mathcal{O}(M^{-\gamma}),\ M\to\infty\}. \tag{19}$$

We refer to $\gamma^{*,\mathit{eff}}(\mathcal{C},\mathcal{D})$ as the effective best M-term approximation rate of $\mathcal{C}$ in the dictionary $\mathcal{D}$.

Note that we required the coefficients c_i in the approximant in Definition V.2 to be polynomially bounded in M. This condition, not present in [14], [60] and easily met for generic $\mathcal{C}$ and $\mathcal{D}$, is imposed for technical reasons underlying the transference results in Section VII. Strictly speaking—relative to [14], [60]—we hence get a subtly different notion of approximation rate. Exploring the implications of this difference is certainly worthwhile, but deemed beyond the scope of this paper.

We next present a central result in best M-term approximation theory stating that for compact $\mathcal{C} \subseteq L^2(\Omega)$, the effective best M-term approximation rate in any dictionary $\mathcal{D}$ 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.3. *[14], [60] Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. The effective best M-term approximation rate of the function class $\mathcal{C} \subseteq L^2(\Omega)$ in the dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ satisfies*

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

In light of this result the following definition is natural (see also [60]).

Definition V.4. *(Kolmogorov-Donoho optimality) Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. If the effective best M-term approximation rate of the function class $\mathcal{C} \subseteq L^2(\Omega)$ in the dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ satisfies*

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

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

As the ideas underlying the proof of Theorem V.3 are essential ingredients in the development of a kindred theory of best M-weight approximation rates for neural networks, we present a detailed proof, which is similar to that in [60]. We perform, however, some minor technical modifications with an eye towards rendering the proof a suitable genesis for the new theory of best M-weight approximation with neural networks, developed in the next section. The spirit of the proof is to construct, for every given $M \in \mathbb{N}$ an encoder that, for each $f \in \mathcal{C}$, maps the indices of the dictionary elements participating in the effective best M-term approximation[3] of f, along with the corresponding coefficients c_i, to a bitstring. This bitstring needs to be of sufficient length for the decoder to be able to reconstruct an approximation to f with an error which is of the same order as that of the best M-term approximation we started from. As elucidated in the proof, this can be accomplished while ensuring that the length of the bitstring is proportional to $M\log(M)$, which upon noting that $\varepsilon = M^{-\gamma}$ implies $M = \varepsilon^{-1/\gamma}$, establishes optimality.

Proof of Theorem V.3. The proof will be based on showing that for every $\gamma \in \mathbb{R}_+$ the following Implication (I) holds: Assume that there exist a constant $C > 0$ and a polynomial π such that for every $M \in \mathbb{N}$, the following holds: For every $f \in \mathcal{C}$, there are an index set $I_{f,M} \subseteq \{1, 2, \ldots, \pi(M)\}$ and coefficients $(c_i)_{i\in I_{f,M}} \subseteq \mathbb{R}$ with $|c_i| \leq \pi(M)$ so that

$$\Big\|f - \sum_{i\in I_{f,M}} c_i\varphi_i\Big\|_{L^2(\Omega)} \leq CM^{-\gamma}. \tag{20}$$

This implies the existence of a constant $C' > 0$ such that for every $M \in \mathbb{N}$, there is an encoder-decoder pair $(E_M, D_M) \in \mathfrak{E}^{\ell(M)} \times \mathfrak{D}^{\ell(M)}$ with $\ell(M) \leq C'M\log(M)$ and

$$\|f - D_M(E_M(f))\|_{L^2(\Omega)} \leq C'M^{-\gamma}. \tag{21}$$

The implication will be proven by explicit construction. For a given $f \in \mathcal{C}$, we pick an M-term approximation according to (20) and encode the associated index set $I_{f,M}$ and weights c_i as follows. First, note that owing to $|I_{f,M}| \leq \pi(M)$, each index in $I_{f,M}$ can be represented by at most $C_\pi\log(M)$ bits; this results in a total of $C_\pi M\log(M)$ bits needed to encode the indices of all dictionary elements participating in the M-term approximation. The encoder and the decoder are assumed to know C_π, which allows stacking of the binary representations of the indices such that the decoder can read them off uniquely from the sequence of their binary representations.

We proceed to the encoding of the coefficients c_i. First, note that even though the c_i are bounded (namely, polynomially in M) by assumption, we did not impose bounds on the norms of the dictionary elements $\{\varphi_i\}_{i\in I_{f,M}}$ participating in the M-term approximation under consideration. Hence, we can not, in general, expect to be able to control the approximation error incurred by reconstructing f from quantized c_i. We can get around this by performing a Gram-Schmidt orthogonalization on the dictionary elements $\{\varphi_i\}_{i\in I_{f,M}}$ and, as will be seen later, using the fact

[3]Note that as we have an infimum in (18) an effective best M-term approximation need not exist, but we can pick an M-term approximation that yields an error arbitrarily close to the infimum.

that the function class $\mathcal{C}$ was assumed to be compact. Specifically, this Gram-Schmidt orthogonalization yields a set of functions $\{\tilde{\varphi}_i\}_{i\in\tilde{I}_{f,\widetilde{M}}}$, with $\widetilde{M} \leq M$, that has the same span as $\{\varphi_i\}_{i\in I_{f,M}}$. Next, we define (implicitly) the coefficients $\tilde{c}_i$ according to

$$\sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i = \sum_{i\in I_{f,M}} c_i\varphi_i. \tag{22}$$

Now, note that

$$\left\| \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i \right\|^2_{L^2(\Omega)} = \left\| f - (f - \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i) \right\|^2_{L^2(\Omega)} \leq \|f\|^2_{L^2(\Omega)} + \left\| f - \sum_{i\in I_{f,M}} c_i\varphi_i \right\|^2_{L^2(\Omega)}.$$

Making use of the orthonormality of the $\tilde{\varphi}_i$, we can conclude that

$$\sum_{i\in\tilde{I}_{f,\widetilde{M}}} |\tilde{c}_i|^2 \leq \sup_{f\in\mathcal{C}} \|f\|^2_{L^2(\Omega)} + C^2M^{-2\gamma}.$$

As $\mathcal{C}$ is compact by assumption, we have $\sup_{f\in\mathcal{C}} \|f\|^2_{L^2(\Omega)} < \infty$, which establishes that the coefficients $\tilde{c}_i$ are uniformly bounded. This, in turn, allows us to quantize them, specifically, we shall round the $\tilde{c}_i$ to integer multiples of $M^{-(\gamma+1/2)}$, and denote the resulting rounded coefficients by $\hat{c}_i$. As the $\tilde{c}_i$ are uniformly bounded, this results in a number of quantization levels that is proportional to $M^{(\gamma+1/2)}$. The number of bits needed to store the binary representations of the quantized coefficients is therefore proportional to $M\log(M)$. Again, the proportionality constant is assumed known to encoder and decoder, which allows us to stack the binary representations of the quantized coefficients in a uniquely decodable manner. The resulting bitstring is then appended to the bitstring encoding the indices of the participating dictionary elements. We finally note that the specific choice of the exponent $\gamma + 1/2$ is informed by the upper bound on the reconstruction error we are allowed, this will be made explicit below in the description of the decoder.

In summary, we have mapped the function f to a bitstring of length $\mathcal{O}(M\log(M))$. The decoder is presented with this bitstring and reconstructs an approximation to f as follows. It first reads out the indices of the set $I_{f,M}$ and the quantized coefficients $\hat{c}_i$. Recall that this is uniquely possible. Next, the decoder performs a Gram-Schmidt orthonormalization on the set of dictionary elements indexed by $I_{f,M}$. The error resulting from reconstructing the function f from the quantized coefficients $\hat{c}_i$ rather than the exact coefficients $\tilde{c}_i$ can be bounded according to

$$\begin{aligned}
\left\| f - \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \hat{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} &= \left\| f - \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i + \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i - \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \hat{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} \\
&\leq \left\| f - \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} + \left\| \sum_{i\in\tilde{I}_{f,\widetilde{M}}} (\tilde{c}_i - \hat{c}_i)\tilde{\varphi}_i \right\|_{L^2(\Omega)} \\
&= \left\| f - \sum_{i\in\tilde{I}_{f,\widetilde{M}}} \tilde{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} + \left(\sum_{i\in\tilde{I}_{f,\widetilde{M}}} |\tilde{c}_i - \hat{c}_i|^2 \right)^{1/2},
\end{aligned} \tag{23}$$

where in the last step we again exploited the orthonormality of the $\tilde{\varphi}_i$. Next, note that due to the choice of the quantizer resolution, we have $|\tilde{c}_i - \hat{c}_i|^2 \leq C'' M^{-2\gamma-1}$ for some constant C''. With $\widetilde{M} \leq M$ this yields

$$\sum_{i \in \tilde{I}_{f,\widetilde{M}}} |\tilde{c}_i - \hat{c}_i|^2 \leq C'' M^{-2\gamma}.$$

Combining (20), (22), and (23), we obtain

$$\left\| f - \sum_{i \in \tilde{I}_{f,\widetilde{M}}} \hat{c}_i \tilde{\varphi}_i \right\|_{L^2(\Omega)} \leq C' M^{-\gamma},$$

for some constant C'. As the length of the bitstring used in this construction is proportional to $M \log(M)$, the claim (21) is established.

Now, we note that the antecedent of Implication (I) holds for all $\gamma < \gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D})$. Assume next, towards a contradiction, that the antecedent holds for a $\gamma > \gamma^*(\mathcal{C})$. This would imply that for any $\gamma' < \gamma$,

$$\inf_{(E,D) \in \mathfrak{E}^L \times \mathfrak{D}^L} \sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \in \mathcal{O}\big(L^{-\gamma'}\big),\ L \to \infty. \tag{24}$$

In particular, (24) would hold for some $\gamma' > \gamma^*(\mathcal{C})$ which, owing to (13) stands in contradiction to the definition of $\gamma^*(\mathcal{C})$. This completes the proof. □

Space		$\mathcal{C}$	Optimal dictionary	$\gamma^*(\mathcal{C})$	
L^2-Sobolev	$W_2^m([0,1])$	$\mathcal{U}(W_2^m([0,1]))$	Fourier/Wavelet basis	m	[75, Sec. 14.2]
Hölder	$C^\alpha([0,1])$	$\mathcal{U}(C^\alpha([0,1]))$	Wavelet basis	α	[75, Sec. 14.2]
Bump Algebra	$B^1_{1,1}([0,1])$	$\mathcal{U}(B^1_{1,1}([0,1]))$	Wavelet basis	1	[75, Sec. 14.2]
Bounded Variation	$BV([0,1])$	$\mathcal{U}(BV([0,1]))$	Haar basis	1	[75, Sec. 14.2]
L^p-Sobolev[4]	$W_p^m(\Omega)$	$\mathcal{U}(W_p^m(\Omega))$	Wavelet frame	$\frac{m}{d}$	[76, Thm. 1.3]
Besov[5]	$B^m_{p,q}(\Omega)$	$\mathcal{U}(B^m_{p,q}(\Omega))$	Wavelet frame	$\frac{m}{d}$	[76, Thm. 1.3]
Modulation[6]	$M^s_{p,p}(\mathbb{R}^d)$	$\mathcal{U}(M^s_{p,p}(\mathbb{R}^d))$	Wilson basis	$(\frac{1}{p} - \frac{1}{2} + \frac{2s}{d})^{-1}$	[77, Thm. 4.4]
Cartoon functions[7]		$\mathcal{E}^\beta([-\frac{1}{2}, \frac{1}{2}]^d)$	α-Curvelet frame[8]	$\frac{\beta(d-1)}{2}$	[23]

Table 1: Optimal exponents and corresponding optimal dictionaries. $\mathcal{U}(X) = \{f \in X : \|f\|_X \leq 1\}$ denotes the unit ball in the space X and $\Omega \subseteq \mathbb{R}^d$ is a Lipschitz domain. Recall that compactness of these unit balls is w.r.t. L^2-norm.

[4] $p \in [1, \infty]$, $m > d(1/p - 1/2)_+$

[5] $p, q \in (0, \infty]$, $m > d(1/p - 1/2)_+$

[6] $1 < p < 2$, $s \in \mathbb{R}_+$

[7] This is actually a set of functions and not a (unit) ball in a Banach space.

[8] For $d = 2$, see [78].

The optimal exponent $\gamma^*(\mathcal{C})$ is known for various function classes such as unit balls in Besov spaces $B^m_{p,q}(\mathbb{R}^d)$ with $p, q \in (0, \infty]$ and $m > d(1/p - 1/2)_+$, where $\gamma^*(\mathcal{C}) = m/d$ (see [76]), and unit balls in (polynomially) weighted modulation spaces $M^s_{p,p}(\mathbb{R}^d)$ with $p \in (1, 2)$ and $s \in \mathbb{R}_+$, where $\gamma^*(\mathcal{C}) = (\frac{1}{p} - \frac{1}{2} + \frac{2s}{d})^{-1}$ (see [77]). A further example is the set of β-cartoon-like functions, which are β-smooth on some bounded d-dimensional domain with sufficiently smooth boundary and zero otherwise. Here, we have $\gamma^*(\mathcal{C}) = \beta(d-1)/2$ (see [79], [78], [23]). These examples along with additional ones are summarized in Table 1. For an extensive summary of metric entropy results and techniques for their derivation, we also refer to [64].

We conclude this section with general remarks on certain formal aspects of the Kolmogorov-Donoho rate-distortion framework. First, we note that for the set $\mathcal{C} \subseteq L^2(\Omega)$ to have a well-defined optimal exponent it must be relatively compact[9]. This follows from the fact that the set over which the minimum in the definition (12) of $L(\varepsilon, \mathcal{C})$ is taken must be nonempty for every $\varepsilon \in (0, \infty)$. To see this, note that every length-$L(\varepsilon, \mathcal{C})$ encoder-decoder pair induces an ε-covering of $\mathcal{C}$ with at most $2^{L(\varepsilon,\mathcal{C})}$ balls (and ball centers $\{D(E(f))\}_{f\in\mathcal{C}}$). It hence follows that $\mathcal{C}$ must be totally bounded and thus relatively compact as a consequence of $L^2(\Omega)$ being a complete metric space [80, Thm. 45.1].

As shown in the proof of Theorem V.3, effective best M-term approximations construct encoder-decoder pairs and thereby induce ε-coverings. By the arguments just made, this implies that also $\gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D})$ is well-defined only for compact function classes $\mathcal{C}$.

A consequence of the compactness requirement on $\mathcal{C}$ is that the spaces in Table 1 either consist of functions on bounded domains or, in the case of modulation spaces, are equipped with a weighted norm. In order to provide intuition on why this must be so, let us consider a function space $(X, \|\cdot\|_X)$ with $X \subseteq L^2(\mathbb{R}^d)$ and $\|\cdot\|_X$ translation invariant. Take $\varepsilon > 0$ and $f \in X$ with $\|f\|_X = 1$ and choose $C > 0$ such that $\|f\|_{L^2([-C,C]^d)} > \frac{4}{5}\|f\|_{L^2(\mathbb{R}^d)}$. Now, consider the family of translates of f given by $f_i(x) := f(x - 2Ci)$, $i \in \mathbb{Z}^d$, and note that $\|f_i\|_X = 1$ for all $i \in \mathbb{Z}^d$ by translation invariance of $\|\cdot\|_X$. Furthermore, we have

$$\|f_i\|_{L^2([-C,C]^d)} = \left(\|f_i\|^2_{L^2(\mathbb{R}^d)} - \|f_i\|^2_{L^2(\mathbb{R}^d\setminus[-C,C]^d)}\right)^{\frac{1}{2}} \leq \left(\|f\|^2_{L^2(\mathbb{R}^d)} - \|f\|^2_{L^2([-C,C]^d)}\right)^{\frac{1}{2}} < \tfrac{3}{5}\|f\|_{L^2(\mathbb{R}^d)}$$

for all $i \in \mathbb{Z}^d\setminus\{0\}$ by construction. This, in turn, implies

$$\|f_i - f_j\|_{L^2(\mathbb{R}^d)} = \|f_{i-j} - f\|_{L^2(\mathbb{R}^d)} \geq \|f_{i-j} - f\|_{L^2([-C,C]^d)} > \tfrac{1}{5}\|f\|_{L^2(\mathbb{R}^d)} \tag{25}$$

for all $i, j \in \mathbb{Z}^d$, with $i \neq j$, by the reverse triangle inequality. As such no ε-ball (w.r.t. $L^2(\mathbb{R}^d)$-norm) with $\varepsilon \leq \frac{1}{10}\|f\|_{L^2(\mathbb{R}^d)}$ can contain more than one of the infinitely many $(f_i)_{i\in\mathbb{Z}^d}$ which are, however, all contained in the unit ball $\mathcal{U}(X)$ of the space $(X, \|\cdot\|_X)$. This implies that $\mathcal{U}(X)$ cannot be totally bounded and thereby not relatively compact (w.r.t. $L^2(\mathbb{R}^d)$-norm). Somewhat nonchalantly speaking, for spaces equipped with translation-invariant norms this issue can be avoided by considering functions that live on a bounded domain, which ensures that

[9]For the sake of simplicity, we assume, however, compactness throughout even though relative compactness (i.e. having a compact closure) would be sufficient.

(25) pertains only to a finite number of translates. Alternatively, for spaces of functions living on unbounded domains once can consider weighted norms that are not translation invariant. Here, the weighting effectively constrains the functions to a bounded domain.

The less restrictive concept of best M-term approximation rate $\gamma^*(\mathcal{C},\mathcal{D})$ (see Definition V.1) is, in apparent contrast, often studied for noncompact function classes $\mathcal{C}$.

In [75, Sec. 15.2] a condition for $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$ and $\gamma^*(\mathcal{C},\mathcal{D})$ to coincide is presented. Specifically, this condition, referred to as tail compactness, is expressed as follows. Let $\mathcal{C} \subseteq L^2(\Omega)$ be bounded and let $\mathcal{D} = \{\varphi_i\}_{i\in\mathbb{N}}$ be an ordered orthonormal basis for $\mathcal{C}$. We say that tail compactness holds if there exist $C,\beta > 0$ such that for all $N \in \mathbb{N}$,

$$\sup_{f\in\mathcal{C}} \left\| f - \sum_{i=1}^{N} \langle f,\varphi_i\rangle \varphi_i \right\|_{L^2(\Omega)} \leq CN^{-\beta}. \tag{26}$$

In order to see that (26) implies $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C},\mathcal{D})$, we consider, for fixed $f \in \mathcal{C}$, the (unconstrained) best M-term approximation $f_M = \sum_{i\in I}\langle f,\varphi_i\rangle\varphi_i$ with $I \subseteq \mathbb{N}$, $|I| = M$. We now modify this M-term approximation by letting $\alpha := \lceil \gamma^*(\mathcal{C},\mathcal{D})/\beta \rceil \in \mathbb{N}$ and removing, in the expansion $f_M = \sum_{i\in I}\langle f,\varphi_i\rangle\varphi_i$, all terms corresponding to indices that are larger than M^α. Recalling that in Definition V.2 the same polynomial π bounds the search depth and the size of the coefficients, it follows that the modified approximation we just constructed obeys a polynomial depth search constraint with constraining polynomial $\pi_\alpha(x) = x^\alpha + S$, where $S := \sup_{f\in\mathcal{C}} \|f\|_{L^2(\Omega)}$. Here, owing to orthonormality of $\mathcal{D}$, S accounts for the size of the expansion coefficients $\langle f,\varphi_i\rangle$. In order to complete the argument, we need to show that the additional approximation error incurred by removing terms in $f_M = \sum_{i\in I}\langle f,\varphi_i\rangle\varphi_i$ is in $\mathcal{O}(M^{-\gamma^*(\mathcal{C},\mathcal{D})})$, i.e., it is of the same order as the error corresponding to the original (unconstrained) best M-term approximation. Due to orthonormality of $\mathcal{D}$ this additional error is given by the norm of $\sum_{i\in I, i>\pi_\alpha(M)}\langle f,\varphi_i\rangle\varphi_i$ and can, by virtue of (26), be bounded as

$$\begin{aligned}\left\| \sum_{i\in I, i>\pi_\alpha(M)} \langle f,\varphi_i\rangle\varphi_i \right\|_{L^2(\Omega)} &\leq \left\| \sum_{i=\pi_\alpha(M)+1}^{\infty} \langle f,\varphi_i\rangle\varphi_i \right\|_{L^2(\Omega)} = \left\| f - \sum_{i=1}^{\pi_\alpha(M)} \langle f,\varphi_i\rangle\varphi_i \right\|_{L^2(\Omega)} \\ &\leq C(\pi_\alpha(M))^{-\beta} \in \mathcal{O}(M^{-\gamma^*(\mathcal{C},\mathcal{D})}),\end{aligned}$$

which establishes the claim. We have hence shown that under tail compactness of arbitrary rate $\beta > 0$, $\gamma^*(\mathcal{C},\mathcal{D}) = \gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$, and hence there is no cost incurred by imposing a polynomial depth search constraint combined with a polynomial bound on the size of the expansion coefficients. We hasten to add that the assumptions stated at the beginning of this paragraph together with what was just established imply that $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$ is, indeed, well-defined. For the more general case of $\mathcal{D}$ a frame, we refer to [60, Sec. 5.4.3] for analogous arguments. Finally, we remark that the tail compactness inequality (26) can be interpreted as quantifying the rate of linear approximation for $\mathcal{C}$ in $\mathcal{D}$. Two examples of pairs $(\mathcal{C},\mathcal{D})$ satisfying tail compactness, namely Besov spaces with wavelet bases and modulation spaces with Wilson bases, are provided in Appendices B and C, respectively.

As already mentioned, a larger optimal exponent $\gamma^*(\mathcal{C})$ leads to faster error decay (specifically according to $L^{-\gamma^*(\mathcal{C})}$) and hence corresponds to a function class of smaller complexity. As such, techniques for deriving lower bounds on the optimal exponent are often based on variations of the approach employed in the proof of Theorem V.3, namely on the explicit construction of encoder-decoder pairs (in the case of the proof of Theorem V.3 by encoding the dictionary elements participating in the M-term approximation). A powerful method for deriving upper bounds on the optimal exponent is the hypercube embedding approach proposed by Donoho in [79]; the basic idea here is to show that the function class $\mathcal{C}$ under consideration contains a sufficiently complex embedded set of orthogonal hypercubes and to then find the exponent corresponding to this set. An interesting alternative technique for deriving optimal exponents was proposed in the context of modulation spaces in [77]. The essence of this approach is to exploit the isomorphism between weighted modulation spaces and weighted mixed-norm sequence spaces [17] and to then utilize results about entropy numbers of operators between sequence spaces.

VI. Approximation with Deep Neural Networks

Inspired by the theory of best M-term approximation with dictionaries, we now develop the new concept of best M-weight approximation through neural networks. At the heart of this theory 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 dictionary is replaced by parsimony in terms of network connectivity. Our development will parallel that for best M-term approximation in the previous section.

Before proceeding to the specifics, we would like to issue a general remark. While the neural network approximation results in Section III were formulated in terms of L^∞-norm, we shall be concerned with L^2-norm approximation here, on the one hand paralleling the use of L^2-norm in the context of best M-term approximation, and on the other hand allowing for the approximation of discontinuous functions by ReLU neural networks, which, owing to the continuity of the ReLU nonlinearity, necessarily realize continuous functions.

We start by introducing the concept of best M-weight approximation rate.

Definition VI.1. *Given $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and a function class $\mathcal{C} \subseteq L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma_M^{\mathcal{N}}(f) := \inf_{\substack{\Phi \in \mathcal{N}_{d,1} \\ \mathcal{M}(\Phi) \leq M}} \|f - \Phi\|_{L^2(\Omega)}. \tag{27}$$

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

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

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

We emphasize that the infimum in (27) 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 with respect

to 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 dictionaries, namely polynomial depth search and polynomially 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. To see this, first note that $\mathcal{W}(\Phi) \leq \mathcal{M}(\Phi)$ and $\mathcal{L}(\Phi) \leq \mathcal{M}(\Phi)$. 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 $\mathcal{O}(\mathcal{M}(\Phi)^3)$ possibilities for the "locations" (in terms of entries in the A_ℓ and the b_ℓ) of the $\mathcal{M}(\Phi)$ nonzero weights. Encoding the locations of the $\mathcal{M}(\Phi)$ nonzero weights hence requires $\log(\binom{C\mathcal{M}(\Phi)^3}{\mathcal{M}(\Phi)}) = \mathcal{O}(\mathcal{M}(\Phi)\log(\mathcal{M}(\Phi)))$ bits. This assumes, however, that the architecture of the network, i.e., the number of layers $\mathcal{L}(\Phi)$ and the N_k are known. Proposition VI.7 below shows that the architecture can, indeed, also be encoded with $\mathcal{O}(\mathcal{M}(\Phi)\log(\mathcal{M}(\Phi)))$ 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 $\mathcal{M}(\Phi)$ 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 $\mathcal{M}(\Phi)^{-\gamma}$, see Definition VI.1, this suggests to restrict $\mathcal{L}(\Phi)$ to growth that is polynomial in $\log(\mathcal{M}(\Phi))$.

The second restriction imposed in the definition of effective best M-term approximation, namely polynomially bounded coefficients, will be imposed in monomorphic manner on the magnitude of the weights. This growth condition will turn out natural in the context of the approximation results we are interested in and will, together with polylogarithmic depth growth, be seen below to allow rate-distortion-optimal quantization of the network weights. We remark, however, that networks with weights growing polynomially in $\mathcal{M}(\Phi)$ can be converted into networks with uniformly bounded weights at the expense of increased—albeit still of polylogarithmic scaling in $\mathcal{M}(\Phi)$—depth (see Proposition A.3). In summary, we will develop the concept of "best M-weight approximation subject to polylogarithmic depth and polynomial weight growth".

We start by introducing the following notation for neural networks with depth and weight magnitude bounded polylogarithmically respectively polynomially w.r.t. their connectivity.

Definition VI.2. *For $M, d, d' \in \mathbb{N}$, and π a polynomial, we define*

$$\mathcal{N}^{\pi}_{M,d,d'} := \{\Phi \in \mathcal{N}_{d,d'} : \mathcal{M}(\Phi) \leq M, \mathcal{L}(\Phi) \leq \pi(\log(M)), \mathcal{B}(\Phi) \leq \pi(M)\}\,.$$

Next, we formalize the notion of effective best M-weight approximation rate subject to polylogarithmic depth and polynomial weight growth.

Definition VI.3. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. We define for $M \in \mathbb{N}$ and π a polynomial*

$$\varepsilon_{\mathcal{N}}^{\pi}(M) := \sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{N}_{M,d,1}^{\pi}} \|f - \Phi\|_{L^2(\Omega)}$$

and

$$\gamma_{\mathcal{N}}^{*,\mathit{eff}}(\mathcal{C}) := \sup\{\gamma \geq 0 \colon \exists \text{ polynomial } \pi \text{ s.t. } \varepsilon_{\mathcal{N}}^{\pi}(M) \in \mathcal{O}(M^{-\gamma}),\, M \to \infty\}.$$

We refer to $\gamma_{\mathcal{N}}^{*,\mathit{eff}}(\mathcal{C})$ as the effective best M-weight approximation rate of $\mathcal{C}$.

We now state the equivalent of Theorem V.3 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 VI.4. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. Then, we have*

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

The key ingredients of the proof of Theorem VI.4 are developed throughout this section and the formal proof appears at the end of the section. Before getting started, we note that, in analogy to Definition V.4, what we just found suggests the following.

Definition VI.5. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. We say that the function class $\mathcal{C} \subseteq L^2(\Omega)$ is* optimally representable by neural networks *if*

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

It is interesting to observe that the fundamental limits of effective best M-term approximation (through dictionaries) and effective best M-weight 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 dictionary under polynomial weight growth of the coefficients and with the participating functions identified subject to a polynomial-depth search constraint in the former, and concatenations of affine functions followed by nonlinearities under polynomial growth constraints on the coefficients of the affine functions and with a polylogarithmic growth constraint on the number of concatenations in the latter case.

We now commence the program developing the proof of Theorem VI.4. As in the arguments in the proof sketch of Theorem V.3, the main idea is to compare the length of the bitstring needed to encode 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, its architecture, i.e., L and 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 appropriately with decreasing ε. To formalize this idea, we start by defining the quantization employed.

Definition VI.6. *Let $m \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$. 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 VI.4 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 VI.7. *Let $d, d' \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\mathcal{C} \subseteq L^2(\Omega)$, and let π be a polynomial. Further, let*

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

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

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

Then,

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

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

$$\begin{aligned}
\ell(\varepsilon) &\leq C_0 \cdot \sup_{f\in\mathcal{C}} \left(\mathcal{M}(\Psi(\varepsilon, f)) \log(\mathcal{M}(\Psi(\varepsilon, f))) + 1\right) (\log(\varepsilon^{-1}))^q \qquad (28)\\
&\leq C_0 \left(\varepsilon^{-1/\gamma} \log(\varepsilon^{-1/\gamma}) + 1\right) (\log(\varepsilon^{-1}))^q \\
&\leq C_1 \left(\varepsilon^{-1/\gamma} (\log(\varepsilon^{-1}))^{q+1} + (\log(\varepsilon^{-1}))^q\right) \in \mathcal{O}\left(\varepsilon^{-1/\nu}\right), \quad \text{for } \varepsilon \to 0,
\end{aligned}$$

where $C_0, C_1, q > 0$ are constants not depending on f, ε and $\gamma > \nu > \gamma^*(\mathcal{C})$. The specific form of the upper bound (28) will become apparent in the construction of the bitstring representing Ψ detailed below.

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 architecture, its topology, and the quantized weights in bitstrings of length $\ell(\varepsilon)$ satisfying (28) while guaranteeing unique reconstruction (of the network). For the sake of notational simplicity, we fix $\varepsilon \in (0, 1/2)$ and $f \in \mathcal{C}$ and set $\Psi := \Psi(\varepsilon, f)$, $M := \mathcal{M}(\Psi)$, and $L := \mathcal{L}(\Psi)$. Recall that the number of nodes in layers $0, \ldots, L$ is denoted by $N_0, \ldots, N_L$ and that $N_0 = d, N_L = d'$ (see Definition II.1). Moreover, note that due to our nondegeneracy assumption (see Remark II.2) we have $\sum_{\ell=0}^{L} N_\ell \leq 2M$ and $L \leq M$. The bitstring representing Ψ is constructed according to the following steps.

Step 1: If $M = 0$, we encode the network by a single 0. Using the convention $0\log(0) = 0$, we then note that (28) 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 given by $M + 1$.

Step 2: We continue by encoding the number of layers which, due to $L \leq M$, requires no more than $\lceil \log(M) \rceil$ bits. We thus reserve the next $\lceil \log(M) \rceil$ bits for the binary representation of L.

Step 3: Next, we store the layer dimensions $N_0, \ldots, N_L$. As $L \leq M$ and $N_\ell \leq M$, for all $\ell \in \{0, \ldots, L\}$, owing to nondegeneracy, we can encode the layer dimensions using $(M+1)\lceil \log(M) \rceil$ bits. In combination with Steps 1 and 2 this yields an overall bitstring of length at most

$$M\lceil \log(M) \rceil + M + 2\lceil \log(M) \rceil + 1. \tag{29}$$

Step 4: We encode the topology of the graph associated with the network Ψ. To this end, we enumerate all nodes by assigning a unique index i to each one of them, starting from the 0-th layer and increasing from left to right within a given layer. The indices range from 1 to $N := \sum_{\ell=0}^{L} N_\ell \leq 2M$. Each of these indices can be encoded by a bitstring of length $\lceil \log(N) \rceil$. We denote the bitstring corresponding to index i by $b(i) \in \{0,1\}^{\lceil \log(N) \rceil}$ and let for all nodes, except for those in the last layer, $n(i)$ 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. For each of these nodes i, we form a bitstring of length $n(i)\lceil \log(N) \rceil$ by concatenating the bitstrings indexing its children. We follow this string with an all-zeros bitstring of length $\lceil \log(N) \rceil$ to signal that all children of the current node have been encoded. Overall, this yields a bitstring of length

$$\sum_{i=1}^{N-d'} (n(i)+1)\lceil \log(N) \rceil \leq 3M\lceil \log(2M) \rceil, \tag{30}$$

where we used $\sum_{i=1}^{N-d'} n(i) \leq M$.

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 $B_\varepsilon := 2(\lceil \pi(\log(\varepsilon^{-1})) \rceil \lceil \log(\varepsilon^{-1}) \rceil + 1)$ bits. For each node $i = 1, \ldots, N$, 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 the current node and that child. 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

$$\sum_{i=1}^{N-d'} (n(i)+1)B_\varepsilon + d'B_\varepsilon \leq 3MB_\varepsilon$$

representing the weights. Combining this with (29) and (30), we find that the overall number of bits needed to encode the network architecture, topology, and weights is no more than

$$3MB_\varepsilon + 3M\lceil \log(2M) \rceil + (M+2)\lceil \log(M) \rceil + M + 1. \tag{31}$$

The network can be recovered by sequentially reading out M, L, the N_ℓ, 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 (31) can be upper-bounded by

$$C_0(M\log(M)+1)(\log(\varepsilon^{-1}))^q$$

for constants $C_0, q > 0$ depending on 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 (28). This concludes the proof. □

Proposition VI.7 states that the connectivity growth rate of networks with quantized weights achieving uniform approximation error ε over a function class $\mathcal{C}$ must exceed $\mathcal{O}\left(\varepsilon^{-1/\gamma^*(\mathcal{C})}\right)$, $\varepsilon \to 0$. As Proposition VI.7 applies to networks that have each weight represented by a finite number of bits scaling polynomially in $\log(\varepsilon^{-1})$, while guaranteeing that the underlying encoder-decoder pair achieves uniform error ε over $\mathcal{C}$, it remains to establish 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. Establishing that this delicate balancing is implied by our technical assumptions is the subject of the remainder of this section. We start with a perturbation result quantifying how the error induced by weight quantization in the network translates to the output function realized by the network.

Lemma VI.8. *Let $d, d', k \in \mathbb{N}$, $D \in \mathbb{R}_+$, $\Omega \subseteq [-D, D]^d$, $\varepsilon \in (0, 1/2)$, let $\Phi \in \mathcal{N}_{d,d'}$ with $\mathcal{M}(\Phi) \leq \varepsilon^{-k}$, $\mathcal{B}(\Phi) \leq \varepsilon^{-k}$, and let $m \in \mathbb{N}$ satisfy*

$$m \geq 3k\mathcal{L}(\Phi) + \log(\lceil D \rceil). \tag{32}$$

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

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

More specifically, the network $\tilde{\Phi}$ can be obtained simply by replacing every weight in Φ by a closest element in $2^{-m\lceil\log(\varepsilon^{-1})\rceil}\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$.

Proof of Theorem VI.8. We first consider the case $\mathcal{L}(\Phi) = 1$. Here, it follows from Definition II.1 that the network simply realizes an affine transformation and hence

$$\sup_{x\in\Omega} \|\Phi(x) - \tilde{\Phi}(x)\|_\infty \leq \mathcal{M}(\Phi)\lceil D \rceil 2^{-m\lceil\log(\varepsilon^{-1})\rceil - 1} \leq \varepsilon.$$

In the remainder of the proof, we can therefore assume that $\mathcal{L}(\Phi) \geq 2$. For simplicity of notation, we set $L := \mathcal{L}(\Phi), M := \mathcal{M}(\Phi)$, and, as usual, write

$$\Phi = W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

with $W_\ell(x) = A_\ell x + b_\ell$, $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, and $b_\ell \in \mathbb{R}^{N_\ell}$. We now consider the partial networks $\Phi^\ell \colon \Omega \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L-1\}$, given by

$$\Phi^\ell := \begin{cases} \rho \circ W_1, & \ell = 1 \\ \rho \circ W_2 \circ \rho \circ W_1, & \ell = 2 \\ \rho \circ W_\ell \circ \rho \circ W_{\ell-1} \circ \cdots \circ \rho \circ W_1, & \ell = 3, \ldots, L-1, \end{cases}$$

and set $\Phi^L := \Phi$. We hasten to add that we decided—for ease of exposition—to deviate from the convention used in Definition II.1 and to have the partial networks include the application of ρ at the end. Now, 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 and b_ℓ by a closest element in $2^{-m\lceil \log(\varepsilon^{-1})\rceil}\,\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$. We denote these replacements by $\tilde{A}_\ell$ and $\tilde{b}_\ell$, respectively, and note that

$$\begin{aligned} \max_{i,j} |A_{\ell,i,j} - \tilde{A}_{\ell,i,j}| &\leq \tfrac{1}{2}\, 2^{-m\lceil \log(\varepsilon^{-1})\rceil} \leq \tfrac{1}{2}\, \varepsilon^m, \\ \max_{i,j} |b_{\ell,i,j} - \tilde{b}_{\ell,i,j}| &\leq \tfrac{1}{2}\, 2^{-m\lceil \log(\varepsilon^{-1})\rceil} \leq \tfrac{1}{2}\, \varepsilon^m. \end{aligned} \tag{33}$$

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

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

We further set $C_0 := \lceil D \rceil$ 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 obtained by applying[10] the 1-Lipschitz function ρ to the sum of a weighted sum of at most $N_{\ell-1}$ components of the vector $\Phi^{\ell-1}(x) \in \mathbb{R}^{N_{\ell-1}}$ and a bias 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

$$C_\ell \leq C_0\, \varepsilon^{-k\ell} \prod_{i=0}^{\ell-1} (N_i + 1). \tag{34}$$

Next, note that the components $(\tilde{\Phi}^1(x))_i, i \in \{1, 2, \ldots, N_1\}$, 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 (33) and the fact that ρ is 1-Lipschitz implies

$$e_1 \leq C_0 N_0 \tfrac{\varepsilon^m}{2} + \tfrac{\varepsilon^m}{2} \leq C_0 (N_0 + 1) \tfrac{\varepsilon^m}{2}. \tag{35}$$

[10]Note that going from Φ_{L-1} to Φ_L the activation function is not applied anymore, which nevertheless leads to the same estimate as the identity mapping is 1-Lipschitz.

Due to ρ and the identity mapping being 1-Lipschitz, we have, for $\ell = 1, \dots, L$,

$$\begin{aligned} e_\ell &= \sup_{x\in\Omega} \|\Phi^\ell(x) - \tilde{\Phi}^\ell(x)\|_\infty = \sup_{x\in\Omega, i\in\{1,\dots,N_\ell\}} |(\Phi^\ell(x))_i - (\tilde{\Phi}^\ell(x))_i| \\ &\leq \sup_{x\in\Omega, i\in\{1,\dots,N_\ell\}} \left| \left[\left(\sum_{j=1}^{N_{\ell-1}} A_{\ell,i,j} (\Phi^{\ell-1}(x))_j \right) + b_{\ell,i} \right] - \left[\left(\sum_{j=1}^{N_{\ell-1}} \tilde{A}_{\ell,i,j} (\tilde{\Phi}^{\ell-1}(x))_j \right) + \tilde{b}_{\ell,i} \right] \right| \\ &\leq \sup_{x\in\Omega, i\in\{1,\dots,N_\ell\}} \left[\left(\sum_{j=1}^{N_{\ell-1}} \left| A_{\ell,i,j} (\Phi^{\ell-1}(x))_j - \tilde{A}_{\ell,i,j} (\tilde{\Phi}^{\ell-1}(x))_j \right| \right) + \left| b_{\ell,i} - \tilde{b}_{\ell,i} \right| \right]. \end{aligned} \tag{36}$$

As $|(\Phi^{\ell-1}(x))_j - (\tilde{\Phi}^{\ell-1}(x))_j| \leq e_{\ell-1}$ and $|(\Phi^{\ell-1}(x))_j| \leq C_{\ell-1}$ for all $x \in \Omega$, $j \in \{1, \dots, N_{\ell-1}\}$ by definition, and $|A_{\ell,i,j}| \leq \varepsilon^{-k}$ by assumption, upon invoking (33), 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 (36), that for all $\ell \in \{2, \dots, L\}$,

$$e_\ell \leq N_{\ell-1}(e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2} + e_{\ell-1}\tfrac{\varepsilon^m}{2}) + \tfrac{\varepsilon^m}{2} \leq (N_{\ell-1}+1)(2e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2}). \tag{37}$$

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

$$e_\ell \leq \tfrac{1}{2}(2^\ell - 1) C_0 \varepsilon^{m-(\ell-1)k} \prod_{i=0}^{\ell-1} (N_i + 1), \tag{38}$$

which we prove by induction. The base case $\ell = 1$ was already established in (35). For the induction step we assume that (38) holds for a given ℓ which, in combination with (34) and (37), 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 (38). Using $2^{L-1} \leq \varepsilon^{-(L-1)}$, $\prod_{i=0}^{L-1}(N_i+1) \leq M^L \leq \varepsilon^{-kL}$, and $m \geq 3kL + \log(\lceil D \rceil)$ by assumption, we get

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

This completes the proof. □

We are now ready to finalize the proof of Theorem VI.4.

Proof of Theorem VI.4. Suppose towards a contradiction that $\gamma^{*,\mathrm{eff}}_{\mathcal{N}}(\mathcal{C}) > \gamma^*(\mathcal{C})$ and let $\gamma \in \big(\gamma^*(\mathcal{C}), \gamma^{*,\mathrm{eff}}_{\mathcal{N}}(\mathcal{C})\big)$. Then, by Definition VI.3, there exist a polynomial π and a constant $C > 0$ such that

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

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 $\Phi_{\varepsilon,f} \in \mathcal{N}^{\pi}_{M_\varepsilon,d,1}$ such that

$$\|f-\Phi_{\varepsilon,f}\|_{L^2(\Omega)} \leq 2\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{N}^{\pi}_{M_\varepsilon,d,1}} \|f-\Phi\|_{L^2(\Omega)} \leq 2CM_\varepsilon^{-\gamma} \leq \frac{\varepsilon}{2}. \tag{39}$$

By Lemma VI.8 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

$$\Big\|\Phi_{\varepsilon,f} - \widetilde{\Phi}_{\varepsilon,f}\Big\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2}. \tag{40}$$

The conditions of Lemma VI.8 are satisfied as M_ε can be upper-bounded by ε^{-k} with a suitably chosen k, the weights in $\Phi_{\varepsilon,f}$ are polynomially bounded in M_ε, and (32) follows from the depth of networks in $\Phi \in \mathcal{N}^{\pi}_{M_\varepsilon,d,1}$ being polylogarithmically bounded in M_ε due to Definition VI.2. Now, defining

$$\Psi\colon \big(0, \tfrac{1}{2}\big) \times \mathcal{C} \to \mathcal{N}_{d,1}, \quad (\varepsilon, f) \mapsto \widetilde{\Phi}_{\varepsilon,f},$$

it follows from (39) and (40), by application of the triangle inequality, that

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

The proof is concluded by noting that $\Psi(\varepsilon, f)$ violates Proposition VI.7. □

We conclude this section with a discussion of the conceptual implications of the results established above. Proposition VI.7 combined with Lemma VI.8 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.

VII. The Transference Principle

We have seen that a wide array of function classes can be approximated in Kolmogorov-Donoho optimal fashion through dictionaries, provided that the dictionary $\mathcal{D}$ is chosen to consort with the function class $\mathcal{C}$ according to $\gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C})$. Examples of such pairs are unit balls in Besov spaces with wavelet bases and unit balls in weighted modulation spaces with Wilson bases. A more extensive list of optimal pairs is provided in Table 1. On the other hand, as shown in [14], Fourier bases are strictly suboptimal—in terms of approximation rate—for balls $\mathcal{C}$ of finite radius in the spaces $BV(\mathbb{R})$ and $W^m_p(\mathbb{R})$.

In light of what was just said, it is hence natural to let neural networks play the role of the dictionary $\mathcal{D}$ and to ask which function classes $\mathcal{C}$ are approximated in Kolmogorov-Donoho-optimal fashion by neural networks. Towards answering this question, we next develop a general framework for transferring results on function approximation through dictionaries to results on approximation by neural networks. This will eventually lead us to a characterization of function classes $\mathcal{C}$ that are optimally representable by neural networks in the sense of Definition VI.5.

We start by introducing the notion of effective representability of dictionaries through neural networks.

Definition VII.1. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ be a dictionary. We call $\mathcal{D}$* 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 $\Phi_{i,\varepsilon} \in \mathcal{N}_{d,1}$ 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.4—of a dictionary $\mathcal{D}$ for a function class $\mathcal{C}$ combined with effective representability of $\mathcal{D}$ by neural networks implies optimal representability of $\mathcal{C}$ by neural networks. The proof is, in essence, effected by noting that every element of the effectively representable $\mathcal{D}$ participating in a best M-term-rate achieving approximation f_M of $f \in \mathcal{C}$ can itself be approximated by neural networks well enough for an overall network to approximate f_M with connectivity $M\pi(\log(M))$. As this connectivity is only polylogarithmically larger than the number of terms M participating in the best M-term approximation f_M, we will be able to conclude that the optimal approximation rate, indeed, transfers from approximation in $\mathcal{D}$ to approximation in neural networks. The conditions on $\mathcal{M}(\Phi_{i,\varepsilon})$ and $\mathcal{B}(\Phi_{i,\varepsilon})$ in Definition VII.1 guarantee precisely that the connectivity increase is at most by a polylogarithmic factor. To see this, we first recall that effective best M-term approximation has a polynomial depth search constraint, which implies that the indices i under consideration are upper-bounded by a polynomial in M. In addition, the approximation error behavior we are interested in is $\varepsilon = M^{-\gamma}$. Combining these two insights, it follows that $\mathcal{M}(\Phi_{i,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(i))$ implies polylogarithmic (in M) connectivity for each network $\Phi_{i,\varepsilon}$ and hence connectivity $M\pi(\log(M))$ for the overall network realizing f_M, as desired. By the same token, $\mathcal{B}(\Phi_{i,\varepsilon}) \leq \pi(\varepsilon^{-1}, i)$ guarantees that the weights of $\Phi_{i,\varepsilon}$ are polynomial in M.

There is another aspect to effective representability by neural networks that we would like to illustrate by way of example, namely that of ordering the dictionary elements. Specifically, we consider, for $d = 1$ and $\Omega = [-\pi, \pi)$, the class $\mathcal{C}$ of real-valued even functions in $\mathcal{C} = L^2(\Omega)$, and take the dictionary as $\mathcal{D} = \{\cos(ix), i \in \mathbb{N}_0\}$. As the index i enumerating the dictionary elements corresponds to frequencies, the basis functions in $\mathcal{D}$ are hence ordered according to increasing frequencies. Next, note that the parameter a in Theorem III.8 corresponds to the frequency index i in our example. As the network $\Psi_{a,D,\varepsilon}$ in Theorem III.8 is of finite width, it hence follows, upon replacing a in the expression for $\mathcal{L}(\Psi_{a,D,\varepsilon})$ by i, that $\mathcal{M}(\Psi_{i,D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(i))$. The condition on the weights for effective representability is satisfied trivially, simply as $\mathcal{B}(\Psi_{i,D,\varepsilon}) \leq 1 \leq \pi(\varepsilon^{-1}, i)$.

We are now ready to state the rate optimality transfer result.

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

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

such that for all $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, 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 we have

$$\mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma}),\ \varepsilon \to 0, \tag{41}$$

with the implicit constant in (41) being independent of f. In particular, it holds that

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

Remark VII.3. *Theorem VII.2 allows us to draw the following conclusion. If $\mathcal{D}$ optimally represents the function class $\mathcal{C}$ in the sense of Definition V.4, 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 VII.1, then, due to Theorem VI.4, which states that $\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) \leq \gamma^*(\mathcal{C})$, we have $\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) = \gamma^*(\mathcal{C})$ and hence $\mathcal{C}$ is optimally representable by neural networks in the sense of Definition VI.5.*

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

$$\left\| f - \sum_{i\in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)} \leq \frac{C M^{-\gamma'}}{2}. \tag{42}$$

Let $A := \max\{1, |\Omega|^{1/2}\}$. Effective representability of $\mathcal{D}$ according to Definition VII.1 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 $\Phi_{i,M} \in \mathcal{N}_{d,1}$ satisfying

$$\|\varphi_i - \Phi_{i,M}\|_{L^2(\Omega)} \leq \tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)} \tag{43}$$

with

$$\begin{aligned}
\mathcal{M}(\Phi_{i,M}) &\leq \pi_2\left(\log\left(\left(\tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)}\right)^{-1}\right), \log(i)\right) \\
&= \pi_2\left((\gamma'+1)\log(M) + \log\left(\tfrac{4A\pi_1(M)}{C}\right), \log(i)\right), \\
\mathcal{B}(\Phi_{i,M}) &\leq \pi_2\left(\left(\tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)}\right)^{-1}, i\right) = \pi_2\left(\tfrac{4A\pi_1(M)}{C} M^{\gamma'+1}, i\right).
\end{aligned} \tag{44}$$

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

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

Due to $\max(I_{f,M}) \leq \pi_1(M)$, (44) and Lemma A.8 imply the existence of a polynomial π_3 such that $\mathcal{L}(\Psi_{f,M}) \leq \pi_3(\log(M))$, $\mathcal{M}(\Psi_{f,M}) \leq M\pi_3(\log(M))$, and $\mathcal{B}(\Psi_{f,M}) \leq \pi_3(M)$, for all $f \in \mathcal{C}$, $M \in \mathbb{N}$, and, owing to (43), we get

$$\left\| \Psi_{f,M} - \sum_{i \in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)} \leq \sum_{i \in I_{f,M}} |c_i| \tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)} \leq \tfrac{CM^{-\gamma'}}{4A} \sum_{i=1}^{|I_{f,M}|} \tfrac{\max_{i \in I_{f,M}} |c_i|}{M\pi_1(M)} \leq \tfrac{CM^{-\gamma'}}{4A}. \tag{45}$$

Lemma VI.8 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 $\widetilde{\Psi}_{f,M} \in \mathcal{N}_{d,1}$ with $(\lceil \pi_4(\log(\frac{4A}{C} M^{\gamma'})) \rceil, \frac{CM^{-\gamma'}}{4A})$-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

$$\left\| \Psi_{f,M} - \widetilde{\Psi}_{f,M} \right\|_{L^\infty(\Omega)} \leq \tfrac{CM^{-\gamma'}}{4A}. \tag{46}$$

As Ω is bounded by assumption, we have

$$\left\| \Psi_{f,M} - \widetilde{\Psi}_{f,M} \right\|_{L^2(\Omega)} \leq |\Omega|^{\frac{1}{2}} \left\| \Psi_{f,M} - \widetilde{\Psi}_{f,M} \right\|_{L^\infty(\Omega)} \leq \tfrac{CM^{-\gamma'}}{4}, \tag{47}$$

for all $f \in \mathcal{C}$, $M \in \mathbb{N}$. Combining (47) with (42) and (45), we get, for all $f \in \mathcal{C}$, $M \in \mathbb{N}$,

$$\begin{aligned} \left\| f - \widetilde{\Psi}_{f,M} \right\|_{L^2(\Omega)} &\leq \left\| f - \sum_{i \in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)} + \left\| \sum_{i \in I_{f,M}} c_i \varphi_i - \Psi_{f,M} \right\|_{L^2(\Omega)} + \left\| \Psi_{f,M} - \widetilde{\Psi}_{f,M} \right\|_{L^2(\Omega)} \\ &\leq CM^{-\gamma'}. \end{aligned} \tag{48}$$

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, (48) yields

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

Next, we note that, for all polynomials π and $0 \leq m < n$,

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

As $1/\gamma' < 1/\gamma$, this establishes

$$\mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(M_\varepsilon \pi_3(\log(M_\varepsilon))) \subseteq \mathcal{O}(\varepsilon^{-1/\gamma}),\ \varepsilon \to 0. \tag{49}$$

Since M_ε and π_3 are independent of f, the implicit constant in (49) does not depend on f.

Next, note that, in general, an (n, η)-quantized network is also (m, δ)-quantized for $n \geq m$ and $\eta \leq \delta$, simply as

$$2^{-m\lceil \log(\delta^{-1}) \rceil} \mathbb{Z} \cap [-\delta^{-m}, \delta^{-m}] \subseteq 2^{-n\lceil \log(\eta^{-1}) \rceil} \mathbb{Z} \cap [-\eta^{-n}, \eta^{-n}].$$

Since $\frac{CM_\varepsilon^{-\gamma'}}{4A} \leq \varepsilon$ this ensures the existence of a polynomial π such that, for every $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, the network $\Psi(\varepsilon, f)$ is $(\lceil \pi(\log(\varepsilon^{-1}))\rceil, \varepsilon)$-quantized, $\mathcal{L}(\Psi(\varepsilon, f)) \leq \pi(\log(\varepsilon^{-1}))$, and $\mathcal{B}(\Psi(\varepsilon, f)) \leq \pi(\varepsilon^{-1})$. With (49) this establishes the first claim of the theorem. In order to verify the second claim, note that $\Psi(\varepsilon, f) \in \mathcal{N}^{\pi}_{\mathcal{M}(\Psi(\varepsilon,f)),d,1}$, for all $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, which implies

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

Therefore, owing to Definition VI.3, we get

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

which concludes the proof. □

Remark VII.4. *We note that Theorem VII.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 dictionaries in the proof of Theorem IX.3, but are omitted here for ease of exposition.*

The last piece needed to complete our program is to establish that the conditions in Definition VII.1 guaranteeing effective representability in neural networks are, indeed, satisfied by a wide variety of dictionaries.

Inspecting Table 1, we can see that all example function classes provided therein are optimally represented either by affine dictionaries, i.e., wavelets, the Haar basis, and curvelets or Weyl-Heisenberg dictionaries, namely Fourier bases and Wilson bases. The next two sections will be devoted to proving effective representability of affine dictionaries and Weyl-Heisenberg dictionaries by neural networks, thus allowing us to draw the conclusion that neural networks are universally Kolmogorov-Donoho optimal approximators for all function classes listed in Table 1.

VIII. Affine Dictionaries are Effectively Representable by Neural Networks

The purpose of this section is to establish that *affine dictionaries*, including wavelets [70], ridgelets [39], curvelets [71], shearlets [72], α-shearlets and more generally α-molecules [69], which contain all aforementioned dictionaries as special cases, are effectively representable by neural networks. Due to Theorem VII.2 and Theorem VI.4, this will then allow us to conclude that any function class that is optimally representable—in the sense of Definition V.4—by an affine dictionary with a suitable generator function is optimally representable by neural networks in the sense of Definition VI.5. By "suitable" we mean that the generator function can be approximated well by ReLU networks in a sense to be made precise below.

In order to elucidate the main ideas underlying the general definition of affine dictionaries that are effectively representable by neural networks, we start with a basic example, namely the Haar wavelet dictionary on the unit interval, i.e., the set of functions

$$\psi_{n,k}\colon [0,1] \mapsto \mathbb{R},\ x \mapsto 2^{\frac{n}{2}}\psi(2^n x - k),\ n \in \mathbb{N}_0,\ k = 0, \dots, 2^n - 1,$$

with

$$\psi\colon \mathbb{R} \to \mathbb{R},\ x \mapsto \begin{cases} 1, & x \in [0, 1/2) \\ -1, & x \in [1/2, 1) \\ 0, & \text{else.} \end{cases}$$

We approximate the piecewise constant mother wavelet ψ through a continuous piecewise linear function realized by a neural network as follows

$$\Psi_\delta(x) := \tfrac{1}{2\delta}\rho(x+\delta) - \tfrac{1}{2\delta}\rho(x-\delta) - \tfrac{1}{\delta}\rho(x-(\tfrac{1}{2}-\delta)) + \tfrac{1}{\delta}\rho(x-(\tfrac{1}{2}+\delta)) + \tfrac{1}{2\delta}\rho(x-(1-\delta)) - \tfrac{1}{2\delta}\rho(x-(1+\delta))$$

and, setting $\delta(\varepsilon) := \varepsilon^2$ for $\varepsilon \in (0, 1/2)$, let

$$\Phi_{n,k,\varepsilon}(x) := 2^{\frac{n}{2}}\Psi_{\delta(\varepsilon)}(2^n x - k),\ n \in \mathbb{N}_0,\ k = 0, \dots, 2^n - 1.$$

The basic idea in the approximation of ψ through Ψ_δ is to let the transition regions around $0, 1/2$, and 1 shrink, as a function of ε, sufficiently fast for the construction to realize an approximation error of no more than ε. Now, a direct calculation yields that, indeed, for $\varepsilon \in (0, 1/2)$,

$$\|\psi_{n,k} - \Phi_{n,k,\varepsilon}\|_{L^2([0,1])} \leq \varepsilon.$$

Moreover, we have $\mathcal{M}(\Phi_{n,k,\varepsilon}) = 18$ and $\mathcal{B}(\Phi_{n,k,\varepsilon}) \leq \max\{2^{\frac{n}{2}}\varepsilon^{-2}, 2^n\}$. In order to establish effective representability by neural networks, we need to order the Haar wavelet dictionary suitably. Specifically, we proceed from coarse to fine scales, i.e., we let $(\varphi_i)_{i\in\mathbb{N}} = \mathcal{D} = \{\mathcal{D}_0, \mathcal{D}_1, \dots\}$, with $\mathcal{D}_n := \{\psi_{n,k} \mapsto \mathbb{R}\colon k = 0, \dots, 2^n - 1\}$, where the ordering within the $\mathcal{D}_n$ may be chosen arbitrarily. Next, note that for every pair $n \in \mathbb{N}_0$, $k \in \{0, \dots, 2^n - 1\}$, there exists a unique index $i \in \mathbb{N}$ such that $\varphi_i = \psi_{n,k} = \psi_{n(i),k(i)}$ and, owing to $|\mathcal{D}_n| = 2^n$, we have $2^{n(i)} \leq i$. Finally, taking $\Phi_{i,\varepsilon} := \Phi_{n(i),k(i),\varepsilon}$ and $\pi(a,b) := a^2 b + b + 18$, the conditions in Definition VII.1 for effective representability by neural networks are readily verified. A more elaborate example, namely spline wavelets, is considered at the end of this section.

We are now ready to proceed to the general definition of affine dictionaries with canonical ordering.

A. Affine Dictionaries with Canonical Ordering

Definition VIII.1. *Let $d, S \in \mathbb{N}$, $\delta > 0$, $\Omega \subseteq \mathbb{R}^d$ be bounded, and let $g_s \in L^\infty(\mathbb{R}^d)$, $s \in \{1, \dots, S\}$, be compactly supported. Furthermore, for $s \in \{1, \dots, S\}$, let $J_s \subseteq \mathbb{N}$ and $A_{s,j} \in \mathbb{R}^{d\times d}$, $j \in J_s$, be full-rank and with eigenvalues bounded below by 1 in absolute value. We define the* affine dictionary $\mathcal{D} \subseteq L^2(\Omega)$ *with generator functions $(g_s)_{s=1}^S$ as*

$$\mathcal{D} := \left\{ g_s^{j,e} := \left(|\det(A_{s,j})|^{\frac{1}{2}} g_s(A_{s,j} \cdot - \delta e) \right) \big|_\Omega : \ s \in \{1, \dots, S\},\ e \in \mathbb{Z}^d,\ j \in J_s, \text{ and } g_s^{j,e} \neq 0 \right\}.$$

Moreover, we define the sub-dictionaries

$$\mathcal{D}_{s,j} := \{ g_s^{j,e} \in \mathcal{D} : e \in \mathbb{Z}^d \text{ and } g_s^{j,e} \neq 0 \}, \quad \text{for } j \in J_s,\ s \in \{1, \dots, S\}$$
$$\mathcal{D}_j := \bigcup_{s \in \{1,\dots,S\} :\ j \in J_s} \mathcal{D}_{s,j}, \quad \text{for } j \in \mathbb{N}.$$

We call an affine dictionary canonically ordered if it is arranged according to

$$(\varphi_i)_{i\in\mathbb{N}} = \mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \dots), \tag{50}$$

where the elements within each $\mathcal{D}_j$ may be ordered arbitrarily, and there exist constants $a, c > 0$ such that

$$\sum_{k=1}^{j-1} |\det(A_{s,k})| \geq c \|A_{s,j}\|_\infty^a, \quad \textit{for all } j \in J_s \backslash \{1\},\ s \in \{1, \dots, S\}. \tag{51}$$

We call an affine dictionary nondegenerate if for every $j \in J_s$, $s \in \{1, \dots, S\}$, the sub-dictionary $\mathcal{D}_{s,j}$ contains at least one element.

Note that for sake of greater generality, we associate possibly different sets $J_s \subseteq \mathbb{N}$ with the generator functions g_s and, in particular, also allow these sets to be finite. The Haar wavelet dictionary example above is recovered as a nondegenerate affine dictionary by taking $d = 1$, $\Omega = [0, 1]$, $S = 1$, $J_s = \mathbb{N}$, $g_1 = \psi$, $\delta = 1$, $A_{1,j} = 2^{j-1}$, $a = 1$, $c = 1/2$, and noting that nondegeneracy is verified as for scale j, the sub-dictionary $\mathcal{D}_{s,j}$ contains 2^{j-1} elements. Moreover, the weights of the networks approximating the individual Haar wavelet dictionary elements grow linearly in the index of the dictionary elements. This is a consequence of the weights being determined by the dilation factor 2^n and $2^{n(i)} \leq i$ due to the ordering we chose. As will be shown below, morally this continues to hold for general nondegenerate affine dictionaries, thereby revealing what informed our definition of canonical ordering. Besides, our notion of canonical ordering is also inspired by the ordering employed in the tail compactness considerations for Besov spaces and orthonormal wavelet dictionaries as detailed in Appendix B. We remark that (51) constitutes a very weak restriction on how fast the size of dilations may grow; in fact, we are not aware of any affine dictionaries in the literature that would violate this condition. Finally, we note that the dilations $A_{s,j}$ are not required to be ordered in ascending size, as was the case in the Haar wavelet dictionary example. Canonical ordering does, however, ensure a modicum of ordering.

B. Invariance to Affine Transformations

Affine dictionaries consist of dilations and translations 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 nonlinearities, it is clear that translations and dilations can be absorbed into the first layer of the network and the transformed function should inherit the approximability properties of the generator function. However, what we will have to understand is how the weights, the connectivity, and the domain of approximation of the resulting network are impacted. The following result makes this quantitative.

Proposition VIII.2. *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 such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|f - \Phi_{D,\varepsilon}\|_{L^p([-D,D]^d)} \leq \varepsilon, \tag{52}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Then, for all full-rank matrices $A \in \mathbb{R}^{d\times d}$, and all $e \in \mathbb{R}^d$, $E \in \mathbb{R}_+$, and $\eta \in (0,1/2)$, there is a network $\Psi_{A,e,E,\eta} \in \mathcal{N}_{d,1}$ satisfying

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

with $\mathcal{M}(\Psi_{A,e,E,\eta}) \leq \pi'(\log(\eta^{-1}), \log(\lceil F \rceil))$ and $\mathcal{B}(\Psi_{A,e,E,\eta}) \leq \max\{\mathcal{B}(\Phi_{F,\eta}), |\det(A)|^{\frac{1}{p}}, \|A\|_\infty, \|e\|_\infty\}$, where $F = dE\|A\|_\infty + \|e\|_\infty$ and π' is of the same degree as π.

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

$$\big\| |\det(A)|^{\frac{1}{p}} f(A\cdot - e) - |\det(A)|^{\frac{1}{p}} \Phi(A\cdot - e) \big\|_{L^p([-E,E]^d)} = \|f - \Phi\|_{L^p(A\cdot[-E,E]^d - e)}. \tag{53}$$

Furthermore, observe that

$$A\cdot[-E,E]^d - e \subseteq [-(dE\|A\|_\infty + \|e\|_\infty), (dE\|A\|_\infty + \|e\|_\infty)]^d = [-F,F]^d. \tag{54}$$

Next, we consider the affine transformations $W_{A,e}(x) := Ax - e$, $W'_A(x) := |\det(A)|^{\frac{1}{p}} x$ as depth-1 networks and take $\Psi_{A,e,E,\eta} := W'_A \circ \Phi_{F,\eta} \circ W_{A,e}$ according to Lemma II.3. Combining (53) and (54) yields

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

The desired bounds on $\mathcal{M}(\Psi_{A,e,E,\eta})$ and $\mathcal{B}(\Psi_{A,e,E,\eta})$ follow directly by construction. □

C. Canonically Ordered Affine Dictionaries are Effectively Representable

The next result establishes that canonically ordered affine dictionaries with generator functions that can be approximated well by neural networks are effectively representable by neural networks.

Theorem VIII.3. *Let $d, S \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ be bounded with nonempty interior, $(g_s)_{s=1}^S \in L^\infty(\mathbb{R}^d)$ compactly supported, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ a nondegenerate canonically ordered affine dictionary with generator functions $(g_s)_{s=1}^S$. Assume that there exists a polynomial π such that, for all $s \in \{1,\dots,S\}$, $\varepsilon \in (0,1/2)$, there is a network $\Phi_{s,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|g_s - \Phi_{s,\varepsilon}\|_{L^2(\mathbb{R}^d)} \leq \varepsilon, \tag{55}$$

with $\mathcal{M}(\Phi_{s,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_{s,\varepsilon}) \leq \pi(\varepsilon^{-1})$. Then, $\mathcal{D}$ is effectively representable by neural networks.

Proof. By Definition VII.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 $\Phi_{i,\eta} \in \mathcal{N}_{d,1}$ satisfying

$$\|\varphi_i - \Phi_{i,\eta}\|_{L^2(\Omega)} \leq \eta, \tag{56}$$

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 we have

$$\varphi_i = g_{s_i}^{j_i,e_i} = \Big(|\det(A_{s_i,j_i})|^{\frac{1}{2}} g_{s_i}(A_{s_i,j_i}\cdot - \delta e_i)\Big)\big|_\Omega,$$

for $s_i \in \{1,\dots,S\}$, $j_i \in J_{s_i}$, and $e_i \in \mathbb{Z}^d$. In order to devise networks satisfying (56), we employ Proposition VIII.2, upon noting that, by virtue of (55), the networks $\Phi_{s,\varepsilon}$ satisfy (52) with $p = 2$, $f = g_s$, for every $D \in \mathbb{R}_+$. Consequently Proposition VIII.2 yields a connectivity bound that is even slightly stronger than needed, as it is independent of i. It remains to ensure that the desired bound on $\mathcal{B}(\Phi_{i,\eta})$ holds. This is the case for $\|A_{s_i,j_i}\|_\infty$ and $\|e_i\|_\infty$ both bounded polynomially in i. In order to verify this, we first bound $\|e_i\|_\infty$ relative to $\|A_{s_i,j_i}\|_\infty$. As the generators $(g_s)_{s=1}^S$ are compactly supported by assumption, there exists $E \in \mathbb{R}_+$ such that, for every $s \in \{1,\dots,S\}$, the support of g_s is contained in $[-E,E]^d$. We thus get, for all $s \in \{1,\dots,S\}$, $j \in J_s$, and $e \in \mathbb{Z}^d$, that

$$\|\delta e\|_\infty \geq \sup_{x\in\Omega} \|A_{s,j}x\|_\infty + E \implies g_s^{j,e}(x) = 0,\ \forall x \in \Omega \implies g_s^{j,e} \notin \mathcal{D}_j.$$

Since Ω is bounded by assumption, there hence exists a constant $c = c(\Omega, (g_s)_{s=1}^S, \delta, d)$ such that, for all $s \in \{1,\dots,S\}$, $j \in J_s$, and $e \in \mathbb{Z}^d$, we have

$$g_s^{j,e} \in \mathcal{D}_j \implies \|e\|_\infty \leq c\|A_{s,j}\|_\infty.$$

It remains to show that $\|A_{s_i,j_i}\|_\infty$ is polynomially bounded in i. We start by claiming that, for every $s \in \{1,\dots,S\}$, there is a constant $c_s := c_s(\Omega,\delta,d) > 0$ such that

$$|\det(A_{s,j})| \leq c_s|\mathcal{D}_{s,j}|, \text{ for all } j \in J_s. \tag{57}$$

To verify this claim, first note that $|\mathcal{D}_{s,j}| \geq 1$, for all $s \in \{1,\dots,S\}, j \in J_s$, owing to the nondegeneracy condition. Thus, for every $s \in \{1,\dots,S\}$, $j \in J_s$, there exist $x_0 \in \Omega$ and $e_0 \in \mathbb{Z}^d$ such that $g_s^{j,e_0}(x_0) \neq 0$, which implies

$$g_s^{j,e}(x_0 + A_{s,j}^{-1}\delta(e-e_0)) = |\det(A_{s,j})|^{\frac{1}{2}} g_s(A_{s,j}x_0 - \delta e_0) = g_s^{j,e_0}(x_0) \neq 0.$$

We can therefore conclude that $x_0 + A_{s,j}^{-1}\delta(e - e_0) \in \Omega$ implies $g_s^{j,e} \in \mathcal{D}_{s,j}$. Consequently, we have

$$|\mathcal{D}_{s,j}| \geq |\{e \in \mathbb{Z}^d \colon x_0 + A_{s,j}^{-1}\delta(e - e_0) \in \Omega\}| = |\{e \in \mathbb{Z}^d \colon A_{s,j}^{-1}\delta e \in \Omega - x_0\}| = |\mathbb{Z}^d \cap \tfrac{1}{\delta}A_{s,j}(\Omega - x_0)|.$$

As Ω was assumed to have nonempty interior, there exists a constant $C = C(\Omega)$ such that

$$|\mathbb{Z}^d \cap \tfrac{1}{\delta}A_{s,j}(\Omega - x_0)| \geq C\, vol\left(\tfrac{1}{\delta}A_{s,j}(\Omega - x_0)\right) = C\,\delta^{-d}|\det(A_{s,j})|\, vol(\Omega).$$

We have hence established the claim (57). Combining (51) and (57), we obtain, for all $s_i \in \{1, \dots, S\}$, $j \in J_s \backslash \{1\}$,

$$c\|A_{s_i,j_i}\|_\infty^a \leq \sum_{k=1}^{j_i-1} |\det(A_{s_i,k})| \leq c_{s_i} \sum_{k=1}^{j_i-1} |\mathcal{D}_{k,s_i}| \leq c_s i,$$

where the last inequality follows from the fact that $\varphi_i \in \mathcal{D}_{j_i,s_i}$ and hence its index i must be larger than the number of elements contained in preceding sub-dictionaries. This ensures that

$$\|A_{s_i,j_i}\|_\infty \leq \left(\frac{1}{c} \max_{s=1,\dots,S} c_s\right)^{\frac{1}{a}} i^{\frac{1}{a}} + \max_{s=1,\dots,S} \|A_{s,1}\|_\infty, \quad \text{for all } i \in \mathbb{N},$$

thereby completing the proof. □

Remark VIII.4. *Theorem VIII.3 is restricted, for ease of exposition, to bounded Ω and compactly supported generator functions g_s. The result can be extended to $\Omega = \mathbb{R}^d$ and to generator functions g_s 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 instead refer to the proofs of Theorems IX.3 and IX.5, which deal with the corresponding technical aspects in the context of approximation of Gabor dictionaries by neural networks.*

We can now put the results together to conclude a remarkable universality and optimality property of neural networks: Consider an affine dictionary generated by functions g_s that can be approximated well by neural networks. If this dictionary provides Kolmogorov-Donoho-optimal approximation for a given function class, then so do neural networks.

Theorem VIII.5. *Let $d, S \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ be bounded with nonempty interior, $(g_s)_{s=1}^S \in L^\infty(\mathbb{R}^d)$ compactly supported, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ a nondegenerate canonically ordered affine dictionary with generator functions $(g_s)_{s=1}^S$. Assume that there exists a polynomial π such that, for all $s \in \{1, \dots, S\}$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{s,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying $\|g_s - \Phi_{s,\varepsilon}\|_{L^2(\mathbb{R}^d)} \leq \varepsilon$ with $\mathcal{M}(\Phi_{s,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_{s,\varepsilon}) \leq \pi(\varepsilon^{-1})$. Then, we have*

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

for all compact 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.4), then $\mathcal{C}$ is optimally representable by neural networks (in the sense of Definition VI.5).

Proof. The first statement follows from Theorem VII.2 and Theorem VIII.3, the second from Theorem VI.4. □

D. Spline wavelets

We next particularize the results developed above to show that neural networks Kolmogorov-Donoho optimally represent all function classes $\mathcal{C}$ that are optimally representable by spline wavelet dictionaries. As spline wavelet dictionaries have B-splines as generator functions, we start by showing how B-splines can be realized through neural networks. For simplicity of exposition, we restrict ourselves to the univariate case throughout.

Definition VIII.6. *Let $N_1 := \chi_{[0,1]}$ and for $m \in \mathbb{N}$, define*

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

where $*$ stands for convolution. We refer to N_m as the univariate cardinal B-spline of order m.

Recognizing that B-splines are piecewise polynomial, we can build on Proposition III.5 to get the following statement on the approximation of B-splines by deep neural networks.

Lemma VIII.7. *Let $m \in \mathbb{N}$. There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ satisfying*

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

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

Proof. The proof is based on the following representation [81, Eq. 19]

$$N_m(x) = \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho((x-k)^m). \tag{58}$$

While N_m is supported on $[0, m]$, the networks Φ_ε can have support outside $[0, m]$ as well. We only need to ensure that Φ_ε is "close" to N_m on $[0, m]$ and at the same time "small" outside the interval $[0, m]$. To accomplish this, we first approximate N_m on the slightly larger domain $[-1, m+1]$ by a linear combination of networks realizing shifted monomials according to (58), and then multiply the resulting network by another one that takes on the value 1 on $[0, m]$ and 0 outside of $[-1, m+1]$. Specifically, we proceed as follows. Proposition III.5 ensures the existence of a constant C_1 such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{m+2,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\Psi_{m+2,\varepsilon}(x) - x^m\|_{L^\infty([-(m+2),m+2])} \leq \tfrac{\varepsilon}{4(m+2)},$$

with $\mathcal{M}(\Psi_{m+2,\varepsilon}) \leq C_1 \log(\varepsilon^{-1})$ and $\mathcal{B}(\Psi_{m+2,\varepsilon}) \leq 1$. Note that we did not make the dependence of $\mathcal{M}(\Psi_{m+2,\varepsilon})$ on m explicit as we consider m to be fixed. Next, let $T_k(x) := x - k$ and observe that $\rho((x-k)^m)$ can be realized as a neural network according to $\rho \circ \Psi_{m+2,\varepsilon} \circ T_k$, where T_k is taken pursuant to Corollary A.2. Next, we define, for $\varepsilon \in (0, 1/2)$, the network

$$\widetilde{\Phi}_\varepsilon := \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho \circ \Psi_{m+2,\varepsilon} \circ T_k$$

and note that

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

for $k = 0, \dots, m+1$. As ρ is 1-Lipschitz, we have, for all $\varepsilon \in (0, 1/2)$,

$$\begin{aligned} \|\widetilde{\Phi}_\varepsilon - N_m\|_{L^\infty([-1,m+1])} &\leq \sum_{k=0}^{m+1} \frac{1}{m!}\binom{m+1}{k} \|\rho \circ \Psi_{m+2,\varepsilon} \circ T_k - \rho \circ T_k^m\|_{L^\infty([-1,m+1])} \\ &\leq 2 \sum_{k=0}^{m+1} \|\Psi_{m+2,\varepsilon}(x) - x^m\|_{L^\infty([-(m+2),m+2])} \leq \tfrac{\varepsilon}{2}. \end{aligned} \tag{59}$$

Let now $\Gamma(x) := \rho(x+1) - \rho(x) - \rho(x-m) + \rho(x-(m+1))$, note that $0 \leq \Gamma(x) \leq 1$, and take $\Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2}$ to be the multiplication network from Lemma III.3. We define $\Phi_\varepsilon := \Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2} \circ (\widetilde{\Phi}_\varepsilon, \Gamma)$ according to Lemma II.3 and Lemma A.7 and note that

$$\|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})} \leq \|\Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2} \circ (\widetilde{\Phi}_\varepsilon, \Gamma) - \widetilde{\Phi}_\varepsilon \cdot \Gamma\|_{L^\infty([-1,m+1])} + \|\widetilde{\Phi}_\varepsilon \cdot \Gamma - N_m\|_{L^\infty([-1,m+1])} \tag{60}$$

as both N_m and Γ vanish outside $[-1, m+1]$ and $\Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2}$ delivers zero whenever at least one of its inputs is zero. Note that the first term on the right-hand-side of (60) is upper-bounded by $\frac{\varepsilon}{2}$ as a consequence of $N_m(x) \leq 1$ and hence $\widetilde{\Phi}_\varepsilon(x) \leq 1 + \frac{\varepsilon}{2}$, for $x \in [-1, m+1]$, owing to (59). For the second term, we split up the interval $[-1, m+1]$ and first note that, for $x \in [0, m]$, $\Gamma(x) = 1$, which implies $\|\widetilde{\Phi}_\varepsilon \cdot \Gamma - N_m\|_{L^\infty([0,m])} = \|\widetilde{\Phi}_\varepsilon - N_m\|_{L^\infty([0,m])} \leq \varepsilon/2$, again owing to (59). For $x \in [-1, m+1] \setminus [0, m]$, we have $N_m(x) = 0$ and $\Gamma(x) \leq 1$, which yields

$$|\widetilde{\Phi}_\varepsilon(x) \cdot \Gamma(x) - N_m(x)| \leq |\widetilde{\Phi}_\varepsilon(x)| \leq |\widetilde{\Phi}_\varepsilon(x) - N_m(x)| + |N_m(x)| = |\widetilde{\Phi}_\varepsilon(x) - N_m(x)| \leq \varepsilon/2,$$

again by (59). In summary, (59) hence ensures that the second term in (60) is also upper-bounded by $\frac{\varepsilon}{2}$ and therefore $\|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})} \leq \varepsilon$. Combining Lemma II.3, Proposition III.3, Corollary A.2, Lemma A.4, and Lemma A.7 establishes the desired bounds on $\mathcal{M}(\Phi_{D,\varepsilon})$ and $\mathcal{B}(\Phi_{D,\varepsilon})$. □

Remark VIII.8. *As both N_m and the approximating networks Φ_ε we constructed in the proof of Lemma VIII.7 are supported in $[-1, m+1]$, we have $\|\Phi_\varepsilon - N_m\|_{L^2(\mathbb{R})} \leq (m+2)^{1/2}\|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})}$, which shows that Lemma VIII.7 continues to hold when the approximation error is measured in $L^2(\mathbb{R})$-norm, albeit with a different constant C.*

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

$$V_n := \mathrm{clos}_{L^2}\Big(\operatorname{span}\{N_m(2^n x - k) : k \in \mathbb{Z}\}\Big),$$

where clos_{L^2} denotes closure with respect to L^2-norm. Spline spaces V_n, $n \in \mathbb{Z}$, constitute a multiresolution analysis [82] of $L^2(\mathbb{R})$ according to

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

Moreover, with the orthogonal complements $(\dots, W_{-1}, W_0, W_1, \dots)$ such that $V_{n+1} = V_n \oplus W_n$, where $\oplus$ denotes the orthogonal sum, we have

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

Theorem VIII.9 ([83, Theorem 1])**.** *Let $m \in \mathbb{N}$. The m-th order spline*

$$\psi_m(x) = \frac{1}{2^{m-1}} \sum_{j=0}^{2m-2} (-1)^j N_{2m}(j+1) \frac{d^m}{dx^m} N_{2m}(2x-j), \tag{61}$$

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

$$\mathcal{W}_m := \{\psi_{k,n}(x) = 2^{n/2}\psi_m(2^n x - k) : n \in \mathbb{N}_0, k \in \mathbb{Z}\} \cup \{\phi_k(x) = N_m(x-k) : k \in \mathbb{Z}\} \tag{62}$$

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

Taking $\Omega \subseteq \mathbb{R}$, $S = 2$, $J_1 = \mathbb{N}$, $J_2 = \{1\}$, $A_{1,j} = 2^{j-1}$ for $j \in \mathbb{N}$, and $A_{2,1} = 1$, we get that

$$\mathcal{D} := \left\{ g_s^{j,e}(x) := \left(|A_j|^{\frac{1}{2}} g_s(A_j \cdot - \delta e) \right)\Big|_\Omega : s \in \{1,2\},\ e \in \mathbb{Z},\ j \in J_s, \text{ and } g_s^{j,e} \neq 0 \right\} = \mathcal{W}_m \tag{63}$$

is a nondegenerate canonically ordered affine dictionary with generators $g_1 = \psi_m$ and $g_2 = N_m$. The canonical ordering condition (51) is satisfied with $a = 1$ and $c = 1/2$. Nondegeneracy follows upon noting that $\operatorname{supp}(\psi_{k,n}) = [2^{-n}k, 2^{-n}(2m-1+k)]$ and $\operatorname{supp}(N_m(\cdot - k)) = [k, m+k]$, which implies that all sub-dictionaries contain at least one element as required.

We have therefore established the following.

Theorem VIII.10. *Let $\Omega \subseteq \mathbb{R}$ be bounded and of nonempty interior and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ a spline wavelet dictionary according to* (63) *ordered per* (50)*. Then, all compact function classes $\mathcal{C} \subseteq L^2(\Omega)$ that are optimally representable by $\mathcal{D}$ (in the sense of Definition V.4) are optimally representable by neural networks (in the sense of Definition VI.5).*

Proof. As the canonical ordering and the nondegeneracy conditions were already verified, it remains to establish that the generators ψ_m and N_m satisfy the antecedent of Theorem VIII.3. To this end, we first devise an alternative representation of (61). Specifically, using the identity [83, 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

$$\psi_m(x) = \sum_{n=1}^{3m-1} q_n N_m(2x-n+1), \tag{64}$$

with

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

As (64) shows that ψ_m is a linear combination of shifts and dilations of N_m, combining Lemma VIII.7 and Remark VIII.8 with Lemma II.6 and Proposition VIII.2 ensures that (55) is satisfied. Application of Theorem VIII.5 then establishes the claim. □

IX. WEYL-HEISENBERG DICTIONARIES

In this section, we consider Weyl-Heisenberg a.k.a. Gabor dictionaries [17], which consist of time-frequency translates of a given generator function. Gabor dictionaries play a fundamental role in time-frequency analysis [17] and in the study of partial differential equations [84]. We start with the formal definition of Gabor dictionaries.

Definition IX.1 (Gabor dictionaries)**.** *Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d)$, and $x, \xi \in \mathbb{R}^d$. We define the translation operator $T_x \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ as*

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

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 dictionary $\mathcal{G}(g, \alpha, \beta, \Omega) \subseteq L^2(\Omega)$ is defined as

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

In order to describe representability in neural networks in the sense of Definition VII.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 $\mathcal{G}(g, \alpha, \beta, \Omega)$ as

$$\mathcal{G}(g, \alpha, \beta, \Omega) = (\mathcal{G}_0(g, \alpha, \beta, \Omega), \mathcal{G}_1(g, \alpha, \beta, \Omega), \dots), \tag{65}$$

where the ordering within the sets $\mathcal{G}_n(g, \alpha, \beta, \Omega)$ is arbitrary. We hasten to add that the specifics of the overall ordering in (65) 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 polynomially in i; this will become apparent in the proof of Theorem IX.3. We note that this ordering is also inspired by that employed in the tail compactness considerations for modulation spaces and Wilson bases as detailed in Appendix C.

As Gabor dictionaries are built from time-shifted and modulated versions of the generator function g, and invariance to time-shifts was already established in Proposition VIII.2, we proceed to showing that the approximation-theoretic properties of the generator function are inherited by its modulated versions. This result can be interpreted as an invariance property to frequency shifts akin to that established in Proposition VIII.2 for affine transformations in the context of affine dictionaries. In summary, neural networks exhibit a remarkable invariance property both to the affine group operations of scaling and translation and to the Weyl-Heisenberg group operations of modulation and translation.

Lemma IX.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 $\Phi_{D,\varepsilon} \in \mathcal{N}_{d,1}$ satisfy*

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

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

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

and $\mathcal{B}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 1$, where $S_f := \max\{1, \|f\|_{L^\infty(\mathbb{R}^d)}\}$.

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

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

due 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. Hence, combining Lemma II.3, Theorem III.8, and Corollary A.2 establishes 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 $\Psi_{D,\xi,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

$$\sup_{t \in [-D,D]^d} |\cos(2\pi\langle \xi, t\rangle) - \Psi_{D,\xi,\varepsilon}(t)| \leq \tfrac{\varepsilon}{6S_f} \tag{66}$$

with

$$\begin{aligned} \mathcal{L}(\Psi_{D,\xi,\varepsilon}) &\leq C_1((\log(\varepsilon^{-1}))^2 + (\log(S_f))^2 + \log(\lceil dD\|\xi\|_\infty \rceil)), \\ \mathcal{M}(\Psi_{D,\xi,\varepsilon}) &\leq C_1((\log(\varepsilon^{-1}))^2 + (\log(S_f))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) + d), \end{aligned} \tag{67}$$

and $\mathcal{B}(\Psi_{D,\xi,\varepsilon}) \leq 1$. Moreover, Proposition III.3 guarantees the existence of a constant $C_2 > 0$ such that for all $\varepsilon \in (0,1/2)$, there is a network $\mu_\varepsilon \in \mathcal{N}_{2,1}$ satisfying

$$\sup_{x,y \in [-S_f - 1/2, S_f + 1/2]} |\mu_\varepsilon(x,y) - xy| \leq \tfrac{\varepsilon}{6} \tag{68}$$

with

$$\mathcal{L}(\mu_\varepsilon), \mathcal{M}(\mu_\varepsilon) \leq C_2(\log(\varepsilon^{-1}) + \log(\lceil S_f \rceil)) \tag{69}$$

and $\mathcal{B}(\mu_\varepsilon) \leq 1$. Using Lemmas II.4 and II.5, we get that the network $\Gamma_{D,\xi,\varepsilon} := (\Psi_{D,\xi,\varepsilon}, \Phi_{D,\varepsilon}) \in \mathcal{N}_{d,2}$ satisfies

$$\mathcal{L}(\Gamma_{D,\xi,\varepsilon}) \leq \max\{\mathcal{L}(\Psi_{D,\xi,\varepsilon}), \mathcal{L}(\Phi_{D,\varepsilon})\},$$
$$\mathcal{M}(\Gamma_{D,\xi,\varepsilon}) \leq 2\,\mathcal{M}(\Psi_{D,\xi,\varepsilon}) + 2\,\mathcal{M}(\Phi_{D,\varepsilon}) + 2\,\mathcal{L}(\Psi_{D,\xi,\varepsilon}) + 2\,\mathcal{L}(\Phi_{D,\varepsilon}),$$

and $\mathcal{B}(\Gamma_{D,\xi,\varepsilon}) \leq 1$. Finally, applying Lemma II.3 to concatenate the networks $\Gamma_{D,\xi,\varepsilon}$ and μ_ε, we obtain the network

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

satisfying

$$\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq \max\{\mathcal{L}(\Psi_{D,\xi,\varepsilon}), \mathcal{L}(\Phi_{D,\varepsilon})\} + \mathcal{L}(\mu_\varepsilon), \tag{70}$$

$$\mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 4\mathcal{M}(\Psi_{D,\xi,\varepsilon}) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Psi_{D,\xi,\varepsilon}) + 4\mathcal{L}(\Phi_{D,\varepsilon}) + 2\mathcal{M}(\mu_\varepsilon), \tag{71}$$

and $\mathcal{B}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 1$. Next, observe that (66) and (68) imply that

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

Combining (67), (69), (71), and (70) we can further see that there exists a constant $C > 0$ such that

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

and $\mathcal{B}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon})) \leq 1$. The results for $\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}$ follow analogously, simply by using $\sin(x) = \cos(x - \pi/2)$. □

Note that Gabor dictionaries necessarily contain complex-valued functions. The theory developed so far was, however, phrased for neural networks with real-valued outputs. As is evident from the proof of Lemma IX.2, this is not problematic when the generator function g is real-valued. For complex-valued generator functions we would need a version of Proposition III.3 that applies to the multiplication of complex numbers. Due 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 by neural networks according to Definition VII.1 would be needed. This can be 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

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

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

We proceed to establish conditions for effective representability of Gabor dictionaries by neural networks.

Theorem IX.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 dictionary with ordering as defined in* (65)*. Assume that Ω is bounded or that $\Omega = \mathbb{R}^d$ and g is compactly supported. Further, suppose that there exists a polynomial π such that for every $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{x,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|g - \Phi_{x,\varepsilon}\|_{L^\infty(x+\Omega)} \leq \varepsilon, \tag{72}$$

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

Proof. We start by noting that owing to (65), 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

$$\|\xi(i)\|_\infty \leq n(i)\beta \leq i\beta \quad \text{and} \quad \|x(i)\|_\infty \leq n(i)\alpha \leq i\alpha. \tag{73}$$

Next, we take the affine transformation $W_x(y) := y - x$ to be a depth-1 network and observe that, due to (72) and Lemma II.3, we have, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$,

$$\|T_x g - \Phi_{-x,\varepsilon} \circ W_x\|_{L^\infty(\Omega)} = \|g - \Phi_{-x,\varepsilon}\|_{L^\infty(-x+\Omega)} \leq \varepsilon, \tag{74}$$

with

$$\begin{aligned} \mathcal{M}(\Phi_{-x,\varepsilon} \circ W_x) &\leq 2(\pi(\log(\varepsilon^{-1}), \log(\|x\|_\infty)) + 2d) \\ \mathcal{B}(\mathcal{M}(\Phi_{-x,\varepsilon} \circ W_x)) &\leq \max\{\mathcal{B}(\Phi_{-x,\varepsilon}), \|x\|_\infty\} \leq \pi(\varepsilon^{-1}, \|x\|_\infty) + \|x\|_\infty. \end{aligned}$$

We first consider the case where Ω is bounded and let $E \in \mathbb{R}_+$ be such that $\Omega \subseteq [-E, E]^d$. Combining (74) with Proposition VIII.2 and Lemma IX.2, 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 $\Phi_{i,\varepsilon} = (\Phi^{\mathrm{Re}}_{i,\varepsilon}, \Phi^{\mathrm{Im}}_{i,\varepsilon}) \in \mathcal{N}^{\mathbb{C}}_{d,1}$ satisfying

$$\|\mathrm{Re}(\mathcal{M}_{\xi(i)} T_{x(i)} g) - \Phi^{\mathrm{Re}}_{i,\varepsilon}\|_{L^\infty(\Omega)} + \|\mathrm{Im}(\mathcal{M}_{\xi(i)} T_{x(i)} g) - \Phi^{\mathrm{Im}}_{i,\varepsilon}\|_{L^\infty(\Omega)} \leq (2E)^{-\frac{d}{2}} \varepsilon, \tag{75}$$

with

$$\begin{aligned} \mathcal{M}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{M}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) &\leq \pi_1(\log(\varepsilon^{-1}), \log(\|\xi(i)\|_\infty), \log(\|x(i)\|_\infty)), \\ \mathcal{B}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{B}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) &\leq \pi_1(\varepsilon^{-1}, \|\xi(i)\|_\infty, \|x(i)\|_\infty). \end{aligned} \tag{76}$$

Note that here we did not make the dependence of the connectivity and the weight upper bounds on d and E explicit as these quantities are irrelevant for the purposes of what we want to show, as long as they are finite, of

course, which is the case by assumption. Likewise, we did not explicitly indicate the dependence of π_1 on g. As $|z| \leq |\mathrm{Re}(z)| + |\mathrm{Im}(z)|$, it follows from (75) that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$,

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

Moreover, (73) and (76) imply the existence of a polynomial π_2 such that

$$\mathcal{M}(\Phi_{i,\varepsilon}^{\mathrm{Re}}), \mathcal{M}(\Phi_{i,\varepsilon}^{\mathrm{Im}}) \leq \pi_2(\log(\varepsilon^{-1}), \log(i)), \quad \mathcal{B}(\Phi_{i,\varepsilon}^{\mathrm{Re}}), \mathcal{B}(\Phi_{i,\varepsilon}^{\mathrm{Im}}) \leq \pi_2(\varepsilon^{-1}, i),$$

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 $\Omega = \mathbb{R}^d$ and g compactly supported, i.e., there exists $E \in \mathbb{R}_+$ such that $\mathrm{supp}(g) \subseteq [-E, E]^d$. This implies

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

Again, combining (74) with Proposition VIII.2 and Lemma IX.2 establishes 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 $\Psi_{x,\xi,\varepsilon}^{\mathrm{Re}}, \Psi_{x,\xi,\varepsilon}^{\mathrm{Im}} \in \mathcal{N}_{d,1}$ satisfying

$$\|\mathrm{Re}(M_\xi T_x g) - \Psi_{x,\xi,\varepsilon}^{\mathrm{Re}}\|_{L^\infty(S_x)} + \|\mathrm{Im}(M_\xi T_x g) - \Psi_{x,\xi,\varepsilon}^{\mathrm{Im}}\|_{L^\infty(S_x)} \leq \tfrac{\varepsilon}{2s_x}, \tag{77}$$

with

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

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 $\Omega = \mathbb{R}^d$, the estimate in (77) is insufficient. In particular, we have no control over the behavior of the networks $\Psi_{x,\xi,\varepsilon}^{\mathrm{Re}}, \Psi_{x,\xi,\varepsilon}^{\mathrm{Im}}$ 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 supported in S_x, and realize the same output for all inputs in S_x. To this end let, for $y \in \mathbb{R}_+$, the network $\alpha_y \in \mathcal{N}_{1,1}$ 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 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.3 there exists a constant C_2 such that, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\mu_{x,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\sup_{y,z \in [-2B, 2B]} |\mu_{x,\varepsilon}(y,z) - yz| \leq \tfrac{\varepsilon}{4s_x}, \tag{78}$$

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

$$\mu_{x,\varepsilon}(0,y) = \mu_{x,\varepsilon}(y,0) = 0, \tag{79}$$

with $\mathcal{M}(\mu_{x,\varepsilon}) \leq C_2(\log(\varepsilon^{-1}) + \log(s_x))$ and $\mathcal{B}(\mu_{x,\varepsilon}) \leq 1$. Note that in the upper bound on $\mathcal{M}(\mu_{x,\varepsilon})$, we did not make the dependence on B explicit as we consider g fixed for the purposes of the proof. Next, as E is 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 $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$.

We now take

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

according to Lemmas II.5 and II.3, which ensures the existence of a polynomial π_4 such that, for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}\mathcal{M}(\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{M}(\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}) &\leq \pi_4(\log(\varepsilon^{-1}), \log(\|x\|_\infty), \log(\|\xi\|_\infty)), \\ \mathcal{B}(\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{B}(\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}) &\leq \pi_4(\varepsilon^{-1}, \|x\|_\infty, \|\xi\|_\infty).\end{aligned} \tag{80}$$

Furthermore,

$$\begin{aligned}\|\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon} - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)} \leq{}& \|\mu_{x,\varepsilon} \circ (\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \chi_x) - \Psi^{\mathrm{Re}}_{x,\xi,\varepsilon} \cdot \chi_x\|_{L^\infty(S_x)} \\ &+ \|\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon} \cdot \chi_x - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)},\end{aligned} \tag{81}$$

where the first term is upper-bounded by $\frac{\varepsilon}{4s_x}$ due to (78). The second term on the right-hand side of (81) is upper-bounded as follows. First, note that for $t \in S_x \setminus [-(\|x\|_\infty + E), \|x\|_\infty + E]^d$, we have $\mathrm{Re}(M_\xi T_x g)(t) = 0$ and $|\chi_x(t)| \leq 1$, which implies

$$\begin{aligned}|\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t) \cdot \chi_x(t) - \mathrm{Re}(M_\xi T_x g)(t)| &\leq |\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t)| \leq |\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t) - \mathrm{Re}(M_\xi T_x g)(t)| + |\mathrm{Re}(M_\xi T_x g)(t)| \\ &= |\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t) - \mathrm{Re}(M_\xi T_x g)(t)|.\end{aligned}$$

As $|\chi_x(t)| = 1$ for $t \in [-(\|x\|_\infty + E), \|x\|_\infty + E]^d$, together with (81), this yields

$$\|\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon} - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)} \leq \tfrac{\varepsilon}{4s_x} + \|\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon} - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)}.$$

The analogous estimate for $\|\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon} - \mathrm{Im}(M_\xi T_x g)\|_{L^\infty(S_x)}$ is obtained in exactly the same manner. Together with (77), we can finally infer that, for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$,

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

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 (79), we get

$$\begin{aligned} &\|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^2(\mathbb{R}^d)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^2(\mathbb{R}^d)} \\ &= \|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^2(S_x)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^2(S_x)} \\ &\leq s_x \|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} + s_x \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \leq \varepsilon. \end{aligned} \tag{82}$$

Consider now, for $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, the complex-valued network $\Gamma_{i,\varepsilon} \in \mathcal{N}^{\mathbb{C}}_{d,1}$ 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, (82) 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 (73) in (80), 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 finalizes the proof. □

Next, we establish the central result of this section. To this end, we first recall that according to Theorem VIII.5 neural networks provide optimal approximations for all function classes that are optimally approximated by affine dictionaries (generated by functions f that can be approximated well by neural networks). While this universality property is significant as it applies to all affine dictionaries, it is perhaps not completely surprising as affine dictionaries are generated by affine transformations and neural networks consist of concatenations of affine transformations and nonlinearities. Gabor dictionaries, on the other hand, 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 dictionaries (again, with generator functions that can be approximated well by neural networks).

Theorem IX.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 dictionary with ordering as defined in* (65)*. Assume that Ω is bounded or that $\Omega = \mathbb{R}^d$ and g is compactly supported. Further, suppose that there exists a polynomial π such that for every $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{x,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|g - \Phi_{x,\varepsilon}\|_{L^\infty(x+\Omega)} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{x,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\|x\|_\infty))$, $\mathcal{B}(\Phi_{x,\varepsilon}) \leq \pi(\varepsilon^{-1}, \|x\|_\infty)$. Then, for all compact function classes $\mathcal{C} \subseteq L^2(\Omega)$, we have

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

In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{G}(g, \alpha, \beta, \Omega)$ (in the sense of Definition V.4), then $\mathcal{C}$ is optimally representable by neural networks (in the sense of Definition VI.5).

Proof. The first statement follows from Theorem VII.2 and Theorem IX.3, the second is by Theorem VI.4. □

We complete the program in this section by showing that the Gaussian function satisfies the conditions on the generator g in Theorem IX.3 for bounded Ω. Gaussian functions are widely used generator functions for Gabor dictionaries owing to their excellent time-frequency localization and their frame-theoretic optimality properties [17]. We hasten to add that the result below can be extended to any generator function g of sufficiently fast decay and sufficient smoothness.

Lemma IX.5. *For $d \in \mathbb{N}$, let $g_d \in L^2(\mathbb{R}^d)$ be given by*

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

There exists a constant $C > 0$ such that, for all $d \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{d,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

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

with $\mathcal{M}(\Phi_{d,\varepsilon}) \leq Cd(\log(\varepsilon^{-1}))^2((\log(\varepsilon^{-1}))^2 + \log(d))$, $\mathcal{B}(\Phi_{d,\varepsilon}) \leq 1$.

Proof. Observe that g_d can be written as the composition $h \circ f_d$ of the functions $f_d \colon \mathbb{R}^d \to \mathbb{R}_+$ and $h \colon \mathbb{R}_+ \to \mathbb{R}$ given by

$$f_d(x) := \|x\|_2^2 = \sum_{i=1}^d x_i^2 \quad \text{and} \quad h(y) := e^{-y}.$$

By Proposition III.3 and Lemma II.6, there exists a constant $C_1 > 0$ such that, for every $d \in \mathbb{N}$, $D \in [1, \infty)$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{d,D,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

$$\sup_{x \in [-D,D]^d} |\Psi_{d,D,\varepsilon}(x) - \|x\|_2^2| \leq \tfrac{\varepsilon}{2}, \tag{83}$$

$$\mathcal{M}(\Psi_{d,D,\varepsilon}) \leq C_1 d(\log(\varepsilon^{-1}) + \log(\lceil D \rceil)), \quad \mathcal{B}(\Psi_{d,D,\varepsilon}) \leq 1. \tag{84}$$

Moreover, as $|\frac{d^n}{dy^n}e^{-y}| = |e^{-y}| \leq 1$ for all $n \in \mathbb{N}$, $y \geq 0$, Lemma A.6 implies the existence of a constant $C_2 > 0$ such that for every $d \in \mathbb{N}$, $D \in [1, \infty)$, $\varepsilon \in (0, 1/2)$, there is a network $\Gamma_{d,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\sup_{y \in [0, dD^2]} |\Gamma_{d,D,\varepsilon}(y) - e^{-y}| \leq \tfrac{\varepsilon}{2}, \tag{85}$$

$$\mathcal{M}(\Gamma_{d,D,\varepsilon}) \leq C_2 dD^2((\log(\varepsilon^{-1}))^2 + \log(d) + \log(\lceil D \rceil)), \quad \mathcal{B}(\Gamma_{D,\varepsilon}) \leq 1. \tag{86}$$

Now, let $D_\varepsilon := \log(\varepsilon^{-1})$ and take $\widetilde{\Phi}_{d,\varepsilon} := \Gamma_{d,D_\varepsilon,\varepsilon} \circ \Psi_{d,D_\varepsilon,\varepsilon}$ according to Lemma II.3. Consequently, it follows from (84) and (86) that there exists a constant $C_2 > 0$ such that for all $d \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, we have $\mathcal{M}(\widetilde{\Phi}_{d,\varepsilon}) \leq C_2 d(\log(\varepsilon^{-1}))^2((\log(\varepsilon^{-1}))^2 + \log(d))$ and $\mathcal{B}(\widetilde{\Phi}_{d,\varepsilon}) \leq 1$. Moreover, as $|e^{-y}| \leq 1$ for all $y \geq 0$, combining (83) and (85) yields for all $\varepsilon \in (0, 1/2)$, $x \in [-D_\varepsilon, D_\varepsilon]^d$,

$$\begin{aligned} |g(x) - \widetilde{\Phi}_{d,\varepsilon}(x)| &= |e^{-\|x\|_2^2} - \Gamma_{d,D_\varepsilon,\varepsilon}(\Psi_{d,D_\varepsilon,\varepsilon}(x))| \\ &\leq |e^{-\|x\|_2^2} - e^{-\Psi_{d,D_\varepsilon,\varepsilon}(x)}| + |e^{-\Psi_{d,D_\varepsilon,\varepsilon}(x)} - \Gamma_{d,D_\varepsilon,\varepsilon}(\Psi_{d,D_\varepsilon,\varepsilon}(x))| \\ &\leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

We can now use the same approach as in the proof of Theorem IX.3 to construct networks $\Phi_{d,\varepsilon}$ supported on the interval $[-D_\varepsilon, D_\varepsilon]^d$ over which they approximate g to within error ε, and obey $\mathcal{M}(\Phi_\varepsilon) \leq Cd(\log(\varepsilon^{-1}))^2((\log(\varepsilon^{-1}))^2 + \log(d))$, $\mathcal{B}(\Phi_{d,\varepsilon}) \leq 1$ for some absolute constant C. Together with $|g(x)| \leq \varepsilon$, for all $x \in \mathbb{R}^d \backslash [-D_\varepsilon, D_\varepsilon]^d$, this completes the proof. □

Remark IX.6. *Note that Lemma IX.5 establishes an approximation result that is even stronger than what is required by Theorem IX.3. Specifically, we achieve ε-approximation over all of $\mathbb{R}^d$ with a network that does not depend on the shift parameter x, while exhibiting the desired growth rates on $\mathcal{M}$ and $\mathcal{B}$, which consequently do not depend on the shift parameter as well. The idea underlying this construction can be used to strengthen Theorem IX.3 to apply to $\Omega = \mathbb{R}^d$ and generator functions of unbounded support, but sufficiently rapid decay.*

We conclude this section with a remark on the neural network approximation of the real-valued counterpart of Gabor dictionaries known as Wilson dictionaries [74], [17] and consisting of cosine-modulated and time-shifted versions of a given generator function, see also Appendix C. The techniques developed in this section, mutatis mutandis, show that neural networks provide Kolmogorov-Donoho optimal approximation for all function classes that are optimally approximated by Wilson dictionaries (generated by functions that can be approximated well by neural networks). Specifically, we point out that the proofs of Lemma IX.2 and Theorem IX.3 explicitly construct neural network approximations of time-shifted and cosine- and sine-modulated versions of the generator g. As identified in Table 1, Wilson bases provide optimal nonlinear approximation of (unit) balls in modulation spaces [85], [74]. Finally, we note that similarly the techniques developed in the proofs of Lemma IX.2 and Theorem IX.3 can be used to establish optimal representability of Fourier bases.

X. Improving Polynomial Approximation Rates to Exponential Rates

Having established that for all function classes listed in Table 1, Kolmogorov-Donoho-optimal approximation through neural networks is possible, this section proceeds to show that neural networks, in addition to their striking Kolmogorov-Donoho universality property, can also do something that has no classical equivalent.

Specifically, as mentioned in the introduction, 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 networks fill this gap.

Let us start by defining one-dimensional "oscillatory textures" according to [18]. To this end, we recall the following definition from Lemma A.6,

$$\mathcal{S}_{[a,b]} = \left\{ f \in C^\infty([a,b],\mathbb{R})\colon \|f^{(n)}(x)\|_{L^\infty([a,b])} \leq n!, \ \textit{for all } n \in \mathbb{N}_0 \right\}.$$

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

$$\mathcal{F}_{D,a} = \left\{ \cos(ag)h\colon g,h \in \mathcal{S}_{[-D,D]} \right\}.$$

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 function g. The best approximation results available in the literature [18] are based on wave-atom dictionaries[11] 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.

We start with our statement on the neural network approximation of oscillatory textures.

Proposition X.2. *There exists a constant $C > 0$ 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 $\Gamma_{f,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying*

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

with $\mathcal{L}(\Gamma_{f,\varepsilon}) \leq C\lceil D\rceil((\log(\varepsilon^{-1}) + \log(\lceil a\rceil))^2 + \log(\lceil D\rceil) + \log(\lceil D^{-1}\rceil))$, $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 32$, $\mathcal{B}(\Gamma_{f,\varepsilon}) \leq 1$.

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

$$\|\Psi_{g_f,\varepsilon} - g_f\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{12\lceil a\rceil}, \quad \|\Psi_{h_f,\varepsilon} - h_f\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{12\lceil a\rceil} \tag{87}$$

[11] To be precise, the results of [18] are concerned with the two-dimensional case, whereas here we focus on the one-dimensional case. Note, however, that all our results are readily extended to the multi-dimensional case.

with

$$\mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon}) \leq C_1 \lceil D \rceil (\log((\tfrac{\varepsilon}{12\lceil a \rceil})^{-1})^2 + \log(\lceil D \rceil) + \log(\lceil D^{-1} \rceil)),$$

$\mathcal{W}(\Psi_{g_f,\varepsilon}), \mathcal{W}(\Psi_{h_f,\varepsilon}) \leq 16$, and $\mathcal{B}(\Psi_{g_f,\varepsilon}), \mathcal{B}(\Psi_{h_f,\varepsilon}) \leq 1$. Furthermore, Theorem III.8 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 $\Phi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\Phi_{a,D,\varepsilon} - \cos(a\,\cdot)\|_{L^\infty([-3/2,3/2])} \leq \tfrac{\varepsilon}{3}, \tag{88}$$

with $\mathcal{L}(\Phi_{a,D,\varepsilon}) \leq C_2((\log(\varepsilon^{-1}))^2 + \log(\lceil 3a/2 \rceil))$, $\mathcal{W}(\Phi_{a,D,\varepsilon}) \leq 9$, and $\mathcal{B}(\Phi_{a,D,\varepsilon}) \leq 1$. Moreover, due to Proposition III.3, there exists a constant $C_3 > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\mu_\varepsilon \in \mathcal{N}_{2,1}$ satisfying

$$\sup_{x,y \in [-3/2,3/2]} |\mu_\varepsilon(x, y) - xy| \leq \tfrac{\varepsilon}{3}, \tag{89}$$

with $\mathcal{L}(\mu_\varepsilon) \leq C_3 \log(\varepsilon^{-1})$, $\mathcal{W}(\mu_\varepsilon) \leq 5$, and $\mathcal{B}(\mu_\varepsilon) \leq 1$. By Lemma II.3 there exists a network Ψ^1 satisfying $\Psi^1 = \Phi_{a,D,\varepsilon} \circ \Psi_{g_f,\varepsilon}$ with $\mathcal{W}(\Psi^1) \leq 16$, $\mathcal{L}(\Psi^1) = \mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon})$, and $\mathcal{B}(\Psi^1) \leq 1$. Furthermore, combining Lemma II.4 and Lemma A.7, we can conclude the existence of a network $\Psi^2(x) = (\Psi^1(x), \Psi_{h_f,\varepsilon}(x)) = (\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x))$ with $\mathcal{W}(\Psi^2) \leq 32$, $\mathcal{L}(\Psi^2) = \max\{\mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon})\}$, and $\mathcal{B}(\Psi^2) \leq 1$. Next, for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, $\varepsilon \in (0, 1/2)$, we define the network $\Gamma_{f,\varepsilon} := \mu_\varepsilon \circ \Psi^2$. By (87), (88), and $\sup_{x \in \mathbb{R}} |\frac{d}{dx} \cos(ax)| = a$, we have, for all $x \in [-D, D]$,

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

Combining this with (87), (89), and $\|\cos\|_{L^\infty([-D,D])}, \|f\|_{L^\infty([-D,D])} \leq 1$ yields for all $x \in [-D, D]$,

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

Finally, by Lemma II.3 there exists a constant C_4 such that for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, $\varepsilon \in (0, 1/2)$, it holds that $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 32$,

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

and $\mathcal{B}(\Gamma_{f,\varepsilon}) \leq 1$. □

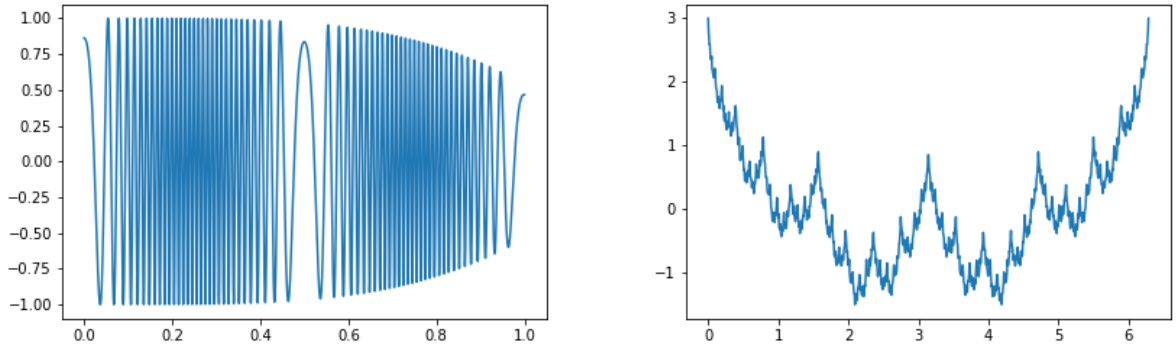


Fig. 4: Left: A function in $\mathcal{F}_{1,100}$. Right: The function $W_{\frac{1}{\sqrt{2}},2}$.

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

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

and let $\alpha = -\frac{\log(p)}{\log(a)}$, see Figure 4 right for an example. It is well known [86] that $W_{p,a}$ possesses Hölder smoothness α which may be made arbitrarily small by suitable choice of a. While classical approximation methods achieve polynomial approximation rates only, it turns out that finite-width deep networks yield exponential approximation rates. This is formalized as follows.

Proposition X.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 $\Psi_{p,a,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying*

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

with $\mathcal{L}(\Psi_{p,a,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^3 + (\log(\varepsilon^{-1}))^2 \log(\lceil a \rceil) + \log(\varepsilon^{-1}) \log(\lceil D \rceil))$, $\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 13$, $\mathcal{B}(\Psi_{p,a,D,\varepsilon}) \leq 1$.

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

$$|S_{N,p,a}(x) - W_{p,a}(x)| \leq \sum_{k=N+1}^{\infty} |p^k \cos(a^k \pi x)| \leq \sum_{k=N+1}^{\infty} p^k = \tfrac{1}{1-p} - \tfrac{1-p^{N+1}}{1-p} \leq 2^{-N}. \tag{90}$$

Let $N_\varepsilon := \lceil \log(2/\varepsilon) \rceil$ for $\varepsilon \in (0,1/2)$. Next, note that Theorem III.8 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 $\phi_{a^k,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\phi_{a^k,D,\varepsilon} - \cos(a^k \pi \cdot)\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{4}, \tag{91}$$

with $\mathcal{L}(\phi_{a^k,D,\varepsilon}) \leq C_1((\log(\varepsilon^{-1}))^2 + \log(\lceil a^k \pi D \rceil))$, $\mathcal{W}(\phi_{a^k,D,\varepsilon}) \leq 9$, $\mathcal{B}(\phi_{a^k,D,\varepsilon}) \leq 1$. Let $A\colon \mathbb{R}^3 \to \mathbb{R}^3$ and $B\colon \mathbb{R}^3 \to \mathbb{R}$ be the affine transformations given by $A(x_1,x_2,x_3) = (x_1, x_1, x_2+x_3)^T$ and $B(x_1,x_2,x_3) = x_2+x_3$, respectively. We now define, for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $k \in \mathbb{N}_0$, $\varepsilon \in (0,1/2)$, the networks

$$\psi_{D,\varepsilon}^{p,a,0}(x) = \begin{pmatrix} x \\ p^0 \phi_{a^0,D,\varepsilon}(x) \\ 0 \end{pmatrix} \quad \text{and} \quad \psi_{D,\varepsilon}^{p,a,k}(x_1,x_2,x_3) = \begin{pmatrix} x_1 \\ p^k \phi_{a^k,D,\varepsilon}(x_2) \\ x_3 \end{pmatrix}, \ k > 0,$$

and, for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, the network

$$\Psi_{p,a,D,\varepsilon} := B \circ \psi_{D,\varepsilon}^{p,a,N_\varepsilon} \circ A \circ \psi_{D,\varepsilon}^{p,a,N_\varepsilon - 1} \circ \cdots \circ A \circ \psi_{D,\varepsilon}^{p,a,0}.$$

Due to (91) we get, for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, $x \in [-D,D]$, that

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

Combining this with (90) establishes, for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$, $x \in [-D,D]$,

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

Applying Lemmas II.3, II.4, and II.5 establishes the existence of a constant C_2 such that for all $p \in (0,1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0,1/2)$,

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

$\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 13$, and $\mathcal{B}(\Psi_{p,a,D,\varepsilon}) \leq 1$. □

We finally note that the restriction $p \in (0,1/2)$ in Proposition X.3 was made for simplicity of exposition and can be relaxed to $p \in (0,r)$, with $r < 1$, while only changing the constant C.

XI. Impossibility results for finite-depth networks

The recent successes of neural networks in machine learning applications have been enabled by various technological factors, but they all have in common the use of deep networks as opposed to shallow networks studied intensely in the 1990s. It is hence of interest to understand whether the use of depth offers fundamental advantages. In this spirit, the goal of this section is to make a formal case for depth in neural network approximation by establishing that, for nonconstant 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 nonperiodic 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 [26].

Definition XI.1 ([26]). *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 of these intervals.*

Lemma XI.2 ([26]). *Every $\Phi \in \mathcal{N}_{1,1}$ is $(2\mathcal{W}(\Phi))^{\mathcal{L}(\Phi)}$-sawtooth.*

Definition XI.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 nonlinearity of f. The next result states that finite-depth networks with width and hence also connectivity scaling polylogarithmically in the "highest frequency" of the periodic function to be approximated can not achieve arbitrarily small approximation error.

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

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

Proof. First note that there exists an even $a \in \mathbb{N}$ such that $a/2 > (2\pi(\log(a)))^L$. Lemma XI.2 now implies that every network $\Phi \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi) \leq L$ and $\mathcal{W}(\Phi) \leq \pi(\log(a))$ is $(2\pi(\log(a)))^L$-sawtooth and therefore consists of no more than $a/2$ 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

$$\begin{aligned}\|f(a\,\cdot) - \Phi\|_{L^\infty([0,u])} &\geq \|f(a\,\cdot) - \Phi\|_{L^\infty([u_1,u_2])} \geq \inf_{c,d\in\mathbb{R}} \|f(x) - (cx+d)\|_{L^\infty([0,2u])} \\ &\geq \sup_{\delta\in[0,u)} \inf_{c,d\in\mathbb{R}} \|f(x) - (cx+d)\|_{L^\infty([\delta,u+\delta])} = \xi(f).\end{aligned}$$

Finally, note that $\xi(f) > 0$ as $\xi(f) = 0$ for u-periodic $f \in C(\mathbb{R})$ necessarily implies that f is constant, which, however, is ruled out by assumption. □

Application of Proposition XI.4 to $f(x) = \cos(x)$ shows that finite-depth networks, owing to $\xi(\cos) > 0$, require faster than polylogarithmic growth of connectivity in a to approximate $x \mapsto \cos(ax)$ with arbitrarily small error, whereas finite-width networks, due to Theorem III.8, can accomplish this with polylogarithmic connectivity growth.

The following result from [87] allows a similar observation for functions that are sufficiently smooth.

Theorem XI.5 ([87]). *Let $[a,b] \subseteq \mathbb{R}$, $f \in C^3([a,b])$, and for $\varepsilon \in (0,1/2)$, let $s(\varepsilon) \in \mathbb{N}$ denote the smallest number such that there exists a piecewise linear approximation of f with $s(\varepsilon)$ pieces and error at most ε in $L^\infty([a,b])$-norm. Then, it holds that*

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

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

Theorem XI.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 $\Phi \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi) \leq L$ and $\mathcal{W}(\Phi) \leq \pi(\log(\varepsilon^{-1}))$, we have*

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

Proof. The proof will be effected by contradiction. Assume that for every $\varepsilon > 0$, there exists a network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_\varepsilon) \leq L$, $\mathcal{W}(\Phi_\varepsilon) \leq \pi(\log(\varepsilon^{-1}))$, and $\|f - \Phi_\varepsilon\|_{L^\infty([a,b])} \leq \varepsilon$. By Lemma XI.2 every (ReLU) neural network realizes a piecewise linear function. Application of Theorem XI.5 hence allows us to conclude the existence of 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 XI.2, Φ_ε is at most $(2\pi(\log(\varepsilon^{-1})))^L$-sawtooth and $\tilde{\pi}(\log(\varepsilon^{-1})) \in o(\varepsilon^{-1/2})$, $\varepsilon \to 0$, for every polynomial $\tilde{\pi}$. □

In summary, we have hence established that any function which is at least three times continuously differentiable (and does not have a vanishing second derivative) cannot be approximated by finite-depth networks with connectivity scaling polylogarithmically in the inverse of the approximation error. Our results in Section III establish that, in contrast, this "is" possible with finite-width deep networks for various interesting types of smooth functions such as polynomials and sinusoidal functions. Further results on the limitations of finite-depth networks akin to Theorem XI.6 were reported in [23].

ACKNOWLEDGMENTS

The authors are indebted to R. Gül and W. Ou for their careful proofreading of the paper, to E. Riegler and the reviewers for their constructive and insightful comments, and to the handling editor, P. Narayan, for his helpful comments and his patience.

APPENDIX A
AUXILIARY NEURAL NETWORK CONSTRUCTIONS

The following three results are concerned with the realization of affine transformations of arbitrary weights by neural networks with weights upper-bounded by 1.

Lemma A.1. *Let $d \in \mathbb{N}$ and $a \in \mathbb{R}$. There exists a network $\Phi_a \in \mathcal{N}_{d,d}$ satisfying $\Phi_a(x) = ax$, with $\mathcal{L}(\Phi_a) \leq \lfloor \log(|a|) \rfloor + 4$, $\mathcal{W}(\Phi_a) \leq 3d$, $\mathcal{B}(\Phi_a) \leq 1$.*

Proof. First note that for $|a| \leq 1$ the claim holds trivially, which can be seen by taking Φ_a to be the affine transformation $x \mapsto ax$ and interpreting it according to Definition II.1 as a depth-1 neural network. Next, we consider the case $|a| > 1$ for $d = 1$, set $K := \lfloor \log(a) \rfloor$, $\alpha := a2^{-(K+1)}$, and define $A_1 := (1, -1)^T \in \mathbb{R}^{2\times 1}$,

$$A_2 := \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2}, \quad A_k := \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad k \in \{3, \ldots, K+3\},$$

and $A_{K+4} := (\alpha, 0, -\alpha)$. Note that $(\rho \circ A_2 \circ \rho \circ A_1)(x) = (\rho(x), \rho(x) + \rho(-x), \rho(-x))$ and $\rho(A_k(x, x+y, y)^T) = 2(x, x+y, y)$, for $k \in \{3, \ldots, K+3\}$. The network $\Psi_a := A_{K+4} \circ \rho \circ \cdots \circ \rho \circ A_1$ hence satisfies $\Psi_a(x) = ax$, $\mathcal{L}(\Psi_a) = \lfloor \log(a) \rfloor + 4$, $\mathcal{W}(\Psi_a) = 3$, and $\mathcal{B}(\Phi_a) \leq 1$. Applying Lemma II.5 to get a parallelization of d copies of Ψ_a completes the proof. □

Corollary A.2. *Let $d, d' \in \mathbb{N}$, $a \in \mathbb{R}_+$, $A \in [-a, a]^{d'\times d}$, and $b \in [-a, a]^{d'}$. There exists a network $\Phi_{A,b} \in \mathcal{N}_{d,d'}$ satisfying $\Phi_{A,b}(x) = Ax + b$, with $\mathcal{L}(\Phi_{A,b}) \leq \lfloor \log(|a|) \rfloor + 5$, $\mathcal{W}(\Phi_{A,b}) \leq \max\{d, 3d'\}$, $\mathcal{B}(\Phi_{A,b}) \leq 1$.*

Proof. Let $\Phi_a \in \mathcal{N}_{d',d'}$ be the multiplication network from Lemma A.1, consider $W(x) := a^{-1}(Ax + b)$ as a 1-layer network, and take $\Phi_{A,b} := \Phi_a \circ W$ according to Lemma II.3. □

Proposition A.3. *Let $d, d' \in \mathbb{N}$ and $\Phi \in \mathcal{N}_{d,d'}$. There exists a network $\Psi \in \mathcal{N}_{d,d'}$ satisfying $\Psi(x) = \Phi(x)$, for all $x \in \mathbb{R}^d$, and with $\mathcal{L}(\Psi) \leq (\lceil \log(\mathcal{B}(\Phi)) \rceil + 5)\mathcal{L}(\Phi)$, $\mathcal{W}(\Psi) \leq \max\{3d', \mathcal{W}(\Phi)\}$, $\mathcal{B}(\Psi) \leq 1$.*

Proof. We write $\Phi = W_{\mathcal{L}(\Phi)} \circ \rho \circ \ldots \circ \rho \circ W_1$ and set $\widetilde{W}_\ell := (\mathcal{B}(\Phi))^{-1} W_\ell$, for $\ell \in \{1, \ldots, \mathcal{L}(\Phi)\}$, and $a := \mathcal{B}(\Phi)^{\mathcal{L}(\Phi)}$. Let $\Phi_a \in \mathcal{N}_{d',d'}$ be the multiplication network from Lemma A.1 and define

$$\widetilde{\Phi} := \widetilde{W}_{\mathcal{L}(\Phi)} \circ \rho \circ \cdots \circ \rho \circ \widetilde{W}_1,$$

and $\Psi := \Phi_a \circ \widetilde{\Phi}$ according to Lemma II.3. Note that $\widetilde{\Phi}$ has weights upper-bounded by 1 and is of the same depth and width as Φ. As ρ is positively homogeneous, i.e., $\rho(\lambda x) = \lambda \rho(x)$, for all $\lambda \geq 0$, $x \in \mathbb{R}$, we have $\Psi(x) = \Phi(x)$, for all $x \in \mathbb{R}^d$. Application of Lemma II.3 and Lemma A.1 completes the proof. □

Next we record a technical Lemma on how to realize a sum of networks with the same input by a network whose width is independent of the number of constituent networks.

Lemma A.4. *Let $d, d' \in \mathbb{N}$, $N \in \mathbb{N}$, and $\Phi_i \in \mathcal{N}_{d,d'}$, $i \in \{1, \dots, N\}$. There exists a network $\Phi \in \mathcal{N}_{d,d'}$ satisfying*

$$\Phi(x) = \sum_{i=1}^{N} \Phi_i(x), \quad \textit{for all } x \in \mathbb{R}^d,$$

with $\mathcal{L}(\Phi) = \sum_{i=1}^{N} \mathcal{L}(\Phi_i)$, $\mathcal{W}(\Phi) \leq 2d + 2d' + \max\{2d, \max_i\{\mathcal{W}(\Phi_i)\}\}$, $\mathcal{B}(\Phi) = \max\{1, \max_i \mathcal{B}(\Phi_i)\}$.

Proof. We set $L_i = \mathcal{L}(\Phi_i)$ and write the networks Φ_i as

$$\Phi_i = W^i_{L_i} \circ \rho \circ W^i_{L_i-1} \circ \rho \circ \cdots \circ \rho \circ W^i_1,$$

with $W^i_\ell(x) = A^i_\ell x + b^i_\ell$, where $A^i_\ell \in \mathbb{R}^{N^i_\ell \times N^i_{\ell-1}}$ and $b^i_\ell \in \mathbb{R}^{N^i_\ell}$. Next, using Lemma II.4, we turn the identity matrices $\mathbb{I}_d$ and $\mathbb{I}_{d'}$ into networks $\mathbb{I}^i_d$ and $\mathbb{I}^i_{d'}$, respectively, of depth L_i and then parallelize these networks, according to Lemma II.5, to get $\Psi_i := (\mathbb{I}^i_d, \mathbb{I}^i_{d'}, \Phi_i)$. Let $V^i_1(x) = E^i_1 x + f^i_1$ and $V^i_{L_i}(x) = E^i_{L_i} x + f^i_{L_i}$ denote the first and last, respectively, affine transformation of the network Ψ_i. By construction we have

$$E^i_1 = \begin{pmatrix} \mathbb{I}_d & 0 & 0 \\ -\mathbb{I}_d & 0 & 0 \\ 0 & \mathbb{I}_{d'} & 0 \\ 0 & -\mathbb{I}_{d'} & 0 \\ 0 & 0 & A^i_1 \end{pmatrix} \in \mathbb{R}^{(2d+2d'+N^i_1)\times(2d+d')}, \quad f^i_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ b^i_1 \end{pmatrix} \in \mathbb{R}^{2d+2d'+N^i_1}$$

and

$$E^i_{L_i} = \begin{pmatrix} \mathbb{I}_d & -\mathbb{I}_d & 0 & 0 & 0 \\ 0 & 0 & \mathbb{I}_{d'} & -\mathbb{I}_{d'} & 0 \\ 0 & 0 & 0 & 0 & A^i_{L_i} \end{pmatrix} \in \mathbb{R}^{(d+2d')\times(2d+2d'+N^i_{L_i-1})}, \quad f^i_{L_i} = \begin{pmatrix} 0 \\ 0 \\ b^i_{L_i} \end{pmatrix} \in \mathbb{R}^{d+2d'}.$$

Next, we define the matrices

$$A_{\text{in}} := \begin{pmatrix} \mathbb{I}_d \\ 0 \\ \mathbb{I}_d \end{pmatrix} \in \mathbb{R}^{(2d+d')\times d}, \quad A := \begin{pmatrix} \mathbb{I}_d & 0 & 0 \\ 0 & \mathbb{I}_{d'} & \mathbb{I}_{d'} \\ \mathbb{I}_d & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(2d+d')\times(d+2d')},$$

$$A_{\text{out}} := \begin{pmatrix} 0 & \mathbb{I}_{d'} & \mathbb{I}_{d'} \end{pmatrix} \in \mathbb{R}^{d'\times(d+2d')},$$

and note that $A_{\text{in}} x = (x, 0, x)$, $A(x, y, z)^T = (x, y + z, x)^T$, and $A_{\text{out}}(x, y, z)^T = y + z$, for $x \in \mathbb{R}^d, y, z \in \mathbb{R}^{d'}$. We construct

- the network $\widetilde{\Psi}_1$ by taking Ψ_1 and replacing E^1_1 with $E^1_1 A_{\text{in}}$, $E^1_{L_1}$ with $A E^1_{L_1}$, and $f^1_{L_1}$ with $A f^1_{L_1}$,
- the network $\widetilde{\Psi}_N$ by taking Ψ_N and replacing $E^N_{L_N}$ with $A_{\text{out}} E^N_{L_N}$ and $f^N_{L_N}$ with $A_{\text{out}} f^N_{L_N}$,
- the networks $\widetilde{\Psi}_i$, $i \in \{2, \dots, N-1\}$ by taking Ψ_i and replacing $E^i_{L_i}$ with $A E^i_{L_i}$ and $f^i_{L_i}$ with $A f^i_{L_i}$.

We can now verify that

$$\Phi = \widetilde{\Psi}_N \circ \widetilde{\Psi}_{N-1} \circ \cdots \circ \widetilde{\Psi}_1,$$

when the compositions are taken in the sense of Lemma II.3. Due to Lemmas II.4 and II.5, we have $\mathcal{L}(\Psi_i) = \mathcal{L}(\Phi_i)$, $\mathcal{W}(\Psi_i) = 2d + 2d' + \mathcal{W}(\Phi_i)$, and $\mathcal{B}(\Psi_i) = \max\{1, \mathcal{B}(\Phi_i)\}$. The proof is finalized by noting that, owing to the structure of the involved matrices, the depth and the weight magnitude remain unchanged by turning Ψ_i into $\widetilde{\Psi}_i$, whereas the width can not increase, but may decrease owing to the replacement of E_1^1 by $E_1^1 A_{\text{in}}$. □

The following lemma shows how to patch together local approximations using multiplication networks and a partition of unity consisting of hat functions. We note that this argument can be extended to higher dimensions using tensor products (which can be realized efficiently through multiplication networks) of the one-dimensional hat function.

Lemma A.5. *Let $\varepsilon \in (0, 1/2)$, $n \in \mathbb{N}$, $a_0 < a_1 < \cdots < a_n \in \mathbb{R}$, $f \in L^\infty([a_0, a_n])$, and*

$$A := \lceil \max\{|a_0|, |a_n|, 2 \max_{i \in \{2,\dots,n-1\}} \tfrac{1}{|a_i - a_{i-1}|}\} \rceil, \quad B := \max\{1, \|f\|_{L^\infty([a_0,a_n])}\}.$$

Assume that for every $i \in \{1, \dots, n-1\}$, there exists a network $\Phi_i \in \mathcal{N}_{1,1}$ with $\|f - \Phi_i\|_{L^\infty([a_{i-1}, a_{i+1}])} \leq \varepsilon/3$. Then, there is a network $\Phi \in \mathcal{N}_{1,1}$ satisfying

$$\|f - \Phi\|_{L^\infty([a_0,a_n])} \leq \varepsilon,$$

with $\mathcal{L}(\Phi) \leq \sum_{i=1}^{n-1} \mathcal{L}(\Phi_i) + Cn(\log(\varepsilon^{-1}) + \log(B) + \log(A))$, $\mathcal{W}(\Phi) \leq 7 + \max\{2, \max_{i \in \{1,\dots,n-1\}} \mathcal{W}(\Phi_i)\}$, $\mathcal{B}(\Phi) = \max\{1, \max_i \mathcal{B}(\Phi_i)\}$, and with $C > 0$ an absolute constant, i.e., independent of $\varepsilon, n, f, a_0, \dots, a_n$.

Proof. We first define the neural networks $(\Psi_i)_{i=1}^{n-1} \in \mathcal{N}_{1,1}$ forming a partition of unity according to

$$\begin{aligned}
\Psi_1(x) &:= 1 - \tfrac{1}{a_2 - a_1}\, \rho(x - a_1) + \tfrac{1}{a_2 - a_1}\, \rho(x - a_2), \\
\Psi_i(x) &:= \tfrac{1}{a_i - a_{i-1}}\, \rho(x - a_{i-1}) - (\tfrac{1}{a_i - a_{i-1}} + \tfrac{1}{a_{i+1} - a_i})\, \rho(x - a_i) + \tfrac{1}{a_{i+1} - a_i}\, \rho(x - a_{i+1}), \quad i \in \{2, \dots, n-2\}, \\
\Psi_{n-1}(x) &:= \tfrac{1}{a_{n-1} - a_{n-2}}\, \rho(x - a_{n-2}) - \tfrac{1}{a_{n-1} - a_{n-2}}\, \rho(x - a_{n-1}).
\end{aligned}$$

Note that $\operatorname{supp}(\Psi_1) = (\infty, a_2)$, $\operatorname{supp}(\Psi_{n-1}) = [a_{n-2}, \infty)$, and $\operatorname{supp}(\Psi_i) = [a_{i-1}, a_{i+1}]$. Proposition A.3 now ensures that, for all $i \in \{1, \dots, n-1\}$, Ψ_i can be realized as a network with $\mathcal{L}(\Psi_i) \leq 2(\lceil \log(A) \rceil + 5)$, $\mathcal{W}(\Psi_i) \leq 3$, and $\mathcal{B}(\Psi_i) \leq 1$. Next, let $\Phi_{B+1/6, \varepsilon/3} \in \mathcal{N}_{2,1}$ be the multiplication network according to Proposition III.3 and define the networks

$$\widetilde{\Phi}_i(x) := \Phi_{B+1/6, \varepsilon/3}(\Phi_i(x), \Psi_i(x))$$

according to Lemma II.5 and Lemma II.3, along with their sum

$$\Phi(x) := \sum_{i=1}^{n-1} \widetilde{\Phi}_i(x)$$

according to Lemma A.4. Proposition III.3 ensures, for all $i \in \{1, \dots, n-1\}$, $x \in [a_{i-1}, a_{i+1}]$, that

$$\begin{aligned}|f(x)\Psi_i(x) - \widetilde{\Phi}_i(x)| &\leq |f(x)\Psi_i(x) - \Phi_i(x)\Psi_i(x)| + |\Phi_i(x)\Psi_i(x) - \Phi_{B+1/6,\varepsilon/3}(\Phi_i(x), \Psi_i(x))| \\ &\leq (\Psi_i(x) + 1)\tfrac{\varepsilon}{3}\end{aligned}$$

and $\operatorname{supp}(\widetilde{\Phi}_i) = [a_{i-1}, a_{i+1}]$. In particular, for every $x \in [a_0, a_n]$, the set

$$I(x) := \{i \in \{1, \dots, n-1\} \colon \widetilde{\Phi}_i(x) \neq 0\}$$

of active indices contains at most two elements. Moreover, we have $\sum_{i \in I(x)} \Psi_i(x) = 1$ by construction, which implies that, for all $x \in \mathbb{R}$,

$$|f(x) - \Phi(x)| = \left| \sum_{i \in I(x)} \Psi_i(x) f(x) - \sum_{i \in I(x)} \tilde{\Phi}_i(x) \right| \leq \sum_{i \in I(x)} (\Psi_i(x) + 1)\tfrac{\varepsilon}{3} \leq \varepsilon.$$

Due to Lemma II.3, Lemma II.5, Proposition III.3, and Lemma A.4, we can conclude that Φ, indeed, satisfies the claimed properties. □

Next, we present an extension of Lemma III.7 to arbitrary (finite) intervals.

Lemma A.6. *For $a, b \in \mathbb{R}$ with $a < b$, let*

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

There exists a constant $C > 0$ such that for all $a, b \in \mathbb{R}$ with $a < b$, $f \in \mathcal{S}_{[a,b]}$, and $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{f,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

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

with $\mathcal{L}(\Psi_{f,\varepsilon}) \leq C \max\{2, (b-a)\}((\log(\varepsilon^{-1}))^2 + \log(\lceil \max\{|a|, |b|\} \rceil) + \log(\lceil \frac{1}{b-a} \rceil))$, $\mathcal{W}(\Psi_{f,\varepsilon}) \leq 16$, $\mathcal{B}(\Psi_{f,\varepsilon}) \leq 1$.

Proof. We first recall that the case $[a, b] = [-1, 1]$ has already been dealt with in Lemma III.7. Here, we will first prove the statement for the interval $[-D, D]$ with $D \in (0, 1)$ and then use this result to establish the general case through a patching argument according to Lemma A.5. We start by noting that for $g \in \mathcal{S}_{[-D,D]}$, the function $f_g \colon [-1, 1] \to \mathbb{R}, x \mapsto g(Dx)$ is in $\mathcal{S}_{[-1,1]}$ due to $D < 1$. Hence, by Lemma III.7, there exists a constant $C > 0$ such that for all $g \in \mathcal{S}_{[-D,D]}$ and $\varepsilon \in (0, 1/2)$, there is a network $\widetilde{\Psi}_{g,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying $\|\widetilde{\Psi}_{g,\varepsilon} - f_g\|_{L^\infty([-1,1])} \leq \varepsilon$, with $\mathcal{L}(\widetilde{\Psi}_{g,\varepsilon}) \leq C(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\widetilde{\Psi}_{g,\varepsilon}) \leq 9$, $\mathcal{B}(\widetilde{\Psi}_{g,\varepsilon}) \leq 1$. The claim is then established by taking the network approximating g to be $\Psi_{g,\varepsilon} := \widetilde{\Psi}_{g,\varepsilon} \circ \Phi_{D^{-1}}$, where $\Phi_{D^{-1}}$ is the scalar multiplication network from Lemma A.1, and noting that

$$\begin{aligned}\|\Psi_{g,\varepsilon}(x) - g(x)\|_{L^\infty([-D,D])} &= \sup_{x \in [-D,D]} |\widetilde{\Psi}_{g,\varepsilon}(\tfrac{x}{D}) - f_g(\tfrac{x}{D})| \\ &= \sup_{x \in [-1,1]} |\widetilde{\Psi}_{g,\varepsilon}(x) - f_g(x)| \leq \varepsilon.\end{aligned}$$

Due to Lemma II.3, we have $\mathcal{L}(\Psi_{g,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil \frac{1}{D} \rceil))$, $\mathcal{W}(\Psi_{g,\varepsilon}) \leq 9$, and $\mathcal{B}(\Psi_{g,\varepsilon}) \leq 1$. We are now ready to proceed to the proof of the statement for general intervals $[a,b]$. This will be accomplished by approximating f on intervals of length no more than 2 and stitching the resulting approximations together according to Lemma A.5. We start with the case $b-a \leq 2$ and note that here we can simply shift the function by $(a+b)/2$ to center its domain around the origin and then use the result above for approximation on $[-D,D]$ with $D \in (0,1)$ or Lemma III.7 if $b-a=2$, both in combination with Corollary A.2 to realize the shift through a neural network with weights bounded by 1. Using Lemma II.3 to implement the composition of the network realizing this shift with that realizing g, we can conclude the existence of a constant $C' > 0$ such that, for all $[a,b] \subseteq \mathbb{R}$ with $b-a \leq 2$, $g \in \mathcal{S}_{[a,b]}$, $\varepsilon \in (0,1/2)$, there is a network satisfying $\|g - \Psi_{g,\varepsilon}\|_{L^\infty([a,b])} \leq \varepsilon$ with $\mathcal{L}(\Psi_{g,\varepsilon}) \leq C'((\log(\varepsilon^{-1}))^2 + \log(\lceil \frac{1}{b-a} \rceil))$, $\mathcal{W}(\Psi_{g,\varepsilon}) \leq 9$, and $\mathcal{B}(\Psi_{g,\varepsilon}) \leq 1$. Finally, for $b-a > 2$, we partition the interval $[a,b]$ and apply Lemma A.5 as follows. We set $n := \lceil b-a \rceil$ and define

$$a_i := a + i\tfrac{b-a}{n}, \quad i \in \{0,\dots,n\}.$$

Next, for $i \in \{1,\dots,n-1\}$, let $g_i \colon [a_{i-1}, a_{i+1}] \to \mathbb{R}$ be the restriction of g to the interval $[a_{i-1}, a_{i+1}]$, and note that $a_{i+1} - a_{i-1} = \frac{2(b-a)}{n} \in (\frac{4}{3}, 2]$. Furthermore, for $i \in \{1,\dots,n-1\}$, let $\Psi_{g_i,\varepsilon/3}$ be the network approximating g_i with error $\varepsilon/3$ as constructed above. Then, for every $i \in \{1,\dots,n-1\}$, it holds that $\|g - \Psi_{g_i,\varepsilon/3}\|_{L^\infty([a_{i-1},a_{i+1}])} \leq \frac{\varepsilon}{3}$ and application of Lemma A.5 yields the desired result. □

We finally record, for technical purposes, slight variations of Lemmas II.5 and II.6 to account for parallelizations and linear combinations, respectively, of neural networks with shared input.

Lemma A.7. *Let $n, d, L \in \mathbb{N}$ and, for $i \in \{1,2,\dots,n\}$, let $d_i' \in \mathbb{N}$ and $\Phi_i \in \mathcal{N}_{d,d_i'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{d,\sum_{i=1}^n d_i'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) = \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) \leq \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i \mathcal{B}(\Phi_i)$, and satisfying*

$$\Psi(x) = (\Phi_1(x), \Phi_2(x), \dots, \Phi_n(x)) \in \mathbb{R}^{\sum_{i=1}^n d_i'},$$

for $x \in \mathbb{R}^d$.

Proof. The claim is established by following the construction in the proof of Lemma II.5, but with the matrix $A_1 = \operatorname{diag}(A_1^1, A_1^2, \dots, A_1^n)$ replaced by

$$A_1 = \begin{pmatrix} A_1^1 \\ \vdots \\ A_1^n \end{pmatrix} \in \mathbb{R}^{(\sum_{i=1}^n N_1^i) \times d},$$

where N_1^i is the dimension of the first layer of Φ_i. □

Lemma A.8. *Let $n, d, d', L \in \mathbb{N}$ and, for $i \in \{1, 2, \dots, n\}$, let $a_i \in \mathbb{R}$ and $\Phi_i \in \mathcal{N}_{d,d'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{d,d'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) \leq \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) \leq \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i\{|a_i|\mathcal{B}(\Phi_i)\}$, and satisfying*

$$\Psi(x) = \sum_{i=1}^n a_i \Phi_i(x) \in \mathbb{R}^{d'},$$

for $x \in \mathbb{R}^d$.

Proof. The proof follows directly from that of Lemma A.7 with the same modifications as those needed in the proof of Lemma II.6 relative to that of Lemma II.5. □

APPENDIX B
TAIL COMPACTNESS FOR BESOV SPACES

We consider the Besov space $B^m_{p,q}([0,1])$ [16] given by the set of functions $f \in L^2([0,1])$ satisfying

$$\|f\|_{m,p,q} := \|(2^{n(m+\frac{1}{2}-\frac{1}{p})}\|(\langle f, \psi_{n,k}\rangle)_{k=0}^{2^n-1}\|_{\ell^p})_{n\in\mathbb{N}_0}\|_{\ell^q} < \infty, \tag{92}$$

with $\mathcal{D} = \{\psi_{n,k} \colon n \in \mathbb{N}_0, k = 0, \dots, 2^n - 1\}$ an orthonormal wavelet basis[12] for $L^2([0,1])$ and ℓ^p denoting the usual sequence norm

$$\|(a_i)_{i\in I}\|_{\ell^p} = \begin{cases} \left(\sum_{i\in I} |a_i|^p\right)^{\frac{1}{p}}, & 1 \leq p < \infty \\ \sup_{i\in I} |a_i|, & p = \infty \end{cases}.$$

The unit ball in $B^m_{p,q}([0,1])$ is

$$\mathcal{U}(B^m_{p,q}([0,1])) = \{f \in L^2([0,1]) \colon \|f\|_{m,p,q} \leq 1\}. \tag{93}$$

For simplicity of notation, we set $a_{n,k}(f) := \langle f, \psi_{n,k}\rangle$ and $A_n(f) := (a_{n,k}(f))_{k=0}^{2^n-1} \in \mathbb{R}^{2^n}$, for $n \in \mathbb{N}_0$. We now want to verify that for $q \in [1,2]$ tail compactness holds for the pair $(\mathcal{U}(B^m_{p,q}([0,1])), \mathcal{D})$ under the ordering $\mathcal{D} = (\mathcal{D}_0, \mathcal{D}_1, \dots)$, where $\mathcal{D}_n := \{\psi_{n,k} \colon k = 0, \dots, 2^n - 1\}$. To this end, we first note that owing to $\sum_{n=0}^N |\mathcal{D}_n| = 2^{N+1} - 1$, we have tail compactness according to (26) if there exist $C, \beta > 0$ such that for all $f \in \mathcal{U}(B^m_{p,q}([0,1]))$, $N \in \mathbb{N}$,

$$\left\| f - \sum_{n=0}^N \sum_{k=0}^{2^n-1} a_{n,k}(f)\psi_{n,k} \right\|_{L^2([0,1])} \leq C(2^{N+1})^{-\beta}. \tag{94}$$

To see that (92) implies (94), we note that by orthonormality of $\mathcal{D}$,

$$\left\| f - \sum_{n=0}^N \sum_{k=0}^{2^n-1} a_{n,k}(f)\psi_{n,k} \right\|_{L^2([0,1])} = \left\| \sum_{n=N+1}^\infty \sum_{k=0}^{2^n-1} a_{n,k}(f)\psi_{n,k} \right\|_{L^2([0,1])} = \left(\sum_{n=N+1}^\infty \sum_{k=0}^{2^n-1} |a_{n,k}(f)|^2 \right)^{\frac{1}{2}}$$
$$= \|(\|A_n(f)\|_{\ell^2})_{n=N+1}^\infty\|_{\ell^2}.$$

[12]The space does not depend on the particular choice of mother wavelet ψ as long as ψ has at least r vanishing moments and is in $C^r([0,1])$ for some $r > m$. For further details we refer to Section 9.2.3 in [16].

As the $A_n(f)$ are finite sequences of length $|\mathcal{D}_n| = 2^n$, it follows, by application of Hölder's inequality, that $\|A_n(f)\|_{\ell^2} \leq 2^{n(\frac{1}{2}-\frac{1}{p})}\|A_n(f)\|_{\ell^p}$. Together with $\|\cdot\|_{\ell^2} \leq \|\cdot\|_{\ell^q}$, for $q \leq 2$, (92) then ensures, for all $f \in \mathcal{U}(B^m_{p,q}([0,1]))$ and $q \in [1,2]$, that

$$\begin{aligned}\|(\|A_n(f)\|_{\ell^2})_{n=N+1}^{\infty}\|_{\ell^2} &\leq \|(2^{n(\frac{1}{2}-\frac{1}{p})}\|A_n(f)\|_{\ell^p})_{n=N+1}^{\infty}\|_{\ell^q} \leq 2^{-(N+1)m}\|(2^{n(m+\frac{1}{2}-\frac{1}{p})}\|A_n(f)\|_{\ell^p})_{n=N+1}^{\infty}\|_{\ell^q}\\ &\leq 2^{-(N+1)m}\|f\|_{m,p,q} \leq (2^{N+1})^{-m},\end{aligned}$$

which establishes (94) with $C = 1$ and $\beta = m$.

APPENDIX C
TAIL COMPACTNESS FOR MODULATION SPACES

We consider tail compactness for unit balls in (polynomially) weighted modulation spaces, which, for $p, q \in [1,\infty)$, are defined as follows

$$M^s_{p,q}(\mathbb{R}) := \{f\colon \|f\|_{M^s_{p,q}(\mathbb{R})} < \infty\},$$

with

$$\|f\|_{M^s_{p,q}(\mathbb{R})} := \left(\int_{\mathbb{R}}\left(\int_{\mathbb{R}} |V_w f(x,\xi)|^p (1+|x|+|\xi|)^{sp}\mathrm{d}x\right)^{\frac{q}{p}}\mathrm{d}\xi\right)^{\frac{1}{q}},$$

where

$$V_w f(x,\xi) := \int_{\mathbb{R}} f(t)\,\overline{w(t-x)}e^{-2\pi i t\xi}\mathrm{d}t, \quad x,\xi \in \mathbb{R},$$

is the short-time Fourier transform of f with respect to the window function[13] $w \in \mathcal{S}(\mathbb{R})$.

Next, let $g \in L^2(\mathbb{R})$ with $\|g\|_{L^2(\mathbb{R})} = 1$ and $g(x) = \overline{g(-x)}$ such that the Gabor dictionary $\mathcal{G}(g, \frac{1}{2}, 1, \mathbb{R})$ is a tight frame [68] for $L^2(\mathbb{R})$. Then, the Wilson dictionary $\mathcal{D} = \{\psi_{k,n}\colon (k,n) \in \mathbb{Z}\times\mathbb{N}_0\}$ with

$$\begin{aligned}\psi_{k,0} &= T_k g, && k \in \mathbb{Z},\\ \psi_{k,n} &= \tfrac{1}{\sqrt{2}}T_{\frac{k}{2}}(M_n + (-1)^{k+n}M_{-n})g, && (k,n) \in \mathbb{Z}\times\mathbb{N},\end{aligned}$$

is an orthonormal basis for $L^2(\mathbb{R})$ (see [17, Thm. 8.5.1]). We have, for every $f \in M^s_{p,q}(\mathbb{R})$, the expansion [17, Thm. 12.3.4]

$$f = \sum_{(k,n)\in\mathbb{Z}\times\mathbb{N}_0} c_{k,n}(f)\psi_{k,n}, \quad \text{where} \quad c_{k,n}(f) = \langle f, \psi_{k,n}\rangle, \quad c(f) \in \ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0),$$

[13]The resulting modulation space does not depend on the specific choice of window function w as long as w is in the Schwartz space $\mathcal{S}(\mathbb{R}) = \{f \in C^\infty(\mathbb{R})\colon \sup_{x\in\mathbb{R}}|x^\alpha f^{(\beta)}(x)| < \infty, \text{ for all } \alpha,\beta \in \mathbb{N}_0\}$, where $f^{(n)}$ stands for the n-th derivative of f.

with $\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)$ the space of sequences $c\in\mathbb{R}^{\mathbb{Z}\times\mathbb{N}_0}$ satisfying

$$\|c\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)} := \left(\sum_{n\in\mathbb{N}_0}\left(\sum_{k\in\mathbb{Z}}|c_{k,n}|^p(1+|\tfrac{k}{2}|+|n|)^{sp}\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}<\infty.$$

Moreover, there exists [17, Thm. 12.3.1] a constant $D\geq 1$ such that, for all $f\in M^s_{p,q}(\mathbb{R})$,

$$\tfrac{1}{D}\|f\|_{M^s_{p,q}(\mathbb{R})}\leq\|c(f)\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)}\leq D\|f\|_{M^s_{p,q}(\mathbb{R})}.$$

In particular, we can characterize the unit ball of $M^s_{p,q}(\mathbb{R})$ according to

$$\mathcal{U}(M^s_{p,q}(\mathbb{R}))=\{f\colon\|c(f)\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)}\leq D\}.$$

We now order the Wilson basis dictionary as follows. Define $\mathcal{D}_0:=\{\psi_{0,0}\}$ and

$$\mathcal{D}_\ell:=\{\psi_{k,n}\colon|k|,n\leq\ell\}\setminus\bigcup_{i=0}^{\ell-1}\mathcal{D}_i$$

for $\ell\geq 1$, and order the overall dictionary according to $\mathcal{D}=(\mathcal{D}_0,\mathcal{D}_1,\dots)$. Owing to $\sum_{\ell=0}^N|\mathcal{D}_\ell|=(2N+1)(N+1)$, we have tail compactness for the pair $(\mathcal{U}(M^s_{p,q}(\mathbb{R})),\mathcal{D})$ if there exist $C,\beta>0$ such that, for all $f\in\mathcal{U}(M^s_{p,q}(\mathbb{R}))$, $N\in\mathbb{N}$,

$$\left\|f-\sum_{n=0}^N\sum_{k=-N}^N c_{k,n}(f)\psi_{k,n}\right\|_{L^2(\mathbb{R})}\leq CN^{-\beta}. \tag{95}$$

We restrict our attention to $p,q\leq 2$ and use orthonormality of $\mathcal{D}$ and the fact that $\|\cdot\|_{\ell^2}\leq\|\cdot\|_{\ell^p}$, for $p\leq 2$, to obtain, for all $f\in\mathcal{U}(M^s_{p,q}(\mathbb{R}))$,

$$\begin{aligned}\left\|f-\sum_{n=0}^N\sum_{k=-N}^N c_{k,n}(f)\psi_{k,n}\right\|_{L^2(\mathbb{R})}&=\left\|\sum_{n>N}\sum_{|k|>N}c_{k,n}(f)\psi_{k,n}\right\|_{L^2(\mathbb{R})}=\left(\sum_{n>N}\sum_{|k|>N}|c_{k,n}(f)|^2\right)^{\frac{1}{2}}\\&\leq\left(\sum_{n>N}\left(\sum_{|k|>N}|c_{k,n}(f)|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}\\&\leq(1+\tfrac{3}{2}N)^{-s}\left(\sum_{n>N}\left(\sum_{|k|>N}|c_{k,n}(f)|^p(1+|\tfrac{k}{2}|+|n|)^{sp}\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}\\&\leq(1+\tfrac{3}{2}N)^{-s}\|c(f)\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)}\leq(3/2)^{-s}DN^{-s},\end{aligned}$$

which establishes tail compactness with $C=(3/2)^{-s}D$ and $\beta=s$.

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] W. McCulloch and W. Pitts, "A logical calculus of ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 1943.

[10] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," *Dokl. Akad. Nauk SSSR*, vol. 114, no. 5, pp. 953–956, 1957.

[11] 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

[12] 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

[13] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 8–45, 2019.

[14] 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

[15] ——, "Unconditional bases and bit-level compression," *Appl. Comput. Harm. Anal.*, vol. 3, pp. 388–392, 1996.

[16] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. USA: Academic Press, Inc., 2008.

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

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

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

[20] D. M. Elbrächter, J. Berner, and P. Grohs, "How degenerate is the parametrization of neural networks with the ReLU activation function?" in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, p. 7788–7799. [Online]. Available: https://arxiv.org/abs/1905.09803

[21] V. Vlačić and H. Bölcskei, "Neural network identifiability for a family of sigmoidal nonlinearities," *Constructive Approximation*, 2021. [Online]. Available: https://arxiv.org/abs/1906.06994

[22] ——, "Affine symmetries and neural network identifiability," *Advances in Mathematics*, vol. 376, no. 107485, pp. 1–72, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001870820305132

[23] 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.

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

[25] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020. [Online]. Available: https://arxiv.org/abs/1708.06633

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

[27] B. Hanin and D. Rolnick, "Deep ReLU networks have surprisingly few activation patterns," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 361–370. [Online]. Available: http://papers.nips.cc/paper/8328-deep-relu-networks-have-surprisingly-few-activation-patterns.pdf

[28] D. Fokina and I. Oseledets, "Growing axons: Greedy learning of neural networks with application to function approximation," 2019. [Online]. Available: https://arxiv.org/abs/1910.12686

[29] C. Schwab and J. Zech, "Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ," *Analysis and Applications*, vol. 17, no. 1, pp. 19–55, 2019.

[30] J. A. A. Opschoor, P. C. Petersen, and C. Schwab, "Deep ReLU networks and high-order finite element methods," *Analysis and Applications*, vol. 18, no. 5, pp. 715–770, 2020. [Online]. Available: https://doi.org/10.1142/S0219530519410136

[31] I. Gühring, G. Kutyniok, and P. Petersen, "Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms," *Analysis and Applications*, vol. 18, no. 5, pp. 803–859, 2020. [Online]. Available: https://doi.org/10.1142/S0219530519410021

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

[33] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" *International Conference on Learning Representations*, 2017. [Online]. Available: https://arxiv.org/abs/1610.04161

[34] A. Gil, J. Segura, and N. M. Temme, *Numerical Methods for Special Functions*. Society for Industrial and Applied Mathematics, 2007. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9780898717822

[35] 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.

[36] ——, "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

[37] 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

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

[39] E. J. Candès, "Ridgelets: Theory and applications," Ph.D. dissertation, Stanford University, 1998.

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

[41] H. N. Mhaskar and C. A. 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.

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

[43] 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

[44] 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

[45] 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

[46] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proceedings of the 29th Conference on Learning Theory*, 2016, pp. 907–940.

[47] 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

[48] 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.

[49] 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.

[50] 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.

[51] 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," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 3, pp. 631–657, 2020.

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

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

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

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

[56] R. DeVore, B. Hanin, and G. Petrova, "Neural network approximation," *arXiv:2012.14501*, 2020.

[57] 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

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

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

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

[61] E. Ott, *Chaos in Dynamical Systems*. Cambridge University Press, 2002.

[62] M. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

[63] R. T. Prosser, "The ε-entropy and ε-capacity of certain time-varying channels," *Journal of Mathematical Analysis and Applications*, vol. 16, pp. 553–573, 1966.

[64] A. Kolmogorov and V. Tikhomirov, "ε-entropy and ε-capacity of sets in function spaces," *Uspekhi Mat. Nauk.*, vol. 14, no. 2, pp. 3–86, 1959.

[65] M. Ehler and F. Filbir, "Metric entropy, n-widths, and sampling of functions on manifolds," *Journal of Approximation Theory*, vol. 225, pp. 41 – 57, 2018.

[66] J. Schmidt-Hieber, "Deep ReLU network approximation of functions on a manifold," *arXiv:1908.00695*, 2019.

[67] H. Mhaskar, "A direct approach for function approximation on data defined manifolds," *Neural Networks*, vol. 132, pp. 253 – 268, 2020.

[68] V. Morgenshtern and H. Bölcskei, *Mathematical Foundations for Signal Processing, Communications, and Networking*, Boca Raton, FL, 2012, ch. A short course on frame theory, pp. 737–789.

[69] 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

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

[71] 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.

[72] 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.

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

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

[75] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2435–2476, 1998.

[76] P. Grohs, A. Klotz, and F. Voigtlaender, "Phase transitions in rate distortion theory and deep learning," *arxiv:2008.01011*, 2020.

[77] A. Hinrichs, I. Piotrowska-Kurczewski, and M. Piotrowski, "On the degree of compactness of embeddings between weighted modulation spaces," *J. Funct. Spaces Appl.*, vol. 6, pp. 303–317, 01 2008.

[78] 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

[79] 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

[80] J. Munkres, *Topology*, ser. Featured Titles for Topology. Prentice Hall, Incorporated, 2000.

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

[82] S. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$," *Trans. Amer. Math. Soc.*, vol. 315, no. 1, pp. 69–87, Sep. 1989.

[83] C. K. Chui and J.-Z. Wang, "On compactly supported spline wavelets and a duality principle," *Trans. Amer. Math. Soc.*, vol. 330, no. 2, pp. 903–915, Apr. 1992.

[84] 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

[85] H. G. Feichtinger, "On a new Segal algebra," *Monatshefte für Mathematik*, vol. 92, pp. 269–289, 1981.

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

[87] 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.