OPTIMAL APPROXIMATION WITH SPARSELY CONNECTED DEEP NEURAL NETWORKS

HELMUT BÖLCSKEI*, PHILIPP GROHS†, GITTA KUTYNIOK‡, AND PHILIPP PETERSEN§

Abstract. We derive fundamental lower bounds on the connectivity and the memory requirements of deep neural networks guaranteeing uniform approximation rates for arbitrary function classes in $L^2(\mathbb{R}^d)$. In other words, we establish a connection between the complexity of a function class and the complexity of deep neural networks approximating functions from this class to within a prescribed accuracy. Additionally, we prove that our lower bounds are achievable for a broad family of function classes. Specifically, all function classes that are optimally approximated by a general class of representation systems—so-called *affine systems*—can be approximated by deep neural networks with minimal connectivity and memory requirements. Affine systems encompass a wealth of representation systems from applied harmonic analysis such as wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules. Our central result elucidates a remarkable universality property of neural networks and shows that they achieve the optimum approximation properties of all affine systems combined. As a specific example, we consider the class of α^{-1}-cartoon-like functions, which is approximated optimally by α-shearlets. We also explain how our results can be extended to the approximation of functions on low-dimensional immersed manifolds. Finally, we present numerical experiments demonstrating that the standard stochastic gradient descent algorithm yields deep neural networks with close-to-optimal approximation rates. Moreover, these results indicate that stochastic gradient descent can learn approximations that are sparse in the representation systems optimally sparsifying the function class the network is trained on.

Keywords. Neural networks, function approximation, optimal sparse approximation, sparse connectivity, wavelets, shearlets

AMS subject classification. 41A25, 82C32, 42C40, 42C15, 41A46, 68T05, 94A34, 94A12

1. Introduction.

Neural networks arose from the seminal work by McCulloch and Pitts [41] in 1943 which, inspired by the functionality of the human brain, introduced an algorithmic approach to learning with the aim of building a theory of artificial intelligence. Roughly speaking, a neural network consists of neurons arranged in layers and connected by weighted edges; in mathematical terms this boils down to a concatenation of affine linear functions and relatively simple non-linearities.

Despite significant theoretical progress in the 1990s [12, 34], the area has seen practical progress only during the past decade, triggered by the drastic improvements in computing power and the availability of vast amounts of training data. Deep neural networks, i.e., networks with large numbers of layers, are now state-of-the-art technology for a wide variety of applications, such as image classification [36], speech recognition [33], or game intelligence [13]. For an in-depth overview, we refer to the survey paper by LeCun, Bengio, and Hinton [39] and the recent book [24].

A neural network effectively implements a non-linear mapping and can be used to either perform classification directly or to extract features that are then fed into a classifier, such as a support vector machine [54]. In the former case, the primary goal is to approximate an unknown classification function based on a given set of input-output value pairs. This is typically accomplished by learning the network's weights through, e.g., the stochastic gradient descent (via backpropagation) algorithm [52]. In a classification task with, say, two classes,

*Department of Information Technology and Electrical Engineering and Department of Mathematics, ETH Zürich, 8092 Zürich, Switzerland. Email-Address: hboelcskei@ethz.ch

†Faculty of Mathematics, University of Vienna, 1090 Vienna, Austria, and Research Platform DataScience@UniVienna, University of Vienna, 1090 Vienna, Austria. Email-Address: philipp.grohs@univie.ac.at

‡Institut für Mathematik and Department of Electrical Engineering & Computer Science, Technische Universität Berlin, 10623 Berlin, Germany. Email-Address: kutyniok@math.tu-berlin.de

§Mathematical Institute, University of Oxford, OX26GG Oxford, United Kingdom. Email-Address: philipp.petersen@maths.ox.ac.uk

the function to be learned would take only two values, whereas in the case of, e.g., the prediction of the temperature in a certain environment, it would be real-valued. It is therefore clear that characterizing to what extent (deep) neural networks are capable of approximating general functions is a question of significant practical relevance.

Neural networks employed in practice often consist of hundreds of layers and may depend on billions of parameters, see for example the work [32] on image classification. Training and operation of networks of this scale entail formidable computational challenges. As a case in point, we mention speech recognition on a smartphone such as, e.g., Apple's SIRI-system, which operates in the cloud. Android's speech recognition system has meanwhile released an offline version based on a neural network with sparse connectivity. The desire to reduce the complexity of network training and operation naturally leads to the question of the fundamental limits on function approximation through neural networks with sparse connectivity. In addition, the network's memory requirements in terms of the number of bits needed to store its topology and weights are of concern in practice.

The purpose of this paper is to understand the connectivity and memory requirements of (deep) neural networks induced by demands on their approximation-theoretic properties. Specifically, defining the complexity of a function class $\mathcal{C}$ as the rate of growth of the minimum number of bits needed to describe any element in $\mathcal{C}$ to within a maximum allowed error approaching zero, we shall be interested in the following question: Depending on the complexity of $\mathcal{C}$, what are the connectivity and memory requirements of a deep neural network approximating every element in $\mathcal{C}$ to within an error of ε? We address this question by interpreting the network as an encoder in Donoho's min-max rate distortion theory [19] and establishing rate-distortion optimality for a broad family of function classes $\mathcal{C}$, namely those classes for which so-called affine systems—a general class of representation systems—yield optimal approximation rates in the sense of non-linear approximation theory [16]. Affine systems encompass a wealth of representation systems from applied harmonic analysis such as wavelets [14], ridgelets [4], curvelets [6], shearlets [31], α-shearlets and more generally α-molecules [27]. Our result therefore uncovers an interesting universality property of deep neural networks; they exhibit the optimal approximation properties of all affine systems combined. The technique we develop to prove our main statements is interesting in its own right as it constitutes a more general framework for transferring results on function approximation through representation systems to results on approximation by deep neural networks.

1.1. Deep Neural Networks. While various network architectures exist in the literature, we focus on the following setup.

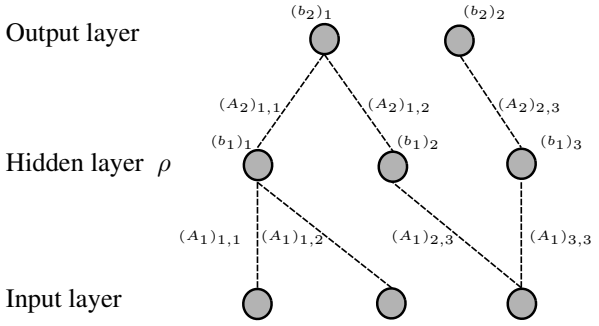
DEFINITION 1.1. *Let $L, d, N_1, \dots, N_L \in \mathbb{N}$ with $L \geq 2$. A map $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by*

$$\Phi(x) = W_L \rho \left(W_{L-1} \rho \left(\dots \rho \left(W_1(x) \right) \right) \right), \quad \text{for } x \in \mathbb{R}^d, \tag{1.1}$$

with affine linear maps $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $1 \leq \ell \leq L$, and the non-linear activation function *ρ acting component-wise, is called a* neural network. *Here, $N_0 := d$ is the* dimension of the 0-th layer referred to as the input layer*, L denotes the* number of layers *(not counting the input layer), $N_1, \dots, N_{L-1}$ stands for the* dimensions of the $L-1$ hidden layers*, and N_L is the* dimension of the output layer. *The affine linear map W_ℓ is defined via $W_\ell(x) = A_\ell x + b_\ell$ with $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and the affine part $b_\ell \in \mathbb{R}^{N_\ell}$. $(A_\ell)_{i,j}$ 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*, while $(b_\ell)_i$ is* the weight associated with the i-th node in the ℓ-th layer. *These assignments are schematized in Figure 1. The total number of nodes is given by $\mathcal{N}(\Phi) := d + \sum_{\ell=1}^{L} N_\ell$. The real numbers $(A_\ell)_{i,j}$ and $(b_\ell)_i$ are said to be the network's edge weights*

and node weights, respectively, and the total number of nonzero edge weights, denoted by $\mathcal{M}(\Phi)$, is the network's connectivity.

The term "network" stems from the interpretation of the mapping Φ as a weighted acyclic directed graph with nodes arranged in $L+1$ hierarchical layers and edges only between adjacent layers. If the network's connectivity $\mathcal{M}(\Phi)$ is small relative to the number of connections possible (i.e., the number of edges in the graph that is fully connected between adjacent layers), we say that the network is *sparsely connected*.



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

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

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.*

Throughout the paper, we consider the case $\Phi : \mathbb{R}^d \to \mathbb{R}$, i.e., $N_L = 1$, which includes situations such as the classification and temperature prediction problem described above. We emphasize, however, that the general results of Sections 3, 4, and 5 are readily generalized to $N_L > 1$.

We denote the class of networks $\Phi : \mathbb{R}^d \to \mathbb{R}$ with exactly L layers, connectivity no more than M, and activation function ρ by $\mathcal{NN}_{L,M,d,\rho}$ with the understanding that for $L = 1$, the set $\mathcal{NN}_{L,M,d,\rho}$ is empty. Moreover, we let $\mathcal{NN}_{\infty,M,d,\rho} := \bigcup_{L\in\mathbb{N}} \mathcal{NN}_{L,M,d,\rho}$, $\mathcal{NN}_{L,\infty,d,\rho} := \bigcup_{M\in\mathbb{N}} \mathcal{NN}_{L,M,d,\rho}$, and $\mathcal{NN}_{\infty,\infty,d,\rho} := \bigcup_{L\in\mathbb{N}} \mathcal{NN}_{L,\infty,d,\rho}$.

Now, given a function $f : \mathbb{R}^d \to \mathbb{R}$, we are interested in the theoretically best possible approximation of f by a network $\Phi \in \mathcal{NN}_{\infty,M,d,\rho}$. Specifically, we will want to know how the approximation quality depends on the connectivity M and what the associated number of bits needed to store the network topology and the corresponding quantized weights is. Clearly, smaller M entails lower computational complexity in terms of evaluating (1.1) and a smaller number of bits translates to reduced memory requirements for storing the network. Such a result benchmarks all conceivable algorithms for learning the network topology and weights.

1.2. Quantifying Approximation Quality. We proceed to formalizing our problem statement and start with a brief review of a widely used framework in approximation theory [17, 16].

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

DEFINITION 1.2 ([17]). *Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a*

representation system $\mathcal{D} = (\varphi_i)_{i\in I} \subset L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,

$$\Gamma_M^{\mathcal{D}}(f) := \inf_{\substack{I_M \subseteq I,\\ \#I_M = M, (c_i)_{i\in I_M}}} \left\| f - \sum_{i\in I_M} c_i\varphi_i \right\|_{L^2(\Omega)}. \tag{1.2}$$

We call $\Gamma_M^{\mathcal{D}}(f)$ the best M-term approximation error of f in $\mathcal{D}$. *Every $f_M = \sum_{i\in I_M} c_i\varphi_i$ attaining the infimum in (1.2) 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 $\gamma^*(\mathcal{C},\mathcal{D})$. Here, $\mathcal{O}(g(\cdot))$ denotes the class of functions bounded asymptotically by g in the sense of standard Landau notation. We say that the best M-term approximation rate of $\mathcal{C}$ in the representation system $\mathcal{D}$ *is $\gamma^*(\mathcal{C},\mathcal{D})$.*

Function classes $\mathcal{C}$ widely studied in the approximation theory literature include unit balls in Lebesgue, Sobolev, or Besov spaces [16], as well as α-cartoon-like functions [27]. A wealth of structured representation systems $\mathcal{D}$ is provided by the area of applied harmonic analysis, starting with wavelets [14], followed by ridgelets [4], curvelets [6], shearlets [31], parabolic molecules [29], and most generally α-molecules [27], which include all previously named systems as special cases. Further examples are Gabor frames [25] and wave atoms [15].

1.3. Approximation by Deep Neural Networks. The main conceptual contribution of this paper is the development of an approximation-theoretic framework for deep neural networks in the spirit of [17]. Specifically, we shall substitute the concept of best M-term approximation with representation systems by best M-edge approximation through neural networks. In other words, parsimony in terms of the number of participating elements of a representation system is replaced by parsimony in terms of connectivity. More formally, we consider the following setup.

DEFINITION 1.3. *Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and an activation function $\rho : \mathbb{R} \to \mathbb{R}$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma_M^{\mathcal{NN}}(f) := \inf_{\Phi\in\mathcal{NN}_{\infty,M,d,\rho}} \|f - \Phi\|_{L^2(\Omega)}. \tag{1.3}$$

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

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

will be denoted by $\gamma^*_{\mathcal{NN}}(\mathcal{C},\rho)$. We say that the best M-edge approximation rate of $\mathcal{C}$ by neural networks with activation function ρ *is $\gamma^*_{\mathcal{NN}}(\mathcal{C},\rho)$.*

We emphasize that the infimum in (1.3) is taken over all networks with fixed activation function ρ, fixed input dimension d, no more than M edges of nonzero weight, and arbitrary number of layers L. In particular, this means that the infimum is taken over all possible network topologies. The resulting best M-edge approximation rate is fundamental as it benchmarks all learning algorithms, i.e., all algorithms that map an input function f and an $\varepsilon > 0$ to a neural network that approximates f with error no more than ε. Our framework hence provides a means for assessing the performance of a given learning algorithm in the sense of allowing to measure how close the M-edge approximation rate induced by the algorithm is to the best M-edge approximation rate $\gamma^*_{\mathcal{NN}}(\mathcal{C},\rho)$.

1.4. Previous Work. The best-known results on approximation by neural networks are the universal approximation theorems of Hornik [34] and Cybenko [12], stating that every measurable function f can be approximated arbitrarily well by a single-hidden-layer ($L = 2$ in our terminology) neural network. The literature on approximation-theoretic properties of networks with a single hidden layer continuing this line of work is abundant. Without any claim to completeness, we mention work on approximation error bounds in terms of the number of neurons for functions with bounded first moments [1], [2], the non-existence of localized approximations [7], a fundamental lower bound on approximation rates [18, 4], and the approximation of smooth or analytic functions [44, 42].

Approximation-theoretic results for networks with multiple hidden layers were obtained in [35, 43] for general functions, in [23] for continuous functions, and for functions together with their derivatives in [47]. In [7] 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 [21] 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 neural networks, first results of a nature similar to those in [21] were reported in [46]. For the activation function $\rho(x) = \max\{0, x\}$—usually referred to as ReLU—it was demonstrated in [58, 50] that deep networks approximate sufficiently smooth functions with rates higher than those achieved by shallow networks.

Convolutional neural networks are obtained as a special case of the general networks considered in this paper, namely by taking the affine linear transformations in Definition 1.1 to have circulant A-matrices. Linking the expressivity properties of neural networks to tensor decompositions, [9, 10] establish the existence of functions that can be realized by relatively small deep convolutional networks but require exponentially larger shallow convolutional networks. Universal approximation theorems for convolutional neural networks are provided in [59, 57]. The approximation-theoretic equivalence results between convolutional networks and general networks established in [49], together with the main findings of the present paper, lead to upper and lower bounds on approximation rates attainable with convolutional networks.

For survey articles on approximation-theoretic aspects of neural networks, we refer the interested reader to [22, 51].

Most closely related to our work is that by Shaham, Cloninger, and Coifman [53], which shows that for functions that are sparse in specific wavelet frames, the best M-edge approximation rate of three-layer neural networks is at least as high as the best M-term approximation rate in piecewise linear wavelet frames.

1.5. Contributions. Our contributions can be grouped into four threads.

- *Fundamental lower bound on connectivity.* We quantify the minimum network connectivity needed to allow approximation of *all* elements of a given function class $\mathcal{C}$ to within a maximum allowed error. On a conceptual level, this result establishes a universal link between the complexity of a given function class and the connectivity required by corresponding approximating neural networks.
- *Transfer from M-term to M-edge approximation.* We develop a general framework for transferring best M-term approximation results in representation systems to best M-edge approximation results for neural networks.

- *Memory requirements*. We characterize the memory requirements needed to store the topology and the quantized weights of optimally-approximating neural networks.
- *Realizability of optimal approximation rates.* An important practical question is how neural networks trained by stochastic gradient descent (via backpropagation) [52] perform relative to the fundamental bounds established in the paper. Interestingly, our numerical experiments indicate that stochastic gradient descent can achieve M-edge approximation rates quite close to the fundamental limit.

1.6. Outline of the Paper. Section 2 introduces the novel concept of effective best M-edge approximation. The fundamental lower bound on connectivity is developed in Section 3. Section 4 describes a general framework for transferring best M-term approximation results in representation systems to best M-edge approximation results for neural networks. In Section 5, we apply this transfer framework to the broad class of affine representation systems, and Section 6 shows that this leads to optimal M-edge approximation rates for cartoon functions. In Section 7, we briefly outline the extension of our main findings to the approximation of functions defined on manifolds. Finally, numerical results assessing the performance of stochastic gradient descent (via backpropagation) relative to our lower bound on connectivity are reported in Section 8.

2. Effective Best M-term and Best M-edge Approximation.

We proceed by introducing M-term approximation via dictionaries and M-edge approximation via neural networks. These concepts do, however, not allow for a meaningful notion of optimality in practice. A remedy is provided by effective best M-term approximation according to [19, 26] and the new concept of effective best M-edge approximation introduced below.

2.1. Effective Best M-term Approximation. The best M-term approximation rate $\gamma^*(\mathcal{C},\mathcal{D})$ according to Definition 1.2 measures the hardness of approximation of a given function class $\mathcal{C}$ by a fixed representation system $\mathcal{D}$. It is sensible to ask whether for a given function class $\mathcal{C}$, there is a fundamental limit on $\gamma^*(\mathcal{C},\mathcal{D})$ when one is allowed to vary over $\mathcal{D}$. As shown in [19, 26], every dense (and countable) $\mathcal{D}\subset L^2(\Omega)$, $\Omega\subset\mathbb{R}^d$, results in $\gamma^*(\mathcal{C},\mathcal{D})=\infty$ for all function classes $\mathcal{C}\subset L^2(\Omega)$. However, identifying the elements in $\mathcal{D}$ participating in the best M-term approximation is infeasible as it entails searching through the infinite set $\mathcal{D}$ and requires, in general, an infinite number of bits to describe the indices of the participating elements. This insight leads to the concept of "best M-term approximation subject to polynomial-depth search" as introduced by Donoho in [19].

DEFINITION 2.1. *Given $d\in\mathbb{N}$, $\Omega\subset\mathbb{R}^d$, a function class $\mathcal{C}\subset L^2(\Omega)$, and a representation system $\mathcal{D}=(\varphi_i)_{i\in I}\subset L^2(\Omega)$, the supremal $\gamma>0$ so that there exist a polynomial π and a constant $D>0$ such that*

$$\sup_{f\in\mathcal{C}} \inf_{\substack{I_M\subset\{1,\dots,\pi(M)\},\\ \#I_M=M,\,(c_i)_{i\in I_M},\,\max_{i\in I_M}|c_i|\le D}} \left\| f-\sum_{i\in I_M} c_i\varphi_i\right\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M\to\infty, \tag{2.1}$$

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

We will demonstrate in Section 3.2 that $\sup_{\mathcal{D}\subset L^2(\Omega)}\gamma^{*,\mathrm{eff}}(\mathcal{C},\mathcal{D})$ is, indeed, finite under quite general conditions on $\mathcal{C}$ and, in particular, depends on the "description complexity" of $\mathcal{C}$. This will allow us to assess the approximation capacity of a given representation system $\mathcal{D}$ for $\mathcal{C}$ by comparing $\gamma^{*,\mathrm{eff}}(\mathcal{C},\mathcal{D})$ to the ultimate limit $\sup_{\mathcal{D}\subset L^2(\Omega)}\gamma^{*,\mathrm{eff}}(\mathcal{C},\mathcal{D})$.

2.2. Effective Best M-edge Approximation. We next aim at establishing a relationship in the spirit of effective best M-term approximation for approximation through deep

neural networks. To this end, we first note that Definition 1.3 encounters problems similar to those identified for approximation by representation systems, namely the quantity $\sup_{\rho:\mathbb{R}\to\mathbb{R}} \gamma^*_{\mathcal{NN}}(\mathcal{C},\rho)$ does not reveal anything tangible about the approximation complexity of $\mathcal{C}$ in deep neural networks, unless further constraints are imposed on the approximating network. To make this point, we first review the following remarkable result:

THEOREM 2.2 ([40]). *There exists a function $\rho:\mathbb{R}\to\mathbb{R}$ that is C^∞, strictly increasing, and satisfies $\lim_{x\to\infty}\rho(x)=1$ and $\lim_{x\to-\infty}\rho(x)=0$, such that for any $d\in\mathbb{N}$, any $f\in C([0,1]^d)$, and any $\varepsilon>0$, there is a neural network Φ with activation function ρ and three layers, of dimensions $N_1=3d$, $N_2=6d+3$, and $N_3=1$, satisfying*

$$\sup_{x\in[0,1]^d}|f(x)-\Phi(x)|\leq\varepsilon. \tag{2.2}$$

We observe that the number of nodes and the number of layers of the approximating network in Theorem 2.2 do not depend on the approximation error ε. In particular, ε can be chosen arbitrarily small while having $\mathcal{M}(\Phi)$ bounded. By density of $C([0,1]^d)$ in $L^2([0,1]^d)$ and hence in all compact subsets of $L^2([0,1]^d)$, this implies the existence of an activation function $\rho:\mathbb{R}\to\mathbb{R}$ such that $\gamma^*_{\mathcal{NN}}(\mathcal{C},\rho)=\infty$ for all compact $\mathcal{C}\subset L^2([0,1]^d), d\in\mathbb{N}$. However, the networks underlying Theorem 2.2 necessarily lead to weights that are (in absolute value) not bounded by $|\pi(\varepsilon^{-1})|$ for a polynomial π, a requirement we will have to impose to get rate-distortion-optimal approximation through neural networks (see Section 3). To see that the weights, indeed, do not obey a polynomial growth bound in ε^{-1}, we note that thanks to Theorem 2.2, there exist $C>0$ and $\gamma>0$ such that

$$\sup_{f\in\mathcal{C}}\inf_{\Phi_M\in\mathcal{NN}_{3,M,d,\rho}}\|f-\Phi_M\|_{L^2(\Omega)}\leq CM^{-\gamma},\ \text{for all } M\in\mathbb{N}. \tag{2.3}$$

Now, as ε in Theorem 2.2 can be made arbitrarily small while the connectivity of the corresponding networks remains upper-bounded by $21d^2+15d+3$, (2.3) would have to hold for arbitrarily large γ, in particular also for $\gamma>\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C},\rho)$, where $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C},\rho)$ is the effective best M-edge approximation rate according to Definition 2.3. By Theorem 3.4 below, however, $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C},\rho)\leq\gamma^*(\mathcal{C})$, where $\gamma^*(\mathcal{C})$ is the optimal exponent according to Definition 3.1. Owing to Definition 2.3, we can therefore conclude that the weights of the network achieving the infimum in (2.3) can not be bounded by a polynomial in $M\sim\varepsilon^{-1}$ whenever $\gamma^*(\mathcal{C})<\infty$. Here and in the sequel, we write $a\sim b$ if the variables a and b are proportional, i.e., there exist uniform constants $c_1,c_2>0$ such that $c_1a\leq b\leq c_2a$.

The observation just made resembles the problem in best M-term approximation which eventually led to the concept of *effective* best M-term approximation, where we restricted the search depth in the representation system $\mathcal{D}$ to be polynomially bounded in M and the coefficients c_i to be bounded according to $\max_{i\in I_M}|c_i|\leq D$. Interpreting the weights in the network as the counterpart of the coefficients c_i in best M-term approximation, we see that the restriction on the search depth corresponds to restricting the size of the indices enumerating the participating weights. The need for such a restriction is obviated by the tree structure of deep neural networks as exposed in detail in the proof of Proposition 3.6. The second restriction will lead us to a growth condition on the weights, which is more generous than the corresponding requirement of the c_i in effective best M-term approximation being bounded.

In summary, this leads to the novel concept of "best M-edge approximation subject to polynomial weight growth" as formalized next.

DEFINITION 2.3. *Given $d\in\mathbb{N}$, $\Omega\subset\mathbb{R}^d$, a function class $\mathcal{C}\subset L^2(\Omega)$, and an activation function $\rho:\mathbb{R}\to\mathbb{R}$, the supremal $\gamma>0$ so that there exist an $L\in\mathbb{N}$ and a polynomial π*

such that

$$\sup_{f\in\mathcal{C}} \inf_{\Phi_M\in\mathcal{NN}^{\pi}_{L,M,d,\rho}} \|f-\Phi_M\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M\to\infty, \tag{2.4}$$

where $\mathcal{NN}^{\pi}_{L,M,d,\rho}$ denotes the class of networks in $\mathcal{NN}_{L,M,d,\rho}$ that have all their weights bounded in absolute value by $|\pi(M)|$, will be referred to as effective best M-edge approximation rate of $\mathcal{C}$ by neural networks *and denoted by $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C},\rho)$.*

We will show in Corollary 3.4 that $\sup_{\rho:\mathbb{R}\to\mathbb{R}} \gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C},\rho)$ is bounded and depends on the "description complexity" of the function class $\mathcal{C}$.

3. Fundamental Bounds on Effective M-Term and M-Edge Approximation Rate.

The purpose of this section is to establish fundamental bounds on effective best M-term and effective best M-edge approximation rates by evaluating the corresponding approximation strategies in the min-max rate distortion theory framework as developed in [19, 26].

3.1. Min-Max Rate Distortion Theory.

Min-max rate distortion theory provides a theoretical foundation for deterministic lossy data compression. We recall the following notions and concepts from [19, 26].

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

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

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

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

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

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

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

DEFINITION 3.1. *Let $d\in\mathbb{N}$, $\Omega\subset\mathbb{R}^d$, and $\mathcal{C}\subset L^2(\Omega)$. Then, for $\varepsilon>0$, the* minimax code length $L(\varepsilon,\mathcal{C})$ *is*

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

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

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

The optimal exponent $\gamma^*(\mathcal{C})$ quantifies the minimum growth rate of $L(\varepsilon,\mathcal{C})$ as the error ε tends to zero and can hence be seen as quantifying the "description complexity" of the function class $\mathcal{C}$. Larger $\gamma^*(\mathcal{C})$ results in smaller growth rate and hence smaller memory requirements for storing signals $f\in\mathcal{C}$ such that reconstruction with uniformly bounded error is possible. The quantity $\gamma^*(\mathcal{C})$ is closely related to the concept of Kolmogorov entropy [48]. Remark 5.10 in [26] makes this connection explicit.

The optimal exponent is known for several function classes, such as subsets of Besov spaces $B^s_{p,q}(\mathbb{R}^d)$ with $1 \leq p, q < \infty, s > 0$, and $q > (s + 1/2)^{-1}$, namely all functions in $B^s_{p,q}(\mathbb{R}^d)$ of bounded norm, see e.g. [8]. If $\mathcal{C}$ is a bounded subset of $B^s_{p,q}(\mathbb{R}^d)$, then we have $\gamma^*(\mathcal{C}) = s/d$. In the present paper, we shall be particularly interested in so-called β-cartoon-like functions, for which the optimal exponent is given by $\beta/2$ (see [20, 28] and Theorem 6.3).

3.2. Fundamental Bound on Effective Best M-Term Approximation Rate. We next recall a result from [19, 26], which says that, for a given function class $\mathcal{C}$, the optimal exponent $\gamma^*(\mathcal{C})$ constitutes a fundamental bound on the effective best M-term approximation rate of $\mathcal{C}$ in any representation system. This gives operational meaning to $\gamma^*(\mathcal{C})$.

THEOREM 3.2 ([19, 26]). *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, $\mathcal{C} \subset L^2(\Omega)$, and assume that the effective best M-term approximation rate of $\mathcal{C}$ in $\mathcal{D} \subset L^2(\Omega)$ is $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D})$. Then, we have*

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

In light of this result the following definition is natural (see also [26]).

DEFINITION 3.3. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and assume that the effective best M-term approximation rate of $\mathcal{C} \subset L^2(\Omega)$ in $\mathcal{D} \subset L^2(\Omega)$ is $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D})$. If*

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

then the function class $\mathcal{C}$ is said to be optimally representable *by $\mathcal{D}$.*

3.3. Fundamental Bound on Effective Best M-Edge Approximation Rate. We now state the first main result of the paper, namely the equivalent of Theorem 3.2 for approximation by deep neural networks. Specifically, we establish that the optimal exponent $\gamma^*(\mathcal{C})$ also constitutes a fundamental bound on the effective best M-edge approximation rate of $\mathcal{C}$. We say below that a function $f : \mathbb{R} \to \mathbb{R}$ is *dominated* by a function $g : \mathbb{R} \to \mathbb{R}$ if $|f(x)| \leq |g(x)|$, for all $x \in \mathbb{R}$.

THEOREM 3.4. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and $\mathcal{C} \subset L^2(\Omega)$. Then, for all $\rho : \mathbb{R} \to \mathbb{R}$ that are Lipschitz-continuous or differentiable with ρ' dominated by a polynomial, we have*

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

The key ingredients of the proof of Theorem 3.4 are developed throughout this section and the formal proof will be stated at the end of the section. Before embarking, we note that, in analogy to Definition 3.3, what we just found suggests the following.

DEFINITION 3.5. *For $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ bounded, we say that the function class $\mathcal{C} \subset L^2(\Omega)$ is* optimally representable by neural networks with activation function $\rho : \mathbb{R} \to \mathbb{R}$, *if*

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

It is remarkable that the fundamental limits of approximation through representation systems and approximation through deep neural networks are determined by the same quantity, although the approximants in the two cases are vastly different, namely linear combinations of elements of a representation system with the participating functions identified subject to a polynomial-depth search constraint in the former, and concatenations of affine functions followed by non-linearities under growth constraints on the weights in the network in the latter case.

A key ingredient of the proof of Theorem 3.4 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.

PROPOSITION 3.6. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, $\rho : \mathbb{R} \to \mathbb{R}$, $c > 0$, and $\mathcal{C} \subset L^2(\Omega)$. Further, let*

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

be a map such that, for each pair $(\varepsilon, f) \in (0, 1/2) \times \mathcal{C}$, every weight of the neural network $\mathbf{Learn}(\varepsilon, f)$ is represented by no more than $\lceil c \log_2(\varepsilon^{-1}) \rceil$ bits while guaranteeing that*

$$\sup_{f \in \mathcal{C}} \|f - \mathbf{Learn}(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon. \tag{3.1}$$

Then,

$$\sup_{f \in \mathcal{C}} \mathcal{M}(\mathbf{Learn}(\varepsilon, f)) \notin \mathcal{O}\left(\varepsilon^{-\frac{1}{\gamma}}\right), \ \varepsilon \to 0, \quad \text{for all } \gamma > \gamma^*(\mathcal{C}). \tag{3.2}$$

Proof. The proof is by contradiction. To this end, let $\gamma > \gamma^*(\mathcal{C})$ and assume that $\sup_{f \in \mathcal{C}} \mathcal{M}(\mathbf{Learn}(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma}), \varepsilon \to 0$. 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

$$\begin{aligned} \ell(\varepsilon) &\leq C_0 \cdot \sup_{f \in \mathcal{C}} [\mathcal{M}(\mathbf{Learn}(\varepsilon, f)) \log_2(\mathcal{M}(\mathbf{Learn}(\varepsilon, f))) + 1] \log_2(\varepsilon^{-1}) \\ &\leq C_0 \cdot \left[\varepsilon^{-\frac{1}{\gamma}} \log_2\left(\varepsilon^{-\frac{1}{\gamma}}\right) + 1\right] \log_2(\varepsilon^{-1}) \\ &\leq C_1 \left(\varepsilon^{-\frac{1}{\gamma}} (\log_2(\varepsilon^{-1}))^2 + \log_2(\varepsilon^{-1})\right) \in \mathcal{O}\left(\varepsilon^{-\frac{1}{\nu}}\right), \text{ for } \varepsilon \to 0, \end{aligned} \tag{3.3}$$

where $C_0, C_1 > 0$ are constants and $\gamma > \nu > \gamma^*(C)$. Such a construction stands in contradiction to the optimality of $\gamma^*(\mathcal{C})$ according to Definition 3.1.

We proceed to the construction of the encoder-decoder pairs $(E_\varepsilon, D_\varepsilon) \in \mathfrak{E}^{\ell(\varepsilon)} \times \mathfrak{D}^{\ell(\varepsilon)}$. Fix $f \in \mathcal{C}$. We enumerate the nodes in $\mathbf{Learn}(\varepsilon, f)$ by assigning natural numbers, henceforth called *indices*, increasing from left to right in every layer as schematized in Figure 2. For the sake of notational simplicity, we also set $\Phi := \mathbf{Learn}(\varepsilon, f)$ and $M := \mathcal{M}(\Phi)$. Without loss of generality, we henceforth assume that M is a power of 2 and larger than 1. The case $M = 0$ will be dealt with in Step 1 below. For all M that are not powers of 2 and for $M = 1$, we make use of the fact that $\mathcal{NN}_{L,M,d,\rho} \subset \mathcal{NN}_{L,M',d,\rho}$, where M' is the smallest power of 2 larger than M, and we encode the network like an M'-edge network. Since $M < M' \leq 2M$ this affects $\ell(\varepsilon)$ by a multiplicative constant only.

We recall that the number of layers of Φ is denoted by L, the number of nodes in these layers is $N_1, \ldots, N_L$ (see Definition 1.1), and d stands for the dimension of the input layer.

Denoting the number of nodes in layer $\ell = 1, ..., L-1$ associated with edges of nonzero weight in the following layer by $\widetilde{N}_\ell$ and setting $\widetilde{N}_L = N_L = 1$, it follows that

$$d + \sum_{\ell=1}^{L} \widetilde{N}_\ell \leq 2\widetilde{M}, \tag{3.4}$$

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

$$L \leq \widetilde{M} \tag{3.5}$$

*Throughout the paper, we say that a weight of $\mathbf{Learn}(\varepsilon, f)$ is represented by no more than K bits, if it is taken from a set that is independent of f and has cardinality at most 2^K.

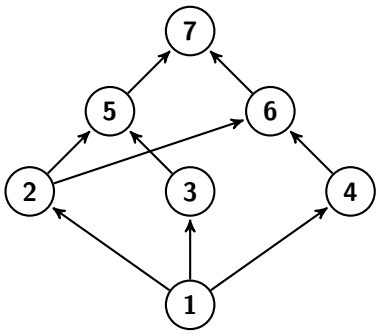


FIG. 2. *Enumeration of nodes as employed in the proof of the theorem.*

as otherwise there would be at least one layer $\ell > 1$ such that

$$A_\ell = 0.$$

As a consequence, the reduced network

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

realizes the same function as the original network Φ but has fewer than L layers. This reduction can be repeated inductively until the resulting reduced network satisfies (3.5).

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

Step 1: If $M = 0$, we encode the network by a leading 0 followed by the bitstring representing the node weight in the last layer. Upon defining $0 \log_2(0) = 0$, we then note that (3.3) holds trivially and we terminate the encoding procedure. Else, we encode the number of nonzero edge weights, M, by starting the overall bitstring with M 1's followed by a single 0. The length of this bitstring is therefore bounded by $\widetilde{M}$.

Step 2: We continue by encoding the number of layers in the network. Thanks to (3.5) this requires no more than $\log_2(\widetilde{M})$ bits. We thus reserve the next $\log_2(\widetilde{M})$ bits for the binary representation of L.

Step 3: Next, we store the dimension d of the input layer and the numbers of nodes $\widetilde{N}_\ell, \ell = 1, \dots, L$, associated with edges of nonzero weight. As by (3.4) $d \leq \widetilde{M}$ and $\widetilde{N}_\ell \leq 2\widetilde{M}$, for all ℓ, we can encode (generously) d and each $\widetilde{N}_\ell$ using $\log_2(\widetilde{M}) + 1$ bits. For the sake of concreteness, we first encode d followed by $\widetilde{N}_1, \dots, \widetilde{N}_L$ in that order. In total, Step 3 requires a bitstring of length

$$\left((L+1) \cdot \left(\log_2\left(\widetilde{M}\right) + 1\right)\right) \leq \left(\widetilde{M} + 1\right) \log_2\left(\widetilde{M}\right) + \widetilde{M} + 1.$$

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

$$\widetilde{M} \log_2\left(\widetilde{M}\right) + 2\log_2\left(\widetilde{M}\right) + 2\widetilde{M} + 1. \tag{3.6}$$

Step 4: We encode the topology of the graph associated with Φ and consider only nodes that contribute to the mapping $\Phi(x)$. Recall that we assigned a unique index i, ranging from 1 to $\widetilde{N} := d + \sum_{\ell=1}^{L} \widetilde{N}_\ell$, to each of these nodes. By (3.4) each of these indices can be encoded by a bitstring of length $\log_2(\widetilde{M}) + 1$. We denote the bitstring corresponding to index i by $b(i) \in \{0,1\}^{\log_2(\widetilde{M})+1}$ and let $n(i)$ be the number of children of the node with index i, i.e.,

the number of nodes in the next layer connected to the node with index i via an edge. (Here, $n(\widetilde{N})=0$.) For each node $i=1,\dots,\widetilde{N}$, we form a bitstring of length $n(i)\cdot(\log_2(\widetilde{M})+1)$ by concatenating the bitstrings $b(j)$ for all j such that there is an edge between i and j. We follow this string with an all-zeros bitstring of length $\log_2(\widetilde{M})+1$ to signal the transition to the node with index $i+1$. The enumeration is concluded with an all-zeros bitstring of length $\log_2(\widetilde{M})+1$ signaling that the last node has been reached. Overall, this yields a bitstring of length

$$\sum_{i=1}^{\widetilde{N}}(n(i)+1)\cdot\left(\log_2\left(\widetilde{M}\right)+1\right)\leq 3\widetilde{M}\cdot\left(\log_2\left(\widetilde{M}\right)+1\right), \tag{3.7}$$

where we used $\sum_{i=1}^{\widetilde{N}} n(i)=M<\widetilde{M}$ and (3.4). Combining (3.6) and (3.7) it follows that we have encoded the overall topology of the network Φ using at most

$$5\widetilde{M}+4\widetilde{M}\log_2\left(\widetilde{M}\right)+2\log_2\left(\widetilde{M}\right)+1 \tag{3.8}$$

bits.

Step 5: We encode the weights of Φ. By assumption, each weight can be represented by a bitstring of length $\lceil c\log_2(\varepsilon^{-1})\rceil$. For each node $i=1,\dots,\widetilde{N}$, we reserve the first $\lceil c\log_2(\varepsilon^{-1})\rceil$ bits to encode its associated node weight and, for each of its children a bitstring of length $\lceil c\log_2(\varepsilon^{-1})\rceil$ to encode the weight corresponding to the edge between that child and its parent node. Concatenating the results in ascending order of child node indices, we get a bitstring of length $(n(i)+1)\cdot(\lceil c\log_2(\varepsilon^{-1})\rceil)$ for node i, and an overall bitstring of length

$$\sum_{i=1}^{\widetilde{N}}(n(i)+1)\cdot\left(\lceil c\log_2\left(\varepsilon^{-1}\right)\rceil\right)\leq 3\widetilde{M}\cdot\lceil c\log_2\left(\varepsilon^{-1}\right)\rceil \tag{3.9}$$

representing the weights of the graph associated with the network Φ.

With (3.8) this shows that the overall number of bits needed to encode the network topology and weights is no more than

$$5\widetilde{M}+4\widetilde{M}\log_2\left(\widetilde{M}\right)+2\log_2\left(\widetilde{M}\right)+1+3\widetilde{M}\cdot\lceil c\log_2\left(\varepsilon^{-1}\right)\rceil. \tag{3.10}$$

The network can be recovered by sequentially reading out M, L, d, the $\widetilde{N}_\ell$, the topology, and the quantized weights from the overall bitstring. It is not difficult to verify that the individual steps in the encoding procedure were crafted such that this yields unique recovery. As (3.10) can be upper-bounded by

$$C_0 M\log_2(M)\log_2\left(\varepsilon^{-1}\right) \tag{3.11}$$

for a constant $C_0>0$ depending on c and d only, we have constructed an encoder-decoder pair $(E,D)\in\mathfrak{E}^{\ell(\varepsilon)}\times\mathfrak{D}^{\ell(\varepsilon)}$ with $\ell(\varepsilon)$ satisfying (3.3). This concludes the proof. ∎

Proposition 3.6 applies to networks that have each weight represented by a finite number of bits scaling according to $\log_2(\varepsilon^{-1})$ while guaranteeing that the underlying encoder-decoder pair achieves uniform error ε over $\mathcal{C}$. We next show that such a compatibility is possible for networks with activation functions that are either Lipschitz or differentiable such that ρ' is dominated by a polynomial.

LEMMA 3.7. *Let $d, L, k, M \in \mathbb{N}, \eta \in (0, 1/2), \Omega \subset \mathbb{R}^d$ be bounded, and let $\rho : \mathbb{R} \to \mathbb{R}$ be either Lipschitz-continuous or differentiable such that ρ' is dominated by a polynomial. Let $\Phi \in \mathcal{NN}_{L,M,d,\rho}$ with $M \leq \eta^{-k}$ and all its weights be bounded (in absolute value) by η^{-k}. Then, there exist $m \in \mathbb{N}$, depending on k, L, and ρ only, and $\widetilde{\Phi} \in \mathcal{NN}_{L,M,d,\rho}$ such that*

$$\left\| \widetilde{\Phi} - \Phi \right\|_{L^\infty(\Omega)} \leq \eta$$

and all weights of $\widetilde{\Phi}$ are elements of $\eta^m \mathbb{Z} \cap [-\eta^{-k}, \eta^{-k}]$.

Proof. We prove the statement for Lipschitz-continuous ρ only. The argument for differentiable activation functions with first derivative not growing faster than polynomial is along similar lines.

Without loss of generality, we can take the number of nonzero node weights of Φ to be upper-bounded by twice the number of nonzero edge weights. This assumption is justified as else there would be nodes that are not connected to the next layer through an edge of nonzero weight and we could replace the corresponding node weights by zero without altering the mapping Φ. This reduction would then lead to the assumption being justified.

Let $m \in \mathbb{N}$, to be specified later, and denote by $\widetilde{\Phi}$ the network that results by replacing all weights of Φ by a closest element in $\eta^m \mathbb{Z} \cap [-\eta^{-k}, \eta^{-k}]$. Set $C_{\max} := \eta^{-k}$ and denote the maximum of 1 and the total number of nonzero edge weights plus nonzero node weights of Φ by C_W. Note that $C_W \leq 3M \leq 3\eta^{-k}$, where the latter inequality is by assumption. For $\ell = 1, \ldots, L-1$, define $\Phi^\ell : \Omega \to \mathbb{R}^{N_\ell}$ as

$$\Phi^\ell(x) := \rho\left(W_\ell \rho\left(\ldots \rho\left(W_1(x)\right)\right)\right), \quad \text{for } x \in \Omega,$$

and $\widetilde{\Phi}^\ell$ accordingly, and let, for $\ell = 1, \ldots, L-1$,

$$e_\ell := \left\| \Phi^\ell - \widetilde{\Phi}^\ell \right\|_{L^\infty(\Omega, \mathbb{R}^{N_\ell})}, \quad e_L := \left\| \Phi - \widetilde{\Phi} \right\|_{L^\infty(\Omega)}.$$

Denote the maximum of 1 and the Lipschitz constant of ρ by C_ρ, set $C_0 := \max\{1, \sup\{|x| : x \in \Omega\}\}$, and let

$$C_\ell := \max\left\{ \left\| \Phi^\ell \right\|_{L^\infty(\Omega, \mathbb{R}^{N_\ell})}, \left\| \widetilde{\Phi}^\ell \right\|_{L^\infty(\Omega, \mathbb{R}^{N_\ell})} \right\}, \quad \text{for } \ell = 1, \ldots, L-1.$$

Then, straightforward, albeit somewhat tedious algebraic manipulations, show that for all $\ell = 2, \ldots, L-1$,

$$e_1 \leq C_0\, C_\rho\, C_W\, \eta^m, \text{ and } e_\ell \leq C_\rho\, C_W\, C_{\ell-1}\, \eta^m + C_\rho\, C_W\, C_{\max}\, e_{\ell-1}. \tag{3.12}$$

Additionally, we observe that

$$e_L \leq C_W\, C_{L-1}\, \eta^m + C_W\, C_{\max}\, e_{L-1}. \tag{3.13}$$

We now bound the quantity C_ℓ for $\ell = 1, \ldots, L-1$. A simple computation, exploiting the Lipschitz-continuity of ρ, yields

$$C_\ell \leq (|\rho(0)| + C_\rho\, C_W\, C_{\max}\, C_{\ell-1}), \quad \text{for all } \ell = 1, \ldots, L-1.$$

Since ρ is continuous on $\mathbb{R}$ we have $|\rho(0)| < \infty$ and thus, by $C_\rho, C_W, C_{\max} \geq 1$, there exists $C' > 0$ such that

$$C_\ell \leq C'\, C_0\, (C_\rho\, C_W\, C_{\max})^\ell, \quad \text{for all } \ell = 1, \ldots, L-1.$$

As C_W and $C_{\max}$ are both bounded by η^{-k-2}, it follows that C_ℓ is bounded by η^{-p} for a $p \in \mathbb{N}$. We can therefore find $n \in \mathbb{N}$ such that for all $\ell = 1, \ldots, L-1$,

$$\max\{C_0 C_\rho C_W, C_W\, C_{\max}, C_W\, C_{L-1}, C_\rho\, C_W\, C_{\ell-1}, C_\rho\, C_W\, C_{\max}\} \leq \frac{\eta^{-n}}{2}. \tag{3.14}$$

Invoking (3.12), we conclude that

$$e_\ell \leq \frac{\eta^{-n}}{2}(\eta^m + e_{\ell-1}), \quad \text{for all } \ell = 1, \ldots, L-1, \tag{3.15}$$

where we set $e_0 = 0$. We proceed by induction to prove that there exists $r \in \mathbb{N}$ such that for all $\ell = 1, \ldots, L-1$,

$$e_\ell \leq \eta^{m-(\ell-1)n-r}. \tag{3.16}$$

Clearly there exists $r \in \mathbb{N}$ such that $e_1 \leq \eta^{m-r}$. Moreover, one easily verifies that the existence of an $r \in \mathbb{N}$ such that (3.16) is satisfied for an $\ell \in \{1, \ldots, L-2\}$, thanks to (3.15), implies the existence of an $r \in \mathbb{N}$ such that (3.16) is satisfied for ℓ replaced by $\ell + 1$. This concludes the induction argument.

Using (3.14) and (3.16) in (3.13), we finally obtain

$$e_L \leq \frac{\eta^{m-n}}{2} + \frac{\eta^{m-(L-1)n-r}}{2},$$

which yields $e_L \leq \eta$ for sufficiently large m. □

Remark 3.8. Note that the weights of the network being elements of $\eta^m \mathbb{Z} \cap [-\eta^{-k}, \eta^{-k}]$ implies that each weight can be represented by no more than $\lceil c \log_2(\eta^{-1}) \rceil$ bits, for some constant $c > 0$.

Proposition 3.6 not only says that the connectivity growth rate of networks achieving uniform approximation error ε over a function class $\mathcal{C}$ must exceed $\mathcal{O}\big(\varepsilon^{-1/\gamma^*(\mathcal{C})}\big)$, $\varepsilon \to 0$, but its proof, by virtue of constructing an encoder-decoder pair that achieves this growth rate also provides an achievability result. We next establish a matching strong converse in the sense of showing that for $\gamma > \gamma^*(\mathcal{C})$, the uniform approximation error remains bounded away from zero for infinitely many $M \in \mathbb{N}$. To simplify terminology in the sequel, we introduce the notion of a polynomially bounded variable.

DEFINITION 3.9. *A real variable X depending on the variables $z_i \in D_i \subset \mathbb{R}$, $i = 1, \ldots, N$, is said to be* polynomially bounded in $z_1, \ldots, z_N$, *if there exists an N-variate polynomial π such that $|X| \leq |\pi(z_1, \ldots, z_N)|$, for all $z_i \in D_i, i = 1, \ldots, N$. A set of real variables $(X_j)_{j \in J}$, each depending on $z_i \in D_i \subset \mathbb{R}$, $i = 1, \ldots, N$, is* uniformly polynomially bounded in $z_1, \ldots, z_N$, *if there exists an N-variate polynomial π such that $|X_j| \leq |\pi(z_1, \ldots, z_N)|$, for all $j \in J$ and all $z_i \in D_i$, $i = 1, \ldots, N$.*

We will refrain from explicitly specifying the D_i in Definition 3.9 whenever they are clear from the context.

Remark 3.10. If $D_i = \mathbb{R} \setminus [-B_i, B_i]$ for some $B_i \geq 1, i = 1, \ldots, N$, then a variable X depending on $z_i \in D_i, i = 1, \ldots, N$, is polynomially bounded in $z_1, \ldots, z_N$ if and only if there exists a $k \in \mathbb{N}$ such that $|X| \leq |z_1^k \cdot z_2^k \cdot \ldots \cdot z_N^k|$, for all $z_i \in D_i$.

PROPOSITION 3.11. *Let $d, L \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, π a polynomial, $\mathcal{C} \subset L^2(\Omega)$, $\rho : \mathbb{R} \to \mathbb{R}$ either Lipschitz-continuous or differentiable such that ρ' is dominated by a polynomial. Then, for all $C > 0$ and $\gamma > \gamma^*(\mathcal{C})$, we have that*

$$\sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{NN}^\pi_{L,M,d,\rho}} \|f - \Phi\|_{L^2(\Omega)} \geq C M^{-\gamma}, \text{ for infinitely many } M \in \mathbb{N}. \tag{3.17}$$

Proof. Let $\gamma > \gamma^*(\mathcal{C})$. Assume, towards a contradiction, that (3.17) holds only for finitely many $M \in \mathbb{N}$. Then, there exists a constant $C > 0$ such that the inequality in (3.17) holds for no $M \in \mathbb{N}$ and hence there exists $C' > 0$ so that

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

Setting $M_\varepsilon := \lceil (\varepsilon/(3C'))^{-1/\gamma} \rceil$, it follows that, for every $f \in \mathcal{C}$ and every $\varepsilon \in (0, 1/2)$, there exists a neural network $\Phi_{\varepsilon,f} \in \mathcal{NN}^{\pi}_{L,M_\varepsilon,d,\rho}$ such that

$$\|f-\Phi_{\varepsilon,f}\|_{L^2(\Omega)} \leq 2 \sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}^{\pi}_{L,M_\varepsilon,d,\rho}} \|f-\Phi\|_{L^2(\Omega)} \leq 2C' M_\varepsilon^{-\gamma} \leq \frac{2\varepsilon}{3}.$$

As the weights of $\Phi_{\varepsilon,f}$ are polynomially bounded in M_ε, they are polynomially bounded in ε^{-1}. By Lemma 3.7 and Remark 3.10, there hence exists a network $\widetilde{\Phi}_{\varepsilon,f}$ whose weights are represented by no more than $\lceil c \log_2(\varepsilon^{-1}) \rceil$ bits, for some constant $c > 0$, satisfying

$$\left\| \Phi_{\varepsilon,f} - \widetilde{\Phi}_{\varepsilon,f} \right\|_{L^2(\Omega)} \leq \frac{\varepsilon}{3}.$$

Defining

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

it follows that

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

The proof is concluded by noting that **Learn** violates Proposition 3.6. ∎

We can now proceed to the proof of Theorem 3.4.

Proof of Theorem 3.4. Suppose towards a contradiction that $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C},\rho) > \gamma^*(\mathcal{C})$. Let $\gamma \in \left(\gamma^*(\mathcal{C}), \gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C},\rho)\right)$. Then, Definition 2.3 implies that there exist a polynomial $\pi, L \in \mathbb{N}$, and $C > 0$ such that

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

This, however, constitutes a contradiction to Proposition 3.11. ∎

We conclude this section with a discussion of the conceptual implications of the results established above. Proposition 3.6 combined with Lemma 3.7 establishes that neural networks with weights polynomially bounded in ε^{-1} and achieving uniform approximation error ε over $\mathcal{C}$ cannot exhibit edge 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.

Note that requiring uniform approximation error ε only (without imposing the constraint of the network's weights being polynomially bounded in ε^{-1}) can lead to arbitrarily large rate γ as exemplified by Theorem 2.2, which proves the existence of networks realizing an arbitrarily small approximation error over $L^2([0,1]^d)$ with a finite number of nodes; in particular, the number of nodes remains constant as $\varepsilon \to 0$. However, as argued right after Theorem 2.2, these networks necessarily lead to weights that are not polynomially bounded in ε^{-1}.

Finally, we remark that the proofs of Theorem 3.4 and Proposition 3.6, by virtue of explicitly constructing encoder-decoder pairs for neural networks, provide a bound on the minimax code length of these networks. This, in turn, implies a bound on the networks' covering numbers, see [26], which, based on classical results from statistical learning theory (see for example [11]), leads to bounds on the generalization error, see e.g. [3].

4. Transitioning from Representation Systems to Neural Networks. The remainder of this paper is devoted to identifying function classes that are optimally representable—according to Definition 3.5—by neural networks. The mathematical technique we develop in the process is interesting in its own right as it constitutes a general framework for transferring results on function approximation through representation systems to results on approximation by neural networks. In particular, we prove that for a given function class $\mathcal{C}$ and an associated representation system $\mathcal{D}$ which satisfies certain technical conditions, there exists a neural network with $\mathcal{O}(M)$ nonzero edge weights that achieves (up to a multiplicative constant) the same uniform error over $\mathcal{C}$ as a best M-term approximation in $\mathcal{D}$. This will finally lead to a characterization of function classes $\mathcal{C}$ that are optimally representable by neural networks in the sense of Definition 3.5.

We start by stating technical conditions on representation systems for the transference principle outlined above to apply.

DEFINITION 4.1. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, $\rho : \mathbb{R} \to \mathbb{R}$, and $\mathcal{D} = (\varphi_i)_{i\in I} \subset L^2(\Omega)$ be a representation system. Then, $\mathcal{D}$ is said to be* representable by neural networks (with activation function ρ)*, if there exist $L, R \in \mathbb{N}$ such that for all $\eta > 0$ and every $i \in I$, there is a neural network $\Phi_{i,\eta} \in \mathcal{NN}_{L,R,d,\rho}$ with*

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

If, in addition, the weights of $\Phi_{i,\eta} \in \mathcal{NN}_{L,R,d,\rho}$ are polynomially bounded in i, η^{-1}, and if ρ is either Lipschitz-continuous or differentiable such that ρ' is dominated by a polynomial, then we say that $\mathcal{D}$ is effectively representable by neural networks (with activation function ρ).

The next result formalizes our transference principle for networks with weights in $\mathbb{R}$.

THEOREM 4.2. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and $\rho : \mathbb{R} \to \mathbb{R}$. Suppose that $\mathcal{D} = (\varphi_i)_{i\in I} \subset L^2(\Omega)$ is representable by neural networks. Let $f \in L^2(\Omega)$ and, for $M \in \mathbb{N}$, let $f_M = \sum_{i\in I_M} c_i\varphi_i$, $I_M \subset I$, $\#I_M = M$, satisfy*

$$\|f - f_M\|_{L^2(\Omega)} \leq \varepsilon,$$

where $\varepsilon \in (0, 1/2)$. Then, there exist $L \in \mathbb{N}$ (depending on $\mathcal{D}$ only) and a neural network $\Phi(f, M) \in \mathcal{NN}_{L,M',d,\rho}$ with $M' \in \mathcal{O}(M)$, satisfying

$$\|f - \Phi(f, M)\|_{L^2(\Omega)} \leq 2\varepsilon. \tag{4.1}$$

In particular, for all function classes $\mathcal{C} \subset L^2(\Omega)$, it holds that

$$\gamma^*_{\mathcal{NN}}(\mathcal{C}, \rho) \geq \gamma^*(\mathcal{C}, \mathcal{D}). \tag{4.2}$$

Proof. By representability of $\mathcal{D}$ according to Definition 4.1, it follows that there exist $L, R \in \mathbb{N}$, such that for each $i \in I_M$ and for $\eta := \varepsilon/\max\{1, \sum_{i\in I_M} |c_i|\}$, there exists a neural network $\Phi_{i,\eta} \in \mathcal{NN}_{L,R,d,\rho}$ with

$$\|\varphi_i - \Phi_{i,\eta}\|_{L^2(\Omega)} \leq \eta. \tag{4.3}$$

Let then $\Phi(f, M)$ be the neural network consisting of the networks $(\Phi_{i,\eta})_{i\in I_M}$ operating in parallel, all with the same input, and summing their one-dimensional outputs (see Figure 3 in Section 8 for an illustration) with weights $(c_i)_{i\in I_M}$ according to

$$\Phi(f, M)(x) := \sum_{i\in I_M} c_i \Phi_{i,\eta}(x), \quad \text{for } x \in \Omega. \tag{4.4}$$

This construction is legitimate as all networks $\Phi_{i,\eta}$ have the same number of layers and the last layer of a neural network according to Definition 1.1 implements an affine function only (without subsequent application of the activation function ρ). Then, the fact that $\Phi(f, M) \in \mathcal{NN}_{L,RM,d,\rho}$ and application of the triangle inequality together with (4.3) yields $\|f_M - \Phi(f, M)\|_{L^2(\Omega)} \leq \varepsilon$. Another application of the triangle inequality according to

$$\|f - \Phi(f, M)\|_{L^2(\Omega)} \leq \|f - f_M\|_{L^2(\Omega)} + \|f_M - \Phi(f, M)\|_{L^2(\Omega)} \leq 2\varepsilon$$

finalizes the proof of (4.1) which by Definitions 1.2 and 1.3 implies (4.2). ∎

Theorem 4.2 shows that we can restrict ourselves to the approximation of the individual elements of a representation system by neural networks with the only constraint being that the number of nonzero edge weights in the individual networks must admit a uniform upper bound. Theorem 4.2 does, however, not guarantee that the weights of the network $\Phi(f, M)$ can be represented with no more than $\lceil c \log_2(\varepsilon^{-1})\rceil$ bits when the overall approximation error is ε. This will again be accomplished through a transfer argument, applied to representation systems $\mathcal{D}$ satisfying slightly more stringent technical conditions.

THEOREM 4.3. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and $\mathcal{C} \subset L^2(\Omega)$. Suppose that the representation system $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ is effectively representable by neural networks. Then, for all $\gamma < \gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$, there exist a polynomial π, constants $c > 0$, $L \in \mathbb{N}$, and a map*

$$\mathbf{Learn} : \left(0, \frac{1}{2}\right) \times L^2(\Omega) \to \mathcal{NN}^{\pi}_{L,\infty,d,\rho},$$

such that for every $f \in \mathcal{C}$ the weights in $\mathbf{Learn}(\varepsilon, f)$ can be represented by no more than $\lceil c \log_2(\varepsilon^{-1})\rceil$ bits while $\|f - \mathbf{Learn}(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon$ and $\mathcal{M}(\mathbf{Learn}(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma})$, for $\varepsilon \to 0$.

Remark 4.4. Theorem 4.3 implies that if $\mathcal{D}$ optimally represents the function class $\mathcal{C}$ in the sense of Definition 3.3 and at the same time is effectively representable by neural networks, then $\mathcal{C}$ is optimally representable by neural networks in the sense of Definition 3.5.

Proof of Theorem 4.3. Let $M \in \mathbb{N}$ and $\gamma < \gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$. According to Definition 2.1, there exist constants $C, D > 0$ and a polynomial π such that for every $f \in \mathcal{C}$, there is a subset $I_M \subset \{1, \dots, \pi(M)\}$, and coefficients $(c_i)_{i\in I_M}$ with $\max_{i\in I_M} |c_i| \leq D$ so that

$$\left\| f - \sum_{i\in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \leq \frac{CM^{-\gamma}}{2} =: \frac{\delta_M}{2}. \tag{4.5}$$

We only need to consider the case $\delta_M \leq 1/2$ as will become clear below. By effective representability according to Definition 4.1, there are $L, R \in \mathbb{N}$ such that for each $i \in I_M$ and with $\eta := \delta_M / \max\{1, 4\sum_{i\in I_M} |c_i|\}$, there exists a neural network $\Phi_{i,\eta} \in \mathcal{NN}_{L,R,d,\rho}$ (with ρ either Lipschitz-continuous or differentiable such that ρ' is dominated by a polynomial) satisfying

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

In addition, the weights of $\Phi_{i,\eta}$ are polynomially bounded in i, η^{-1}. Let then $\Phi(f,M) \in \mathcal{NN}_{L,RM,d,\rho}$ be the neural network consisting of the networks $(\Phi_{i,\eta})_{i\in I_M}$ operating in parallel, according to (4.4). We conclude that

$$\left\| \sum_{i\in I_M} c_i\varphi_i - \Phi(f,M) \right\|_{L^2(\Omega)} \leq \frac{\delta_M}{4}.$$

As the weights of the networks $\Phi_{i,\eta}$ are polynomially bounded in i, η^{-1} and $i \leq \pi(M), \delta_M \sim M^{-\gamma}$, it follows that the weights of $\Phi(f,M)$ are polynomially bounded in δ_M^{-1},

$$\left\| \Phi(f,M) - \widetilde{\Phi}(f,M) \right\|_{L^2(\Omega)} \leq \frac{\delta_M}{4},$$

and all weights of $\widetilde{\Phi}(f,M)$ can be represented with no more than $\lceil c\log_2(\delta_M^{-1})\rceil$ bits, for some $c>0$. Moreover, we have

$$\begin{aligned} \left\| f - \widetilde{\Phi}(f,M) \right\|_{L^2(\Omega)} &\leq \left\| f - \sum_{i\in I_M} c_i\varphi_i \right\|_{L^2(\Omega)} + \left\| \sum_{i\in I_M} c_i\varphi_i - \Phi(f,M) \right\|_{L^2(\Omega)} \\ &\quad + \left\| \Phi(f,M) - \widetilde{\Phi}(f,M) \right\|_{L^2(\Omega)} \leq \delta_M = CM^{-\gamma}. \end{aligned} \tag{4.6}$$

For $\varepsilon \in (0,1/2)$, we now set

$$\mathbf{Learn}(\varepsilon, f) := \widetilde{\Phi}(f, M_\varepsilon),$$

where

$$M_\varepsilon := \left\lceil \left(\frac{C}{\varepsilon}\right)^{\frac{1}{\gamma}} \right\rceil. \tag{4.7}$$

With this choice of M_ε, we have $CM_\varepsilon^{-\gamma} \leq \varepsilon$, which, when used in (4.6), yields

$$\|f - \mathbf{Learn}(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon. \tag{4.8}$$

Since, by construction, $\mathbf{Learn}(\varepsilon, f)$ has RM_ε edges and, moreover, $M_\varepsilon \leq C^{1/\gamma}\varepsilon^{-1/\gamma}+1 \leq 2C^{1/\gamma}\varepsilon^{-1/\gamma}$, it follows that $\mathbf{Learn}(\varepsilon, f)$ has at most $2RC^{1/\gamma}\varepsilon^{-1/\gamma}$ edges. Moreover, as all weights of $\mathbf{Learn}(\varepsilon, f)$ can be represented by no more than $\lceil c\log_2(\delta_{M_\varepsilon}^{-1})\rceil$ bits, it follows from $\delta_{M_\varepsilon} \sim M_\varepsilon^{-\gamma} \sim \varepsilon$ that they can be represented by no more than $\lceil c'\log_2(\varepsilon^{-1})\rceil$ bits, for some $c'>0$. This concludes the proof. ∎

5. All Affine Representation Systems are Effectively Representable by Neural Networks. This section shows that a large class of representation systems, namely *affine systems*, defined below, is effectively representable by neural networks. Affine systems include wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules. Combined with Theorem 4.3 the results in this section establish that any function class that is optimally represented by an arbitrary affine system is optimally represented by neural networks in the sense of Definition 3.5.

Clearly, such strong statements are possible only under restrictions on the choice of the activation function for the approximating neural networks.

5.1. Choice of Activation Function. We consider two classes of activation functions, namely sigmoidal functions and smooth approximations of rectified linear units. We start with the formal definition of sigmoidal activation functions as considered in [12, 43, 45, 7].

DEFINITION 5.1. *A continuous function $\rho : \mathbb{R} \to \mathbb{R}$ is called a* sigmoidal function of order $k \in \mathbb{N}$, $k \geq 2$, *if there exists $C > 0$ such that*

$$\lim_{x\to-\infty} \frac{1}{x^k}\rho(x) = 0, \quad \lim_{x\to\infty} \frac{1}{x^k}\rho(x) = 1, \quad \textit{and} \quad |\rho(x)| \leq C(1+|x|)^k, \textit{ for } x \in \mathbb{R}.$$

A differentiable function ρ is called strongly sigmoidal of order k, *if there exist constants $a, b, C > 0$ such that*

$$\left|\frac{1}{x^k}\rho(x)\right| \leq C|x|^{-a}, \textit{ for } x < 0, \quad \left|\frac{1}{x^k}\rho(x) - 1\right| \leq Cx^{-a}, \textit{ for } x \geq 0, \quad \textit{and}$$

$$|\rho(x)| \leq C(1+|x|)^k, \ \left|\frac{d}{dx}\rho(x)\right| \leq C|x|^b, \quad \textit{for } x \in \mathbb{R}.$$

One of the most widely used activation functions is the so-called rectified linear unit (ReLU) given by $x \mapsto \max\{0, x\}$. The second class of activation functions we consider here are smooth versions of the ReLU.

DEFINITION 5.2. *Let $\rho : \mathbb{R} \to \mathbb{R}^+$, $\rho \in C^\infty(\mathbb{R})$, satisfy*

$$\rho(x) = \begin{cases} 0, & \textit{for } x \leq 0, \\ x, & \textit{for } x \geq K, \end{cases}$$

for some constant $K > 0$. Then, we call ρ an admissible smooth activation function.

The reason for considering these two specific classes of activation functions resides in the fact that neural networks based thereon allow economical representations of multivariate bump functions, which, in turn, leads to effective representation of all affine systems (built from bump functions) by neural networks. Approximation of multivariate bump functions using sparsely connected neural networks is a classical topic in neural network theory [38]. What is new here is the aspect of quantized weights and rate-distortion optimality.

Note that the smooth variants of the ReLU we consider here allow us to build on existing approximation results for bump functions. We emphasize, however, that smooth functions can be approximated—in a rate-distortion optimal fashion—by networks based on the regular ReLU activation function provided that their depth is allowed to scale poly-logarithmically in the inverse of the approximation error ε [30]. However, in Sections 5.1 and 6 we require smooth generators of affine systems to be approximated by neural networks with the number of weights remaining constant as $\varepsilon \to 0$. This is not possible with the regular ReLU activation function. We emphasize, though, that all results in this paper apart from those in Sections 5.1 and 6 are also valid for the regular ReLU activation function.

A class of bump functions of particular importance in wavelet theory are B-splines. In [7] it was shown that B-splines can be parsimoniously approximated by neural networks with sigmoidal activation functions. It is instructive to recall this result. To this end, for $m \in \mathbb{N}$, we denote the univariate cardinal B-spline of order $m \in \mathbb{N}$ by N_m, i.e., $N_1 = \chi_{[0,1]}$, where $\chi_{[0,1]}$ denotes the characteristic function of the interval $[0, 1]$, and $N_{m+1} = N_m * \chi_{[0,1]}$, for all $m \geq 1$. Multivariate B-splines are simply tensor products of univariate B-splines. Specifically, we denote, for $d \in \mathbb{N}$, the d-dimensional cardinal B-spline of order m by N_m^d.

THEOREM 5.3 ([7], Thm. 4.2). *Let $d, m, k \in \mathbb{N}$, and take ρ to be a sigmoidal function of order $k \geq 2$. Further, let $L := \lceil \log_2(md - d)/\log_2(k) \rceil + 1$. Then, there is $M \in \mathbb{N}$,*

possibly dependent on d, m, k, such that for all $D, \varepsilon > 0$, there exists a neural network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{L,M,d,\rho}$ with

$$\|N_m^d - \Phi_{D,\varepsilon}\|_{L^2([-D,D]^d)} \leq \varepsilon.$$

Additionally, we will need to control the weights in the approximating networks $\Phi_{D,\varepsilon}$. We next show that this is, indeed, possible for strongly sigmoidal activation functions.

THEOREM 5.4. *Let $d, m, k \in \mathbb{N}$, and let ρ be strongly sigmoidal of order $k \geq 2$. Further, let $L := \lceil \log_2(md - d)/\log_2(k) \rceil + 1$. Then, there is an $M \in \mathbb{N}$, possibly dependent on d, m, k, such that for all $D, \varepsilon > 0$, there exists a neural network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{L,M,d,\rho}$ with*

$$\|N_m^d - \Phi_{D,\varepsilon}\|_{L^2([-D,D]^d)} \leq \varepsilon.$$

Moreover, the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D, ε^{-1}.

Proof. The neural network $\Phi_{D,\varepsilon}$ in Theorem 5.3 is explicitly constructed in [7]. Carefully following the steps in that construction and making explicit use of the strong sigmoidality of ρ, as opposed to plain sigmoidality as in [7], yields the desired result. ∎

Remark 5.5. We observe that the number of edges of the approximating network in Theorem 5.4 does not depend on the approximation error ε.

While Theorem 5.3 demonstrates that a B-spline of order m can be approximated to arbitrary accuracy by a neural network based on a sigmoidal activation function and of depth depending on m, d, and the order of sigmoidality of the activation function, we next establish that for admissible smooth activation functions, exact representation of a general class of bump functions is possible with a network of three layers only. Before proceeding, we define for $f \in L^1(\mathbb{R}^d)$, $d \in \mathbb{N}$, the *Fourier transform* of f by

$$\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-2\pi i \langle x, \xi \rangle} dx, \text{ for } \xi \in \mathbb{R}^d.$$

THEOREM 5.6. *Let ρ be an admissible smooth activation function. Then, for all $d \in \mathbb{N}$, there exist $M \in \mathbb{N}$ and a neural network $\Phi_\rho \in \mathcal{NN}_{3,M,d,\rho}$ such that*

(i) Φ_ρ is compactly supported,
(ii) $\Phi_\rho \in C^\infty(\mathbb{R})$, and
(iii) $\widehat{\Phi}_\rho(\xi) \neq 0$, for all $\xi \in [-3, 3]^d$.

Proof. We start by constructing an auxiliary function as follows. For $0 < p_1 \leq p_2 \leq p_3$ such that $p_1 + p_2 = p_3$, define $t : \mathbb{R} \to \mathbb{R}$ as

$$t(x) := \rho(x) - \rho(x - p_1) - \rho(x - p_2) + \rho(x - p_3), \ x \in \mathbb{R}. \tag{5.1}$$

Then, $t \in C^\infty$ is compactly supported. Letting $q = \|t\|_{L^\infty(\mathbb{R})}$, we define $g : \mathbb{R}^d \to \mathbb{R}$ according to

$$g(x) := \rho\left(\sum_{i=1}^d t(x_i) - (d-1) \cdot q\right), \ x \in \mathbb{R}^d. \tag{5.2}$$

By construction, $g \in C^\infty$ is compactly supported. Moreover, g can be realized through a three-layer neural network thanks to its two-step design per (5.1) and (5.2). Since $g \geq 0$ and $g \neq 0$, it follows that $|\hat{g}(0)| > 0$. By continuity of $\hat{g}$ there exists a $\delta > 0$ such that $|\hat{g}(\xi)| > 0$ for all $\xi \in [-\delta, \delta]^d$. We now set

$$\varphi := g\left(3\left(\frac{\cdot}{\delta}\right)\right),$$

and note that φ can be realized through a three-layer neural network $\Phi_\rho \in \mathcal{NN}_{3,M,d,\rho}$, for some $M \in \mathbb{N}$. As $|\hat{\varphi}(\xi)| > 0$, for all $\xi \in [-3,3]^d$, Φ_ρ satisfies the desired assumptions. ∎

5.2. Invariance to Affine Transformations. We next leverage Theorems 5.4 and 5.6 to demonstrate that a wide class of representation systems built through affine transformations of B-splines and bump functions as constructed in Theorem 5.6 is effectively representable by neural networks. As a first step towards this general result, we show that representability—in the sense of Definition 4.1—of a single function f by neural networks is invariant to the operation of taking finite linear combinations of affine transformations of f.

PROPOSITION 5.7. *Let $d \in \mathbb{N}$, $\rho : \mathbb{R} \to \mathbb{R}$, and $f \in L^2(\mathbb{R}^d)$. Assume that there exist $M, L \in \mathbb{N}$ such that for all $D, \varepsilon > 0$, there is $\Phi_{D,\varepsilon} \in \mathcal{NN}_{L,M,d,\rho}$ with*

$$\|f - \Phi_{D,\varepsilon}\|_{L^2([-D,D]^d)} \leq \varepsilon. \tag{5.3}$$

Let $A \in \mathbb{R}^{d\times d}$ be full-rank and $b \in \mathbb{R}^d$. Then, there exists $M' \in \mathbb{N}$, depending on M and d only, such that for all $E, \eta > 0$, there is $\Psi_{E,\eta} \in \mathcal{NN}_{L,M',d,\rho}$ with

$$\left\| |\det(A)|^{\frac{1}{2}} f(A\cdot - b) - \Psi_{E,\eta} \right\|_{L^2([-E,E]^d)} \leq \eta.$$

Moreover, if the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D, ε^{-1}, then the weights of $\Psi_{E,\eta}$ are polynomially bounded in $\|A\|_\infty, E, \|b\|_\infty, \eta^{-1}$, where $\|A\|_\infty$ and $\|b\|_\infty$ denote the max-norm of A and b, respectively.

Proof. By a change of variables, we have for every $\Phi \in \mathcal{NN}_{L,M,d,\rho}$ that

$$\left\| |\det(A)|^{\frac{1}{2}} (f(A\cdot - b) - \Phi(A\cdot - b)) \right\|_{L^2([-E,E]^d)} = \|f - \Phi\|_{L^2(A\cdot[-E,E]^d - b)}, \tag{5.4}$$

and there exists a constant M' depending on M and d only such that $|\det(A)|^{1/2}\Phi(A\cdot - b) \in \mathcal{NN}_{L,M',d,\rho}$. We furthermore have that

$$A\cdot[-E,E]^d - b \subset [-(dE\|A\|_\infty + \|b\|_\infty), (dE\|A\|_\infty + \|b\|_\infty)]^d. \tag{5.5}$$

We now set $F = dE\|A\|_\infty + \|b\|_\infty$ and $\Psi_{E,\eta} := |\det(A)|^{1/2}\Phi_{F,\eta}(A\cdot - b)$ and observe that

$$\begin{aligned} \left\| |\det(A)|^{\frac{1}{2}} f(A\cdot - b) - \Psi_{E,\eta} \right\|_{L^2([-E,E]^d)} &= \|f - \Phi_{F,\eta}\|_{L^2(A\cdot[-E,E]^d - b)} \\ &\leq \|f - \Phi_{F,\eta}\|_{L^2([-F,F)]^d)} \leq \eta, \end{aligned}$$

where we applied the same reasoning as in (5.4) in the first equality and used (5.5) in the first inequality and (5.3) in the second. Moreover, we see that if the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D, ε^{-1}, then the weights of $\Psi_{E,\eta}$ are polynomially bounded in $\|A\|_\infty, |\det(A)|, E, \|b\|_\infty, \eta^{-1}$. Since $|\det(A)|$ is polynomially bounded in $\|A\|_\infty$, it follows that the weights of $\Psi_{E,\eta}$ are polynomially bounded in $\|A\|_\infty, E, \|b\|_\infty, \eta^{-1}$. This yields the claim. ∎

Next, we show that representability by neural networks is preserved under finite linear combinations of translates.

PROPOSITION 5.8. *Let $d \in \mathbb{N}$, $\rho : \mathbb{R} \to \mathbb{R}$, and $f \in L^2(\mathbb{R}^d)$. Assume that there exist $M, L \in \mathbb{N}$ such that for all $D, \varepsilon > 0$, there is $\Phi_{D,\varepsilon} \in \mathcal{NN}_{L,M,d,\rho}$ with*

$$\|f - \Phi_{D,\varepsilon}\|_{L^2([-D,D]^d)} \leq \varepsilon. \tag{5.6}$$

Let $r \in \mathbb{N}$, $(c_i)_{i=1}^r \subset \mathbb{R}$, and $(d_i)_{i=1}^r \subset \mathbb{R}^d$. Then, there exists $M' \in \mathbb{N}$, depending on M, d, and r only, such that for all $E, \eta > 0$, there is $\Psi_{E,\eta} \in \mathcal{NN}_{L,M',d,\rho}$ with

$$\left\| \sum_{i=1}^r c_i f(\cdot - d_i) - \Psi_{E,\eta} \right\|_{L^2([-E,E]^d)} \leq \eta. \tag{5.7}$$

Moreover, if the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D, ε^{-1}, then the weights of $\Psi_{E,\eta}$ are polynomially bounded in

$$\sum_{i=1}^r |c_i|, E, \max_{i=1,\dots,r} \|d_i\|_\infty, \eta^{-1}.$$

Proof. Let $E, \eta > 0$. We start by noting that, for all $D, \varepsilon > 0$,

$$\begin{aligned} &\left\| \sum_{i=1}^r c_i f(\cdot - d_i) - \sum_{i=1}^r c_i \Phi_{D,\varepsilon}(\cdot - d_i) \right\|_{L^2([-E,E]^d)} \\ &\leq \left(\sum_{i=1}^r |c_i| \right) \cdot \|f - \Phi_{D,\varepsilon}\|_{L^2([-(E+d^*),(E+d^*)]^d)}, \end{aligned}$$

where $d^* = \max_{i=1,\dots,r} \|d_i\|_\infty$. Setting $D = E + d^*$ and $\varepsilon = \eta / \max\{1, \sum_{i=1}^r |c_i|\}$, and noting that for every $\Phi \in \mathcal{NN}_{L,M,d,\rho}$, the function

$$\Psi := \sum_{i=1}^r c_i \Phi(\cdot - d_i)$$

is in $\mathcal{NN}_{L,M',d,\rho}$ with $M' \in \mathbb{N}$ depending on d, r, and M only, it follows that the network

$$\Psi_{E,\eta} := \sum_{i=1}^r c_i \Phi_{D,\varepsilon}(\cdot - d_i)$$

satisfies (5.7). Finally, if the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D, ε^{-1}, then the weights of $\Psi_{E,\eta}$ are polynomially bounded in $\sum_{i=1}^r |c_i|, E, d^*, \eta^{-1}$.

Based on the invariance results in Propositions 5.7 and 5.8, we now construct neural networks which approximate functions with a given number of vanishing moments with arbitrary accuracy. The resulting construction will be crucial in establishing representability of affine representation systems (see Definition 5.11) by neural networks.

DEFINITION 5.9. *Let $R, d \in \mathbb{N}$, and $k \in \{1, \dots, d\}$. A function $g \in C(\mathbb{R}^d)$ is said to possess R* directional vanishing moments *in the x_k-direction if, for all $\ell \in \{0, \dots, R-1\}$ and all $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_d \in \mathbb{R}$,*

$$\int_{\mathbb{R}} x_k^\ell g(x_1, \dots, x_k, \dots, x_d) dx_k = 0.$$

The next result establishes that functions with an arbitrary number of vanishing moments in a given coordinate direction can be built from suitable linear combinations of translates of a given continuous function with compact support.

LEMMA 5.10. *Let $R, d \in \mathbb{N}$, $B > 0$, $k \in \{1, \dots, d\}$, and $f \in C(\mathbb{R}^d)$ with compact support. Then, the function*

$$g(x_1, \dots, x_d) := \sum_{\ell=0}^{R-1} \binom{R-1}{\ell} (-1)^\ell f\left(x_1, \dots, x_k - \frac{\ell}{B}, \dots, x_d\right) \tag{5.8}$$

has R directional vanishing moments in the x_k-direction. Moreover, if $\hat{f}(\xi) \neq 0$ for all $\xi \in [-B, B]^d \setminus \{0\}$, then

$$\hat{g}(\xi) \neq 0, \quad \text{for all } \xi \in [-B, B]^d \text{ with } \xi_k \neq 0. \tag{5.9}$$

Proof. For simplicity of exposition, we consider the case $B = 1$ only. Taking the Fourier transform of (5.8) yields

$$\hat{g}(\xi) = \sum_{\ell=0}^{R-1} \binom{R-1}{\ell} (-1)^\ell e^{-2\pi i \ell \xi_k} \hat{f}(\xi) = \left(1 - e^{-2\pi i \xi_k}\right)^{R-1} \cdot \hat{f}(\xi) \tag{5.10}$$

which implies

$$\left(\frac{\partial^\ell}{\partial \xi_k^\ell} \hat{g}\right)_{\xi_k = 0} = 0, \quad \text{for all } \ell \in \{0, \dots, R-1\}.$$

But, by Definition 5.9, this says precisely that g possesses the desired vanishing moments. Statement (5.9) follows by inspection of (5.10). ∎

5.3. Affine Representation Systems. We are now ready to introduce the general family of representation systems announced earlier in the paper as *affine systems*. This class includes all representation systems based on affine transformations of a given "mother function". Special cases of affine systems are wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules, as well as tensor products thereof. The formal definition of affine systems is as follows.

DEFINITION 5.11. *Let $d, r, S \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and let $f \in L^2(\mathbb{R}^d)$ be compactly supported. Let $\delta > 0$, $(c_i^s)_{i=1}^r \subset \mathbb{R}$, for $s = 1, \dots, S$, and let $(d_i)_{i=1}^r \subset \mathbb{R}^d$. Further, let $A_j \in \mathbb{R}^{d \times d}$, $j \in \mathbb{N}$, be full-rank, with the absolute values of the eigenvalues of A_j bounded below by 1. Consider the compactly supported functions*

$$g_s := \sum_{i=1}^r c_i^s f(\cdot - d_i), \quad s = 1, \dots, S.$$

We define the affine system $\mathcal{D} \subset L^2(\Omega)$ *corresponding to $(g_s)_{s=1}^S$ according to*

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

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

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

$$g_s^{j,b} \in \mathcal{D} \implies \|b\|_\infty \leq c_b \|A_j\|_\infty. \tag{5.11}$$

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

$$\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} = (\mathcal{D}_{1,1},\dots,\mathcal{D}_{S,1},\mathcal{D}_{1,2},\dots,\mathcal{D}_{S,2},\dots)\,, \tag{5.12}$$

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

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

$$\sum_{s=1}^{S} |\mathcal{D}_{s,j}| \geq c_{\mathrm{o}}|\det(A_j)|, \text{ for all } j\in\mathbb{N}. \tag{5.13}$$

The next result establishes that all affine systems whose generator functions can be approximated to within arbitrary accuracy by neural networks are (effectively) representable by neural networks.

THEOREM 5.12. *Let $d\in\mathbb{N}$, $\rho:\mathbb{R}\to\mathbb{R}$, $\Omega\subset\mathbb{R}^d$ be bounded, and $\mathcal{D}=(\varphi_i)_{i\in\mathbb{N}}\subset L^2(\Omega)$ an affine system with generator function f. Suppose that there exist constants $L,R\in\mathbb{N}$ such that for all $D,\varepsilon>0$, there is $\Phi_{D,\varepsilon}\in\mathcal{NN}_{L,R,d,\rho}$ with*

$$\|f-\Phi_{D,\varepsilon}\|_{L^2([-D,D]^d)} \leq \varepsilon. \tag{5.14}$$

Then, $\mathcal{D}$ is representable by neural networks with activation function ρ. If, in addition, the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D,ε^{-1}, and if there exist $a>0$ and $c>0$ such that

$$1\geq c\|A_1\|_\infty, \quad \sum_{k=1}^{j-1}|\det(A_k)| \geq c\|A_j\|_\infty^a, \quad \text{for all } j\in\mathbb{N}, j\geq 2, \tag{5.15}$$

then $\mathcal{D}$ is effectively representable by neural networks with activation function ρ.

Proof. Let $(g_s)_{s=1}^S$ be as in Definition 5.11. If $g_s=0$ for all $s\in\{1,\dots,S\}$, then $\mathcal{D}=\varnothing$ and the result is trivial. Hence, we can assume that there exists at least one $s\in\{1,\dots,S\}$ such that $g_s\neq 0$, implying that (5.13) holds.

Pick D such that $\Omega\subset[-D,D]^d$. We first show that (5.14) implies representability of $\mathcal{D}$ by neural networks with activation function ρ. To this end, we need to establish the existence of constants $L,R\in\mathbb{N}$ such that for all $i\in\mathbb{N}$ and all $\eta>0$, there exist $\Phi_{i,\eta}\in\mathcal{NN}_{L,R,d,\rho}$ with

$$\|\varphi_i-\Phi_{i,\eta}\|_{L^2(\Omega)}\leq\eta. \tag{5.16}$$

The elements of $\mathcal{D}$ consist of dilations and translations of f according to

$$\varphi_i = |\det(A_{j_i})|^{\frac{1}{2}}\left(\sum_{k=1}^{r} c_k^{s_i} f(A_{j_i}\cdot - \delta b_i - d_k)\right)_{|\Omega}, \tag{5.17}$$

for some $r\in\mathbb{N}$ independent of i, and $s_i\in\{1,\dots,S\}$, $j_i\in\mathbb{N}$, and $b_i\in\mathbb{Z}^d$. Thus (5.16) follows directly by Propositions 5.7 and 5.8.

It remains to show that the weights of $\Phi_{D,\varepsilon}$ in (5.14) polynomially bounded in D,ε^{-1} implies that $\mathcal{D}$ is effectively representable by neural networks with activation function ρ, which, by Definition 4.1, means that the weights of $\Phi_{i,\eta}$ are polynomially bounded in i,η^{-1}. Propositions 5.7 and 5.8 state that the weights of $\Phi_{i,\eta}$ are polynomially bounded in

$$\|A_{j_i}\|_\infty, D, \|b_i\|_\infty, \sum_{k=1}^{r}|c_k|, \max_{k=1,\dots,r}\|d_k\|_\infty, \eta^{-1}.$$

Thanks to (5.11) we have $\|b_i\|_\infty \in \mathcal{O}(\|A_{j_i}\|_\infty)$. Moreover, the quantities D, $\sum_{k=1}^r |c_k|$, and $\max_{k=1,\dots,r} \|d_k\|_\infty$ do not depend on i. We can thus conclude that the weights of $\Phi_{i,\eta}$ are polynomially bounded in

$$\|A_{j_i}\|_\infty, \eta^{-1}. \tag{5.18}$$

To complete the proof, we need to show that the quantities $\|A_{j_i}\|_\infty$ are polynomially bounded in i. To this end, we first observe that φ_i according to (5.17) satisfies $\varphi_i \in \mathcal{D}_{s_i,j_i}$ for some $s_i \in \{1,\dots,S\}$. Thanks to (5.13) and the canonical ordering (5.12), there exists a constant $c > 0$ such that

$$i \geq c \sum_{k=1}^{j_i-1} |\det(A_k)|,\ j_i \geq 2.$$

We finally appeal to (5.15) to conclude that $\|A_{j_i}\|_\infty$ is polynomially bounded in i for all $j_i \in \mathbb{N}$, which, together with (5.18), establishes the desired result. ∎

We remark that condition (5.15) is very weak; in fact, we are not aware of an affine system in the literature that would violate it.

We now proceed to what is probably the central result of this paper, namely that neural networks provide optimal approximations for all function classes that are optimally approximated by any affine system with generator function that can be approximated to within arbitrary accuracy by neural networks.

THEOREM 5.13. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, $\rho : \mathbb{R} \to \mathbb{R}$, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ be an affine system with generator function f. Assume that there exist $L, R \in \mathbb{N}$ such that for all $D, \varepsilon > 0$, there is $\Phi_{D,\varepsilon} \in \mathcal{NN}_{L,R,d,\rho}$ satisfying $\|f - \Phi_{D,\varepsilon}\|_{L^2([-D,D]^d)} \leq \varepsilon$. Then, for all function classes $\mathcal{C} \subset L^2(\Omega)$, we have*

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

If, in addition, there is a bivariate polynomial $\widetilde{\pi}$ such that the weights of $\Phi_{D,\varepsilon}$ are bounded by $|\widetilde{\pi}(D, \varepsilon^{-1})|$, there exist $a > 0$ and $c > 0$ such that (5.15) holds, and $\mathcal{C}$ is optimally represented by $\mathcal{D}$ (according to Definition 3.3), then for all $\gamma < \gamma^*(\mathcal{C})$, there exist a constant $c > 0$, a polynomial π, and a map

$$\mathbf{Learn} : \left(0, \frac{1}{2}\right) \times L^2(\Omega) \to \mathcal{NN}^\pi_{L,\infty,d,\rho},$$

such that for every $f \in \mathcal{C}$ the weights in $\mathbf{Learn}(\varepsilon, f)$ can be represented by no more than $\lceil c \log_2(\varepsilon^{-1}) \rceil$ bits while $\|f - \mathbf{Learn}(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon$ and $\mathcal{M}(\mathbf{Learn}(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma})$, $\varepsilon \to 0$.

Proof. The proof follows directly by combining Theorem 5.12 with Theorems 4.2 and 4.3. ∎

Theorem 5.13 reveals a remarkable universality and optimality property of neural networks: All function classes that can be optimally represented by an affine system with generator f satisfying (5.14) are also optimally representable by neural networks.

6. α-Shearlets and Cartoon-Like Functions. We next present an explicit pair $(\mathcal{C}, \mathcal{D})$ of function class and representation system satisfying $\gamma^*_{\mathcal{NN}}(\mathcal{C}, \rho) = \gamma^*(\mathcal{C}, \mathcal{D})$. Specifically, we take α-shearlets as representation system $\mathcal{D} \subset L^2(\mathbb{R}^2)$ and α^{-1}-cartoon-like functions as function class $\mathcal{C}$. Cartoon-like functions are piecewise smooth functions with only two pieces. These pieces are separated by a smooth interface. In a sense, they can be understood

as a prototype of a two-dimensional classification function with two homogeneous areas corresponding to two classes. Understanding neural network approximation of this function class is hence relevant to classification tasks in machine learning. We point out that the definition of α-shearlets in this paper differs slightly from that in [27]. Concretely, relative to [27] our definition replaces α^{-1} by α so that α-shearlets are a special case of α-molecules, whereas in [27] α-shearlets are a special case of α^{-1}-molecules. We will need dilation and shearing matrices defined as

$$D_{\alpha,a} := \begin{pmatrix} a & 0 \\ 0 & a^{\alpha} \end{pmatrix}, \qquad J := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}.$$

This leads us to the following definition which is a slightly modified version of the corresponding definition in [56].

DEFINITION 6.1 ([56]). *For $\delta \in \mathbb{R}^+$, $\alpha \in [0,1]$, and $f, g \in L^2(\mathbb{R}^2)$, the* cone-adapted α-shearlet system $\mathcal{SH}_\alpha(f,g,\delta)$ *generated by $f, g \in L^2(\mathbb{R}^2)$ is defined as*

$$\mathcal{SH}_\alpha(f,g,\delta) := \mathcal{SH}^0_\alpha(f,g,\delta) \cup \mathcal{SH}^1_\alpha(f,g,\delta),$$

where

$$\mathcal{SH}^0_\alpha(f,g,\delta) := \big\{f(\cdot - \delta t) : t \in \mathbb{Z}^2\big\},$$

$$\mathcal{SH}^1_\alpha(f,g,\delta) := \Big\{2^{\ell\frac{1+\alpha}{2}} g(S_k D_{\alpha,2^\ell} J^\tau \cdot - \delta t) : \ell \in \mathbb{N}_0,\ |k| \le \lceil 2^{\ell(1-\alpha)} \rceil,\ t \in \mathbb{Z}^2,\ k \in \mathbb{Z},\ \tau \in \{0,1\}\Big\}.$$

Our interest in α-shearlets stems from the fact that they optimally represent α^{-1}-cartoon-like functions in the sense of Definition 3.3.

DEFINITION 6.2. *Let $\beta \in [1,2)$ and $\nu > 0$. Define*

$$\mathcal{E}^\beta(\mathbb{R}^2;\nu) = \{f \in L^2(\mathbb{R}^2) : f = f_0 + \chi_B f_1\},$$

*where f_0, $f_1 \in C^\beta(\mathbb{R}^2)$, $\operatorname{supp} f_0$, $\operatorname{supp} f_1 \subset (0,1)^2$, $B \subset [0,1]^2$, χ_B denotes the characteristic function of B, $\partial B \in C^\beta$, and $\|f_1\|_{C^\beta}, \|f_2\|_{C^\beta}$, $\|\partial B\|_{C^\beta} < \nu$. The elements of $\mathcal{E}^\beta(\mathbb{R}^2;\nu)$ are called β-*cartoon-like functions.

This function class was originally introduced in [20] as a model class for functions governed by curvilinear discontinuities of prescribed regularity. In this sense, β-cartoon-like functions provide a convenient model for images governed by edges or for the solutions of transport equations which often exhibit curvilinear singularities.

The optimal exponent $\gamma^*(\mathcal{E}^\beta(\mathbb{R}^2;\nu))$ was found in [20, 28]:

THEOREM 6.3. *For $\beta \in [1,2]$ and $\nu > 0$, we have*

$$\gamma^*(\mathcal{E}^\beta(\mathbb{R}^2;\nu)) = \frac{\beta}{2}.$$

Proof. The proof of [20, Theorem 2] demonstrates that a general function class $\mathcal{C}$ has optimal exponent $\gamma^*(\mathcal{C}) = (2-p)/2p$ if $\mathcal{C}$ contains a copy of ℓ^p_0. The result now follows, since by [28], the function class $\mathcal{E}^\beta(\mathbb{R}^2;\nu)$ does, indeed, contain a copy of ℓ^p_0 for $p = 2/(\beta+1)$. □

Using Proposition 3.6, this result allows us to conclude that neural networks achieving uniform approximation error ε over the class $\mathcal{C}$ of cartoon-like functions, with weights represented by no more than $\lceil c \log_2(\varepsilon^{-1}) \rceil$ bits, for some constant $c > 0$, yield an effective best

M-edge approximation rate of at most $\beta/2$. Theorem 6.8 below demonstrates achievability for $\beta = 1/\alpha$, with $\alpha \in [1/2, 1]$.

The following theorem states that α-shearlets yield optimal best M-term approximation rates for α^{-1}-cartoon-like functions.

THEOREM 6.4 ([56], Theorem 6.3 and Remark 6.4). *Let $\alpha \in [1/2, 1]$, $\nu > 0$, $f \in C^{12}(\mathbb{R}^2)$, $g \in C^{32}(\mathbb{R}^2)$, both compactly supported and such that*

(i) $\widehat{f(\xi)} \neq 0$, for all $|\xi| \leq 1$,

(ii) $\widehat{g(\xi)} \neq 0$, for all $\xi = (\xi_1, \xi_2)^T \in \mathbb{R}^2$ such that $1/3 \leq |\xi_1| \leq 3$ and $|\xi_2| \leq |\xi_1|$,

(iii) g has at least seven vanishing moments in the x_1-direction, i.e.,

$$\int_{\mathbb{R}} x_1^\ell g(x_1, x_2) dx_1 = 0, \quad \text{for all } x_2 \in \mathbb{R},\ \ell \in \{0, \dots, 6\}.$$

Then, there exists $\delta^* > 0$ such that for all $\delta < \delta^*$, the function class $\mathcal{E}^{1/\alpha}(\mathbb{R}^2; \nu)$ is optimally represented by $\mathcal{SH}_\alpha(f, g, \delta)$.

Remark 6.5. The assumptions on the smoothness and the number of vanishing moments of f and g in Theorem 6.4 follow from [56, Eq. 4.9] with $s_1 = 3/2, s_0 = 0, p_0 = q_0 = 2/3$, and $|\beta| \leq 4$. While these particular choices allow the statement of the theorem to be independent of α, it is possible to weaken the assumptions, if a fixed α is considered. For example, for $\alpha = 1/2$ the smoothness assumptions on f and g reduce to $f \in C^{11}, g \in C^{28}$.

As our approximation results for neural networks pertain to bounded domains, we require a definition of cartoon-like functions on bounded domains.

DEFINITION 6.6. *Let $(0, 1)^2 \subset \Omega \subset \mathbb{R}^2$, $\alpha \in [1/2, 1]$, and $\nu > 0$. We define the set of α^{-1}-cartoon-like functions on Ω by*

$$\mathcal{E}^{\frac{1}{\alpha}}(\Omega; \nu) := \left\{ f_{|\Omega} : f \in \mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2; \nu) \right\}.$$

*Additionally, for $\delta > 0$, $f, g \in L^2(\mathbb{R}^2)$, we define an α-*shearlet system on Ω *according to*

$$\mathcal{SH}_\alpha(f, g, \delta; \Omega) := \left\{ \phi_{|\Omega} : \phi \in \mathcal{SH}_\alpha(f, g, \delta) \right\}.$$

Remark 6.7. It is straightforward to check that if $\mathcal{E}^{1/\alpha}(\mathbb{R}^2; \nu)$ is optimally represented by $\mathcal{SH}_\alpha(f, g, \delta)$, then $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ is optimally represented by $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$.

We proceed to the main statement of this section.

THEOREM 6.8. *Suppose that $(0, 1)^2 \subset \Omega \subset \mathbb{R}^2$ is bounded and $\rho : \mathbb{R} \to \mathbb{R}$ is either strongly sigmoidal of order $k \geq 2$ (see Definition 5.1) or an admissible smooth activation function (see Definition 5.2). Then, for every $\alpha \in [1/2, 1]$, the function class $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ is optimally representable by a neural network with activation function ρ.*

Proof. Let $\alpha \in [1/2, 1]$ and $\nu > 0$. We first consider the case of ρ strongly sigmoidal of order $k \geq 2$. Since the two-dimensional cardinal B-spline of order 34, denoted by N_{34}^2, is 32 times continuously differentiable and $\widehat{N_{34}^2}(0) \neq 0$ by construction, we conclude that there exists $c > 0$ such that $f := N_{34}^2(c\cdot)$ satisfies $f \in C^{32}(\mathbb{R}^2)$ and $\hat{f} \neq 0$ for all $\xi \in [-3, 3]^2$. Application of Lemma 5.10 then yields the existence of $(c_i)_{i=1}^7 \subset \mathbb{R}$, $(d_i)_{i=1}^7 \subset \mathbb{R}^2$ such that $g := \sum_{i=1}^7 c_i f(\cdot - d_i)$ is compactly supported, has seven vanishing moments in the x_1-direction, and $\hat{g}(\xi) \neq 0$ for all $\xi \in [-3, 3]^2$ such that $\xi_1 \neq 0$. Then, by Theorem 6.4 and Remark 6.7, there exists $\delta > 0$ such that $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$ is optimal for $\mathcal{E}^{1/\alpha}(\Omega; \nu)$. We define

$$\{A_j : j \in \mathbb{N}\} := \left\{ S_k D_{\alpha, 2^\ell} J^\tau : \ell \in \mathbb{N}_0, |k| \leq \lceil 2^{\ell(1-\alpha)} \rceil, \tau \in \{0, 1\} \right\},$$

where we order $(A_j)_{j\in\mathbb{N}}$ such that $|\det(A_j)| \leq |\det(A_{j+1})|$, for all $j \in \mathbb{N}$. This construction implies that the α-shearlet system $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$ is an affine system with generator function f. Thanks to Theorem 5.4, there exist $L, R \in \mathbb{N}$ such that for all $D, \varepsilon > 0$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{L,R,d,\rho}$ with

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

Moreover, the weights of $\Phi_{D,\varepsilon}$ are polynomially bounded in D, ε^{-1}. It is not difficult to verify that (5.15) holds, and hence Theorem 5.12 yields that $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$ is effectively representable by neural networks with activation function ρ. Finally, since $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ is optimally representable by $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$, we conclude with Theorem 4.3 that $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ is optimally representable by neural networks with activation function ρ.

It remains to establish the statement for admissible smooth ρ. In this case, by Theorem 5.6 there exist $M \in \mathbb{N}$ and a neural network in $\mathcal{NN}_{3,M,d,\rho}$ which realizes a compactly supported $f \in C^\infty(\mathbb{R})$ satisfying $\hat{f}(\xi) \neq 0$, for all $\xi \in [-3, 3]^2$. Lemma 5.10 applied to this f then yields a function g that can be realized by a neural network in $\mathcal{NN}_{3,M',d,\rho}$, for some $M' \in \mathbb{N}$, has seven vanishing moments in the x_1-direction, is compactly supported, and satisfies $g \in C^\infty(\mathbb{R})$, and $\hat{g}(\xi) \neq 0$, for all $\xi \in [-3, 3]^2$ such that $\xi_1 \neq 0$. By Theorem 6.4 and Remark 6.7, there exists $\delta > 0$ such that $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ is optimally representable by $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$. Note that $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$ is an affine system with generator function f. Since f can be realized with zero error by a neural network, Theorem 5.12 yields that $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$ is effectively representable by neural networks with admissible smooth activation function ρ. Optimality of $\mathcal{SH}_\alpha(f, g, \delta; \Omega)$ for $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ implies, with Theorem 4.3, that $\mathcal{E}^{1/\alpha}(\Omega; \nu)$ is optimally representable by neural networks with admissible smooth activation function ρ. ∎

Remark 6.9. Theorem 6.4 requires the generators of the shearlet system guaranteeing optimal representability of $\mathcal{E}^{1/\alpha}(\Omega; \nu)$, for $1/2 \leq \alpha \leq 1, \nu > 0$, $\Omega \subset \mathbb{R}^2$ to be very smooth. On the other hand, Theorem 6.8 demonstrates that optimally-approximating neural networks are not required to be particularly smooth. Indeed, Theorem 6.8 holds for networks with differentiable but not necessarily twice differentiable activation function. As the proof of Theorem 6.8 reveals, such weak assumptions suffice thanks to Theorem 5.4, which demonstrates that it is possible to approximate arbitrarily smooth B-splines (in L^2-norm) to within error ε by neural networks with a number of weights that does not depend on ε as long as the activation function is strongly sigmoidal.

Remark 6.10. We observe from the proof of Theorem 6.8 that the depth of the networks required to achieve optimal approximation depends on the activation function only. Indeed, for an admissible smooth activation function, inspection of Theorem 5.6 reveals that networks with three layers can produce optimal approximations in Theorem 6.8. On the other hand, if a sigmoidal activation function is employed, Theorem 5.4 shows that the construction in Theorem 6.8 requires a certain minimum depth depending on the order of sigmoidality.

7. Generalization to Manifolds. Frequently, a function f to be approximated by a neural network models phenomena on (possibly low-dimensional) immersed submanifolds $\Gamma \subset \mathbb{R}^d$ of dimension $m < d$. We next briefly outline how our main results can be extended to this situation. Since analogous results, for the case of wavelets as representation systems, appear already in [53], we will allow ourselves to be somewhat informal.

Suppose that $f : \Gamma \to \mathbb{R}$ is compactly supported. Let $(U_i)_{i\in\mathbb{N}} \subset \Gamma$ be an open cover of Γ such that for each $i \in \mathbb{N}$ the manifold patch U_i can be parametrized as the graph of a function over a subset of the Euclidean coordinates, i.e., there exist coordinates $x_{d_1}, \ldots, x_{d_m}$, open

sets $V_i \subset \mathbb{R}^m$, and smooth mappings

$$\gamma_\ell : \mathbb{R}^m \to \mathbb{R}, \quad \ell \in \{1, \dots, d\} \setminus \{d_1, \dots, d_m\}$$

such that

$$U_i = \{\Xi_i(x_{d_1}, \dots, x_{d_m}) := (\gamma_1(x_{d_1}, \dots, x_{d_m}), \dots, x_{d_1}, \dots, \gamma_d(x_{d_1}, \dots, x_{d_m})) : (x_{d_1}, \dots, x_{d_m}) \in V_i\}.$$

Take a smooth partition of unity $(h_i)_{i\in\mathbb{N}}$, where $h_i : \Gamma \to \mathbb{R}$ is smooth with $\operatorname{supp}(h_i) \subset \overline{U_i}$ and $\sum_{i\in\mathbb{N}} h_i = 1$. Define the localization of f to U_i by $f_i := f h_i$ such that

$$f = \sum_{i\in\mathbb{N}} f_i. \tag{7.1}$$

Every $f_i : U_i \to \mathbb{R}$ can be reparametrized to

$$\tilde{f}_i : \begin{cases} \mathbb{R}^m & \to & \mathbb{R} \\ (x_{d_1}, \dots, x_{d_m}) & \mapsto & f_i \circ \Xi_i(x_{d_1}, \dots, x_{d_m}). \end{cases}$$

Suppose that there exist $L, M \in \mathbb{N}$ and neural networks $\tilde{\Phi}_i \in \mathcal{NN}_{L,M,m,\rho}$ such that

$$\|\tilde{f}_i - \tilde{\Phi}_i\|_{L^2(V_i)} \leq \varepsilon. \tag{7.2}$$

Then, we can construct a neural network $\Phi_i \in \mathcal{NN}_{L,M+md,d,\rho}$ according to

$$\Phi_i(x) := \widetilde{\Phi}_i(P_i x),$$

where P_i denotes the orthogonal projection of x onto the coordinates $(x_{d_1}, \dots, x_{d_m})$. Since P_i is linear, Φ_i is a neural network. Moreover, since P_i is the inverse of the diffeomorphism Ξ_i, we get

$$\|\Phi_i - f_i\|_{L^2(U_i)} \leq C\varepsilon,$$

with $C > 0$ depending on the curvature of $\Gamma|_{U_i}$ only. Now we may build a neural network Φ by setting $\Phi := \sum_{i\in\mathbb{N}} \Phi_i$. Combining (7.2) with the observation that, owing to the compact support of f, only a finite number of summands appears in the definition of f, we have constructed a neural network Φ which approximates f on Γ. In summary, we observe the following.

Whenever a function class $\mathcal{C}$ is invariant with respect to diffeomorphisms (in our construction the functions Ξ_i) and multiplication by smooth functions (in our construction the functions h_i), then approximation results on $\mathbb{R}^m$ can be lifted to approximation results on m-dimensional submanifolds $\Gamma \subset \mathbb{R}^d$.

Such invariances are, in particular, satisfied for all function classes characterized by a certain smoothness behavior, for example, the class of cartoon-functions as studied in Section 6.

8. Numerical Results. Our theoretical results show that neural networks realizing uniform approximation error ε over a function class $\mathcal{C} \subset L^2(\mathbb{R}^d)$, $d \in \mathbb{N}$, must obey a fundamental lower bound on the growth rate (as $\varepsilon \to 0$) of the number of edges of nonzero weight. One of the most widely used learning algorithms is stochastic gradient descent with the gradient computed via backpropagation [52]. The purpose of this section is to investigate how this algorithm fares relative to our lower bound.

Interestingly, our numerical experiments below indicate that for a fixed, sparsely connected, network topology inspired by the construction of bump functions according to (5.1) and (5.2), and with the ReLU as activation function, the stochastic gradient descent algorithm generates neural networks that achieve M-edge approximation rates quite close to the fundamental limit. The network topology we prescribe is depicted in Figure 3. The rationale

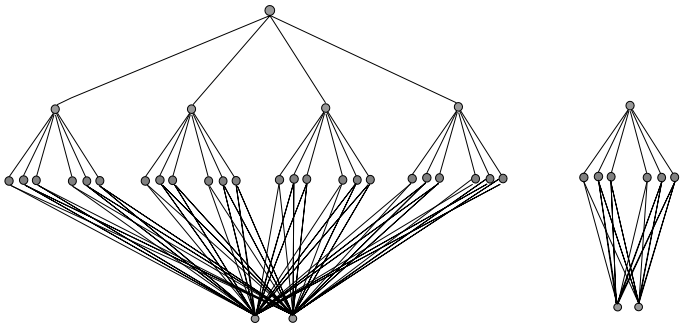


FIG. 3. *Left: Topology of the neural network trained using stochastic gradient descent. The network consists of a weighted sum of four subnetworks. Right: A single subnetwork.*

for choosing this topology is as follows. As mentioned before, admissible smooth activation functions consist of smooth functions which equal a ReLU outside a compact interval. For this class of activation functions, the associated α-shearlet generators were constructed from a function g as specified in (5.2). Choosing $p_1 = p_2 = 1$ and $p_3 = 2$ in (5.1) yields hat functions t. This construction implies that six nodes are required in the first layer in each subnetwork. In Figure 3, we see four network realizations of g in parallel. The output layer realizes a linear combination of the subnetworks. We now train the network using the stochastic gradient descent algorithm. Following (5.2) the weights of the second layer remain fixed, and the weights in the first and third layers only are trained. Training is performed for two different functions, where one is a function with a line singularity (Figure 4(a)), and the other one is a cartoon-like function (Figure 5(a)). Specifically, we train the network by drawing samples (x_1, x_2) from an equispaced grid in $[-1, 1]^2$. The resulting error is then backpropagated through the network. We repeat this procedure for different network sizes, i.e., for different numbers of subnetworks. We start by discussing the results for the function with a line singularity depicted in Figure 4(a). The approximation error corresponding to the trained neural network is shown in Figure 4(b). The faster than linear decay of the approximation error in the semi-logarithmic scale indicates faster than exponential decay with respect to the number of edges. This is consistent with the best M-term approximation rate that ridgelets yield for piecewise constant functions with line singularities, see [5].

It is interesting to observe that the trained subnetworks yield α-molecules for $\alpha = 0$ (see Figures 4(c)-(e)). These functions are constant along one direction and vary along another, hence can be considered part of a ridgelet system, which is, in fact, an optimally sparsifying representation system for line singularities. Moreover, the orientation of the three learned ridge functions matches that of the original function.

In the second experiment, we draw samples from the function depicted in Figure 5(a) below, which exhibits a curvilinear singularity. Figures 5(c)-(e) show that the corresponding trained subnetworks resemble anisotropic molecules with different scales and of different orientations. We report, without showing the results, that the decay rate of the corresponding approximation error obtained when simply training with different network sizes did not come close to the rate of M^{-1} predicted by our theory. However, with a slight adaptation one ob-

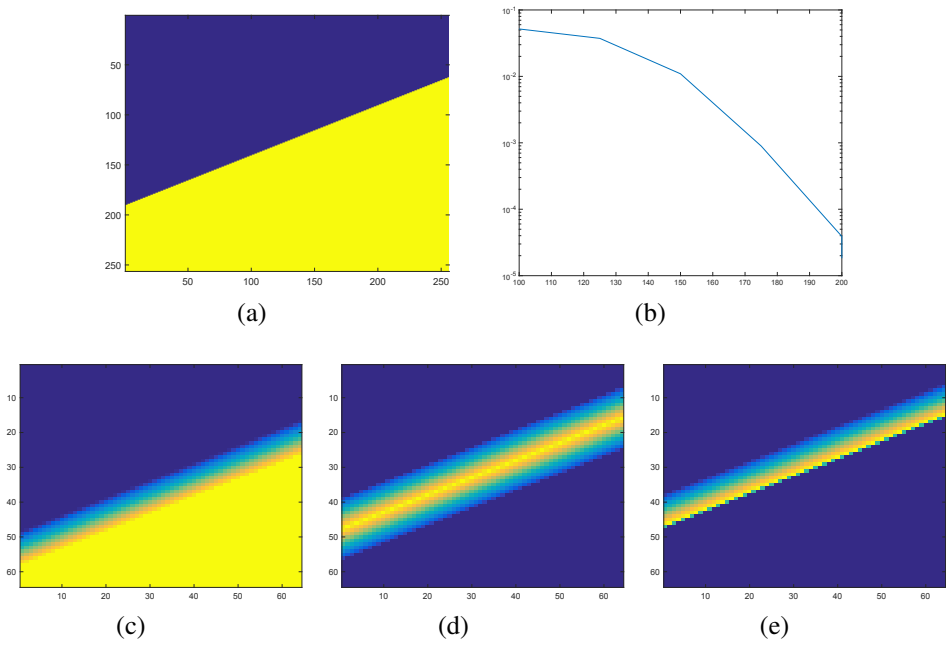


FIG. 4. *(a): Function with a line singularity. (b): Approximation error (vertical axis) as a function of the number of edges (horizontal axis). (c)-(e): The functions obtained by restricting to the subnetworks with the largest weights in modulus in the final layer.*

tains the result of Figure 5(b), which demonstrates a decay of roughly M^{-1}. The specifics of this adaptation are as follows: We first train a large network with ~ 10000 edges, again by stochastic gradient descent. Then, the weights in the last layer are optimized using the Lasso [55] to obtain a sparse weight vector c^*. We then pick the M largest coefficients of c^* and compute the corresponding weighted sum of the associated subnetworks. The resulting approximation error is shown in Figure 5(b). Finally, we investigate whether the approximation characteristics delivered by this procedure are similar to what would be obtained by best M-term approximation with standard shearlet systems. Recall that shearlet elements at high scales tend to cluster around singularities [31, 37]. Figures 5(g)-(i) depict the corresponding results. Specifically, Figure 5(g) shows the weighted sum of those subnetworks that have the largest support. In Figure 5(h), we show weighted sums of subnetworks with medium-sized support, and in Figure 5(i) we sum up only the subnetworks with the smallest supports. We observe that, indeed, subnetworks of large support approximate the smooth part of the underlying function, whereas the subnetworks associated with small supports resolve the jump singularity.

Acknowledgments. The authors would like to thank J. Bruna, E. Candès, M. Genzel, S. Güntürk, Y. LeCun, K.-R. Müller, H. Rauhut, and F. Voigtländer for interesting discussions, and D. Perekrestenko and R. Gül for very detailed and insightful comments on the manuscript. G. K. and P. P. are grateful to the Faculty of Mathematics at the University of Vienna for the hospitality and support during their visits. Moreover, G. K. thanks the Department of Mathematics at Stanford University whose support allowed for completion of a portion of this work. G. K. acknowledges partial support by the Einstein Foundation Berlin, the Einstein Center for Mathematics Berlin (ECMath), the European Commission-Project DEDALE (contract no. 665044) within the H2020 Framework Program, DFG Grant KU

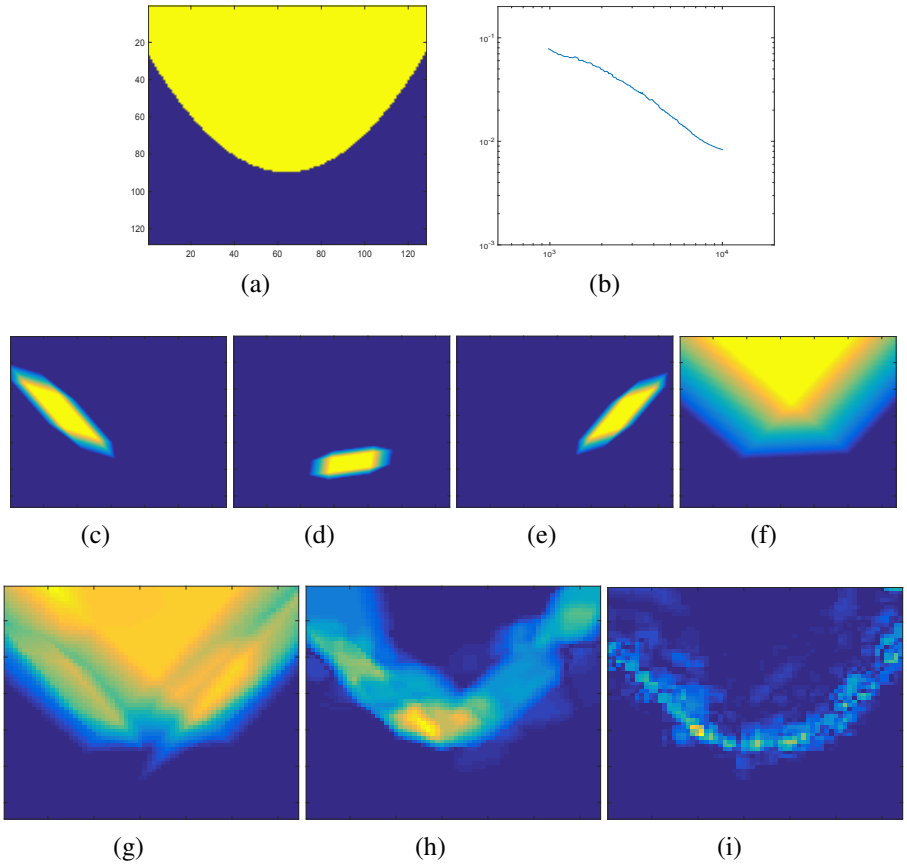


FIG. 5. *(a): Function with curvilinear singularity to be approximated by the neural network. (b): Approximation error (vertical axis) as a function of the number of edges (horizontal axis). (c)-(f): Shearlet-like subnetworks. (g): Reconstruction using only the 10 subnetworks whose associated functions have the largest supports. (h): Reconstruction using only subnetworks whose associated functions have medium-sized support. (i): Reconstruction using only subnetworks with associated functions of very small support.*

1446/18, DFG-SPP 1798 Grants KU 1446/21 and KU 1446/23, and by the DFG Research Center MATHEON "Mathematics for Key Technologies". G. K. and P. P acknowledge support by the DFG Collaborative Research Center TRR 109 "Discretization in Geometry and Dynamics".

REFERENCES

[1] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.
[2] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, 14(1):115–133, 1994.
[3] J. Berner, P. Grohs, and A. Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *arXiv preprint arXiv:1809.03062*, 2018.
[4] E. J. Candès. *Ridgelets: Theory and Applications*. Ph.D thesis, Stanford University, 1998.
[5] E. J. Candès. Ridgelets and the representation of mutilated Sobolev functions. *SIAM J. Math. Anal.*, 33(2):347–368, 2001.
[6] 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.*, 57:219–266, 2002.

[7] C. K. Chui, X. Li, and H. N. Mhaskar. Neural networks for localized approximation. *Math. Comp.*, 63(208):607–623, 1994.
[8] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore. Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, 11(2):192–226, 2001.
[9] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on Learning Theory*, pages 698–728, 2016.
[10] N. Cohen and A. Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning*, pages 955–963, 2016.
[11] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.
[12] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.
[13] 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*, 529(7587):484–489, 2016.
[14] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.
[15] L. Demanet and L. Ying. Wave atoms and sparsity of oscillatory patterns. *Appl. Comput. Harmon. Anal.*, 23(3):368–387, 2007.
[16] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.
[17] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer, 1993.
[18] R. A. DeVore, K. Oskolkov, and P. Petrushev. Approximation by feed-forward neural networks. *Ann. Numer. Math.*, 4:261–287, 1996.
[19] D. L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.*, 1(1):100 – 115, 1993.
[20] D. L. Donoho. Sparse components of images and optimal atomic decompositions. *Constr. Approx.*, 17(3):353–382, 2001.
[21] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 907–940, 2016.
[22] S. Ellacott. Aspects of the numerical analysis of neural networks. *Acta Numer.*, 3:145–202, 1994.
[23] K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.
[24] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.
[25] K. Gröchenig. *Foundations of time-frequency analysis*. Springer Science & Business Media, 2013.
[26] P. Grohs. Optimally sparse data representations. In *Harmonic and Applied Analysis*, pages 199–248. Springer, 2015.
[27] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer. α-molecules. *Appl. Comput. Harmon. Anal.*, 41(1):297–336, 2016.
[28] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer. Cartoon approximation with α-curvelets. *J. Fourier Anal. Appl.*, 22(6):1235–1293, 2016.
[29] P. Grohs and G. Kutyniok. Parabolic molecules. *Found. Comput. Math.*, 14:299–337, 2014.
[30] P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Bölcskei. Deep neural network approximation theory. *IEEE Trans. Inf. Theory, submitted*, 2018.
[31] K. Guo, G. Kutyniok, and D. Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines (Athens, GA, 2005)*, pages 189–201. Nashboro Press, Nashville, TN, 2006.
[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
[33] 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.*, 29(6):82–97, 2012.
[34] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.
[35] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
[37] G. Kutyniok and W.-Q. Lim. Compactly supported shearlets are optimally sparse. *Journal of Approximation Theory*, 163(11):1564 – 1589, 2011.
[38] Y. LeCun. *Modèles connexionnistes de l'apprentissage*. PhD thesis, These de Doctorat, Université Paris 6,

1987.
[39] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
[40] V. Maiorov and A. Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1):81–91, 1999.
[41] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943.
[42] H. Mhaskar and C. Micchelli. Degree of approximation by neural and translation networks with a single hidden layer. *Adv. Appl. Math.*, 16(2):151–183, 1995.
[43] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80, Feb 1993.
[44] H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.*, 8(1):164–177, 1996.
[45] H. N. Mhaskar and C. Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Adv. Appl. Math.*, 13:350–373, 1992.
[46] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.
[47] T. Nguyen-Thien and T. Tran-Cong. Approximation of functions and their derivatives: A neural network implementation with applications. *Appl. Math. Model.*, 23(9):687–704, 1999.
[48] E. Ott. *Chaos in dynamical systems*. Cambridge University Press, Cambridge, second edition, 2002.
[49] P. Petersen and F. Voigtlaender. Equivalence of approximation by convolutional neural networks and fully–connected networks. arXiv:1809.00973.
[50] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296 – 330, 2018.
[51] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numer.*, 8:143–195, 1999.
[52] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct. 1986.
[53] U. Shaham, A. Cloninger, and R. R. Coifman. Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal.*, 44(3):537–557, May 2018.
[54] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.
[55] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288, 1996.
[56] F. Voigtlaender and A. Pein. Analysis sparsity versus synthesis sparsity for α-shearlets. arXiv:1702.03559.
[57] D. Yarotsky. Universal approximations of invariant maps by neural networks. *arXiv:1804.10306v1*.
[58] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.
[59] D.-X. Zhou. Universality of deep convolutional neural networks. arXiv:1805.10769.