Memory-optimal neural network approximation

Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, Philipp Petersen

Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, Philipp Petersen, "Memory-optimal neural network approximation," Proc. SPIE 10394, Wavelets and Sparsity XVII, 103940Q (24 August 2017); doi: 10.1117/12.2272490

SPIE.

Event: SPIE Optical Engineering + Applications, 2017, San Diego, California, United States

Memory-Optimal Neural Network Approximation

Helmut Bölcskei[a], Philipp Grohs[b], Gitta Kutyniok[c], and Philipp Petersen[c]

[a]Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland;
[b]Faculty of Mathematics, University of Vienna, 1090 Vienna, Austria;
[c]Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany

ABSTRACT

We summarize the main results of a recent theory—developed by the authors—establishing fundamental lower bounds on the connectivity and memory requirements of deep neural networks as a function of the complexity of the function class to be approximated by the network. These bounds are shown to be achievable. 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, shearlets, ridgelets, α-shearlets, and more generally α-molecules. This result elucidates a remarkable universality property of deep neural networks and shows that they achieve the optimum approximation properties of all affine systems combined. Finally, we present numerical experiments demonstrating that the standard stochastic gradient descent algorithm generates deep neural networks which provide close-to-optimal approximation rates at minimal connectivity. Moreover, stochastic gradient descent is found to actually learn approximations that are sparse in the representation system optimally sparsifying the function class the network is trained on.

Keywords: Deep neural networks, function approximation, optimal sparse approximation, connectivity, shearlets.

This paper is a summary of https://arxiv.org/abs/1705.01714 with some minor additions and embellishments.

1. INTRODUCTION

Neural networks arose from the seminal work by McCulloch and Pitts [1] 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 (potentially learned) affine linear functions and relatively simple non-linearities.

Despite significant theoretical progress in the 1990s [2, 3], the area has seen practical advances only during the past decade, triggered by the drastic improvements in computing power, and, in particular, the availability of vast amounts of training data. Specifically, deep neural networks, i.e., networks with large numbers of layers are now state-of-the-art technology for a wide variety of practical applications, such as image classification [4], speech recognition [5], or game intelligence [6], to name a few. For an in-depth overview we refer to the survey paper by LeCun, Bengio, and Hinton [7] and the recent book [8].

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 [9]. In the former case, the primary goal is to approximate an unknown classification function based on a given set of corresponding input-output value pairs. This is typically accomplished by learning the network's weights through, e.g., the

Further author information: (Send correspondence to G. Kutyniok)
H.B.: E-mail: boelcskei@nari.ee.ethz.ch, Telephone: + 41 44 6323433
P.G.: E-mail: philipp.grohs@univie.ac.at, Telephone: +43 14 27755741
G.K.: E-mail: kutyniok@math.tu-berlin.de, Telephone: +49 30 31425758
P.P.: E-mail: petersen@math.tu-berlin.de, Telephone: +49 30 31425748

Wavelets and Sparsity XVII, edited by Yue M. Lu, Dimitri Van De Ville, Manos Papadakis, Proc. of SPIE
Vol. 10394, 103940Q · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2272490

standard gradient descent via backpropagation algorithm [10]. In a classification task with, say, two classes, 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.

Deep neural networks employed in practice often consist of hundreds of layers and may depend on billions of parameters, see for example the work [11] on image classification. Training and operation of networks of this scale entail formidable computational challenges which often still present a bottleneck. As a case in point, we mention speech recognition on a smart phone such as, e.g., Apple's SIRI-system, which operates in the cloud. Android's speech recognition system has meanwhile released an offline version that is based on a neural network with sparse connectivity, meaning that the number of edges with nonzero weights is small.

The desire to reduce the complexity of network training and operation naturally leads to the question of function approximation through neural networks with sparse connectivity. In addition, the network's memory requirements in terms of the number of bits needed for its storage are of concern in practice.

The purpose of this paper is to summarize recent results by the authors [12] (with minor additions and embellishments) which characterize the approximation-theoretic properties of deep neural networks with sparse connectivity and under memory constraints. Specifically, defining the complexity of a signal class $\mathcal{C}$ as the number of bits needed to describe any element in $\mathcal{C}$ to within a prescribed accuracy, we shall ask the following question:

> *Given a signal class $\mathcal{C}$, how does the complexity of a neural network that approximates every function in $\mathcal{C}$ to within a prescribed accuracy depend on the complexity of $\mathcal{C}$?*

Interpreting the network as an encoder in Donoho's min-max rate distortion theory [13], we establish fundamental lower bounds on connectivity and memory requirements for a network to guarantee uniform approximation rates for a given signal class $\mathcal{C}$. Moreover, we demonstrate that these bounds can be attained for a broad family of signal classes, namely those that can be optimally approximated by a general class of representation systems—so-called affine systems. Affine systems include wavelets, shearlets, ridgelets, α-shearlets, and more generally α-molecules. This result reveals an interesting universality property of deep neural networks; they achieve the optimum approximation properties of all affine systems combined. The technique we develop to proof this statement is interesting in its own right as it establishes a general framework for transferring statements on function approximation through representation systems to analogous results for approximation by neural networks.

1.1 Deep Neural Networks

While various network architectures exist in the literature, we focus on the following setup:

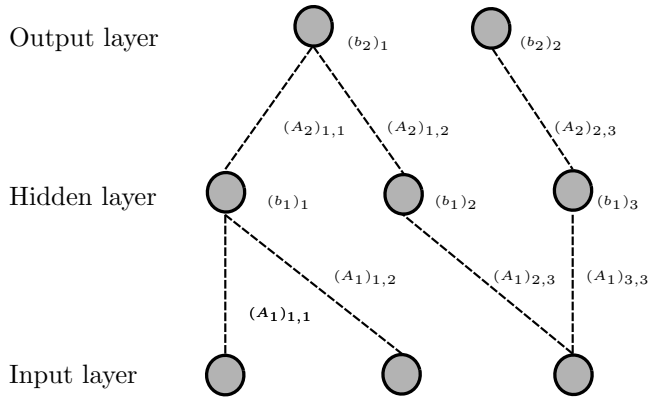
DEFINITION 1.1. *Let $L, d, N_1, \dots, N_L \in \mathbb{N}$. A map $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by*

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

is called a neural network. *It is composed of affine linear maps $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $1 \leq \ell \leq L$, where $N_0 := d$, and non-linear functions—often referred to as* rectifiers*—ρ acting component-wise. Here, d is the* dimension of the input layer*, L denotes the* number of layers*, $N_1, \dots, N_{L-1}$ stands for the* dimensions of the $L-1$ hidden layers*, and N_L is the* dimension of the output layer.

The term "network" arises from the interpretation of the mapping Φ as a weighted acyclic directed graph with nodes arranged in L hierarchical layers and edges only between adjacent layers. In fact, the affine linear map W_ℓ is defined by a matrix $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and an affine part $b_\ell \in \mathbb{R}^{N_\ell}$ via $W_\ell(x) = A_\ell x + b_\ell$. $(A_\ell)_{i,j}$ is 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 neuron in the ℓ-th layer. This assignment is depicted in Figure 1. We refer to the nodes of the graph as neurons and note that the total number of neurons is given by $N := d + \sum_{j=1}^{L} N_j$.

The real numbers $(A_\ell)_{i,j}$ and $(b_\ell)_i$ are said to be the network's weights; the total number of non-zero edge weights, denoted by M, is the network's connectivity. If M is small relative to the number of connections possible



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

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

Figure 1. Assignment of the weights $(A_\ell)_{i,j}$ and $(b_\ell)_i$ to the neurons and edges. The network has sparse connectivity.

(i.e., the number of edges in the graph that is fully connected between adjacent layers), we say that the network is sparsely connected.

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 prediction problems described above. All our results can readily be generalized to $N_L > 1$.

We denote the class of networks $\Phi : \mathbb{R}^d \to \mathbb{R}$ with no more than L layers, no more than M non-zero edge weights, and rectifier ρ by $\mathcal{NN}_{L,M,d,\rho}$. Moreover, we let

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

Now, given a function $f : \mathbb{R}^d \to \mathbb{R}$, we can ask how well a neural network $\Phi \in \mathcal{NN}_{L,M,d,\rho}$ can approximate f. Clearly, this depends on the algorithm chosen to learn the network's weights. But one can also take the following vantage point: The best possible approximation of f by $\mathcal{NN}_{L,M,d,\rho}$ provides a fundamental lower bound on the approximation error *independently of the learning algorithm*. We shall be particularly interested in the dependence of this bound on the connectivity M, and the number of bits required to encode the network. Clearly, smaller M entails lower computational complexity in terms of evaluating (1).

1.2 Quantifying Approximation Quality

We next briefly review a widely used approach for characterizing the approximation quality of functions given restricting conditions on the approximant, in our case the neural network, with the restriction imposed by its connectivity.

Fix $\Omega \subset \mathbb{R}^d$. A common approach to function approximation is to consider a class of functions $\mathcal{C} \subseteq L^2(\Omega)$, termed *signal class* and a corresponding complete system $\mathcal{D} := (\varphi_i)_{i\in I} \subseteq L^2(\Omega)$, termed *(representation) system*, or *dictionary*, with the restriction on the approximant imposed by a limit on the number of elements in $\mathcal{D}$ allowed to participate in the approximation. One then studies the *error of best M-term approximation* of $f \in \mathcal{C}$:

DEFINITION 1.2. *[14] Given $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, a signal class $\mathcal{C} \subseteq L^2(\Omega)$, and a representation system $\mathcal{D}$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

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

We call $\Gamma_M^{\mathcal{D}}(f)$ the best M-term approximation error of f with respect to $\mathcal{D}$. *Every $f_M = \sum_{i\in I_M} c_i\varphi_i$ attaining the infimum is referred to as a* best M-term approximation *of f in the system $\mathcal{D}$. The supremum of $\gamma > 0$ such that there exists $C > 0$ with*

$$\sup_{f\in\mathcal{C}} \Gamma_M^{\mathcal{D}}(f) \le CM^{-\gamma}, \qquad \text{for all } M \in \mathbb{N},$$

determines the optimal M-term approximation rate of $\mathcal{C}$ in the system $\mathcal{D}$*, and will henceforth be referred to as* $\gamma^*(\mathcal{C},\mathcal{D})$.

A wealth of structured representation systems $\mathcal{D}$ is provided by the area of applied harmonic analysis, starting with wavelets [15], followed by ridgelets [16], curvelets [17], shearlets [18], parabolic molecules [19], and most generally α-molecules [20], which include all previously named systems as special cases. Other examples include tensor product wavelets [21], Gabor frames [22], and wave atoms [23].

For α-shearlet systems $\mathcal{D}$, optimal sparse approximation properties have been completely characterized for the signal class $\mathcal{C}$ of so-called α-cartoon-like functions $\mathcal{E}^{\frac{1}{\alpha}}([0,1]^2)$, these are piecewise $C^{1/\alpha}(\mathbb{R}^2)$ functions on the unit square with a $C^{1/\alpha}$ discontinuity curve and $\alpha \in [\frac{1}{2},1]$. Under weak technical conditions on the α-shearlet system, it was shown in [24,25] that $\gamma^*(\mathcal{C},\mathcal{D}) = \frac{1}{2\alpha}$ for $\mathcal{C} = \mathcal{E}^{\frac{1}{\alpha}}([0,1]^2)$.

1.3 Approximation by Deep Neural Networks

We now substitute the concept of M-term approximation with representation systems by approximation through neural networks with M edges, i.e., sparsity in terms of the number of participating elements of a representation system is replaced by sparsity in terms of connectivity.

More formally, we consider the following setup.

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

$$\Gamma_M^{\mathcal{NN}}(f) := \inf_{\Phi \in \mathcal{NN}_{\infty,M,d,\rho}} \|f - \Phi(f)\|_{L^2(\Omega)}. \tag{3}$$

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

$$\sup_{f\in\mathcal{C}} \Gamma_M(f) \leq CM^{-\gamma}, \qquad \text{for all } M \in \mathbb{N}, \tag{4}$$

exists, determines the optimal M-edge approximation rate of $\mathcal{C}$ achievable by neural networks $\mathcal{NN}_{\infty,\infty,d,\rho}$*, and will henceforth be referred to as* $\gamma^*_{\mathcal{NN}}(\mathcal{C})$.

We emphasize that the infimum in (3) is taken over all networks with no more than M edges of non-zero weight. In particular, this means that the optimum is taken over all possible edge positions and associated weights.

Knowledge of the optimal M-edge approximation rate hence provides a bound on the approximation rate of a sparsely connected deep neural network. This bound is fundamental as it must be met by all learning algorithms. While we do not evaluate specific learning algorithms, our framework provides a means for assessing the quality of a given learning algorithm in the sense of measuring how close the rate induced by the algorithm comes to the optimal M-edge approximation rate.

1.4 Previous Work and Contributions

The theory of (M-edge) approximation with neural networks is a classical topic and we refer to [3,16,26–31] for a non-exhaustive list of relevant references. We hasten to add that these works consider exclusively the problem of studying lower bounds on γ in (4) for function classes $\mathcal{C}$ such as unit balls in Sobolev spaces, functions with bounded first moments, or continuously differentiable functions.

Our work [12] generalizes the results available in the literature in several respects:

- We study the impact of network weight quantization and characterize the associated memory requirements. While this may appear a mundane matter, we will see in Section 2.2 that the M-edge approximation rate alone is in general devoid of information related to the storage complexity of an approximating network. In contrast, our results directly lead to bounds on the storage complexity of networks that deliver optimal M-edge approximations.

- Under the (natural) assumption that the weights of an M-edge approximating network can be efficiently quantized, a notion made precise below, we provide upper bounds on γ in (4) for arbitrary signal classes $\mathcal{C}$. These bounds are universal in the sense of applying to all learning algorithms.
- We characterize explicitly the optimal M-edge approximation rate for a wide family of signal classes $\mathcal{C}$.

2. EFFECTIVE M-TERM AND M-EDGE APPROXIMATION AND BOUNDS ON UNIVERSAL APPROXIMATION

2.1 Effective M-term Approximation

Any dictionary $\mathcal{D}$ that is dense in $L^2(\mathbb{R}^d)$ will achieve $\gamma^*(\mathcal{C},\mathcal{D}) = \infty$ *independently of the signal class* $\mathcal{C}$. The catch, of course, is that (i) finding an optimal M-term approximation in such a dictionary is computationally intractable, and (ii) storage of the coefficients and indices of a corresponding optimal M-term approximation in such a dictionary may require more bits than direct encoding of the signal f. We summarize:

The quantity $\gamma^*(\mathcal{C},\mathcal{D})$ per se, in general, does not contain any information on the complexity of $\mathcal{C}$.

To address these issues, we define a variation of the concept of "best M-term approximation subject to polynomial-depth search" introduced in [13] and further developed in [32].

DEFINITION 2.1. *Let $\Omega \subset \mathbb{R}^d$. Consider the signal class $\mathcal{C} \subset L^2(\Omega)$ and the representation system $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$. $\mathcal{C}$ is said to have* effective M-term approximation rate $\gamma > 0$ in $\mathcal{D}$ *if there exist a univariate polynomial π and constants $C, D > 0$ such that for all $M \in \mathbb{N}$ and $f \in \mathcal{C}$*

$$\|f - \sum_{i\in I_M} c_i\varphi_i\|_{L^2(\Omega)} \leq CM^{-\gamma}, \tag{5}$$

for some index set $I_M \subset \{1,\dots,\pi(M)\}$ with $\#I_M = M$ and the coefficients $(c_i)_{i\in I_M}$ satisfy $\max_{i\in I_M}|c_i| \leq D$.

The supremum of all $\gamma > 0$ such that $\mathcal{C}$ has effective M-term approximation rate γ in $\mathcal{D}$ determines the optimal effective M-term approximation rate of $\mathcal{C}$ in $\mathcal{D}$ *and will henceforth be referred to as $\gamma^{*,\mathit{eff}}(\mathcal{C},\mathcal{D})$.*

It can be shown [13] that an effective M-term approximation of an element in a signal class $\mathcal{C}$ in a dictionary $\mathcal{D}$ can be stored with $O(M\cdot\log(M))$ bits, i.e., in optimal storage complexity (up to logarithmic factors).

2.2 Effective M-edge Approximation

We now turn to optimal M-edge approximations in neural networks. Our starting point is the following remarkable result.

THEOREM 2.2. *[33, Theorem 4] There exists a C^∞-function $\rho:\mathbb{R}\to\mathbb{R}$ which is strictly increasing and sigmoidal (i.e., $\lim_{x\to\infty}\rho(x) = 1$ and $\lim_{x\to-\infty}\rho(x) = 0$) such that for every $d \in \mathbb{N}$, every bounded $\Omega \subset \mathbb{R}^d$, every $f \in C([0,1]^d)$, and every $\epsilon > 0$ there exists a neural network Φ with rectifier ρ and two hidden layers of dimensions $N_1 = 3d$ and $N_2 = 6d+3$ satisfying*

$$\sup_{x\in\Omega}|f(x) - \Phi(x)| \leq \epsilon. \tag{6}$$

We note that the number of neurons, edges, and layers of the approximating network in Theorem 2.2 all do not depend on the approximation error ϵ. A direct consequence of Theorem 2.2 is the existence of a rectifier ρ such that $\gamma^*_{\mathcal{NN}}(\mathcal{C}) = \infty$ *independently of the signal class* $\mathcal{C}$. Again, the catch is that storing the weights of a corresponding optimal M-edge approximating network may require more bits than direct encoding of the signal f. In other words, while the representation of the approximating network Φ in (6) only requires a fixed number of real numbers, the associated storage complexity may be intractable. We summarize:

The quantity $\gamma^*_{\mathcal{NN}}(\mathcal{C})$ does, in general, not contain any information on the complexity of $\mathcal{C}$.

To address these issues, the following definition was introduced in [12].

DEFINITION 2.3. *Let $\Omega \subset \mathbb{R}^d$. Consider the signal class $\mathcal{C} \subset L^2(\Omega)$ and the rectifier ρ. $\mathcal{C}$ is said to have* effective *M-edge approximation rate $\gamma > 0$ in neural networks with rectifier ρ if there exists a univariate polynomial π such that there is a constant $C > 0$ so that for all $M \in \mathbb{N}$ and all $f \in \mathcal{C}$*

$$\|f - \Phi\|_{L^2(\Omega)} \leq CM^{-\gamma} \tag{7}$$

for some $\Phi \in \mathcal{NN}_{\infty,M,d,\rho}$ with the weights of Φ all bounded in absolute value by $\pi(M)$.

The supremum of all $\gamma > 0$ such that $\mathcal{C}$ has effective M-edge approximation rate γ determines the optimal effective M-edge approximation rate of $\mathcal{C}$ by neural networks (with rectifier ρ) *and will be denoted as $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C})$.*

Using results from min-max rate distortion theory, briefly reviewed below, we will be able to show that, in contrast to $\gamma^*_{\mathcal{NN}}(\mathcal{C})$, the quantity $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C})$ accurately reflects the complexity of the function class $\mathcal{C}$.

2.3 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 [13, 32]. For a detailed description of min-max rate-distortion theory, we refer to the excellent survey article [34].

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

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

the set of *binary encoders mapping elements of $\mathcal{C}$ to bit strings of length ℓ*, and we let

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

be the set of *binary decoders mapping bit strings of length ℓ to elements of $L^2(\Omega)$*. An encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ is said to *achieve distortion $\epsilon > 0$ over the function class $\mathcal{C}$* if

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

This means that the worst case error incurred by applying the encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ to an element of $\mathcal{C}$ is upper-bounded by ϵ, often also expressed as the uniform error over $\mathcal{C}$ being bounded by ϵ.

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 distortion $\epsilon > 0$ over the function class $\mathcal{C}$, and its asymptotic behavior as made precise in the following definition.

DEFINITION 2.4. *Let $d \in \mathbb{N}$ and $\mathcal{C} \subset L^2(\Omega)$. Then, for $\epsilon > 0$ the* minimax code length *$L(\epsilon, \mathcal{C})$ is given by*

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

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

$$\gamma^*(\mathcal{C}) := \inf\{\gamma \in \mathbb{R} : L(\epsilon, \mathcal{C}) = O(\epsilon^{-\gamma})\}.$$

The optimal exponent $\gamma^*(\mathcal{C})$ describes how fast $L(\epsilon, \mathcal{C})$ tends to infinity as ϵ decreases. For function classes $\mathcal{C}_1$ and $\mathcal{C}_2$, $\gamma^*(\mathcal{C}_1) < \gamma^*(\mathcal{C}_2)$ says that asymptotically, i.e., for $\epsilon \to 0$, the length of the encoding bit string for $\mathcal{C}_2$ is larger than that for $\mathcal{C}_1$. In other words, a smaller exponent indicates smaller description complexity. The optimal exponent $\gamma^*(\mathcal{C})$ therefore determines the minimal memory requirements for storing signals $f \in \mathcal{C}$ such that reconstruction with a uniformly bounded error is possible.

We mention that sometimes in the literature the reciprocal of $\gamma^*(\mathcal{C})$ is termed the optimal exponent. 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 $B^s_{p,q}(\mathbb{R}^d)$ whose norm is bounded by a constant $C > 0$ [35].

For this class we have $\gamma^*(\mathcal{C}) = \frac{d}{s}$. We will be particularly interested in so-called β-cartoon-like functions, for which the optimal exponent is given by $\frac{2}{\beta}$.

The optimal exponent $\gamma^*(\mathcal{C})$ is related to the quantities $\gamma^{*,\mathrm{eff}}(\mathcal{C},\mathcal{D})$, $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C})$ as follows.

THEOREM 2.5. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ bounded, and $\mathcal{C} \subset L^2(\Omega)$ a signal class. Then, the following statements hold.*

(i) For every dictionary $\mathcal{D}$, we have

$$\gamma^{*,eff}(\mathcal{C},\mathcal{D}) \leq \frac{1}{\gamma^*(\mathcal{C})}.$$

(ii) Suppose that the rectifier ρ is either globally Lipschitz-continuous, or differentiable with derivative of at most polynomial growth. Then,

$$\gamma^{*,eff}_{\mathcal{NN}}(\mathcal{C}) \leq \frac{1}{\gamma^*(\mathcal{C})}.$$

Proof. The result (i) is well known, a proof can be found in [32]. For the proof of (ii) we refer to [12]. In both cases the main idea is to show that for all $\gamma < \gamma^{*,\mathrm{eff}}(\mathcal{C},\mathcal{D})$, respectively $\gamma < \gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C})$, an effective M-term, respectively an effective M-edge, approximation can be encoded by a bit string of length $O(M\log(M))$ with distortion at most $O(M^{-\gamma})$. In the neural network case, such an encoding can be achieved by a quantization procedure whose analysis requires a novel weight-perturbation result for neural networks [12]. Based on these encodings, one can directly appeal to the definition of the optimal exponent to get the desired inequalities. □

2.4 Universal Lower Bounds for Neural Network Approximation

Theorem 2.5 readily provides a lower bound on the connectivity and the memory requirements of any learning algorithm guaranteeing uniform approximation rates for a given signal class. For a given network Φ, we denote its connectivity, i.e., the number of nonzero edge weights, by $\mathcal{M}(\Phi)$. Then, the following result holds (see [12, Theorem 2.7]).

THEOREM 2.6. *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, $\rho : \mathbb{R} \to \mathbb{R}$ a rectifier that is either globally Lipschitz-continuous, or differentiable with derivative of at most polynomial growth, and let $\mathcal{C} \subset L^2(\Omega)$. 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 $(\epsilon, f) \in (0,\frac{1}{2}) \times \mathcal{C}$, every weight of the neural network $\mathbf{Learn}(\epsilon, f)$ is polynomially bounded in ϵ^{-1}, and

$$\sup_{f\in\mathcal{C}} \|f - \mathbf{Learn}(\epsilon, f)\|_{L^2(\Omega)} \leq \epsilon. \tag{8}$$

Then,

$$\sup_{\epsilon\in(0,\frac{1}{2})} \epsilon^{\gamma} \cdot \sup_{f\in\mathcal{C}} \mathcal{M}(\mathbf{Learn}(\epsilon, f)) = \infty, \quad \text{for all } \gamma < \gamma^*(\mathcal{C}). \tag{9}$$

This result quantifies the minimum network connectivity needed to allow approximation of *all* elements in $\mathcal{C}$ to within an error of ϵ. In particular, it says that the minimum (complexity) exponent at which the number of edges in the network needs to scale is given by $\gamma^*(\mathcal{C})$. It furthermore establishes a universal link between the connectivity of the approximating network and the complexity of the function class that it is to approximate.

3. OPTIMAL NEURAL NETWORK APPROXIMATION

The previous section established a lower bound on the connectivity (and memory requirements) of neural networks guaranteeing uniform approximation rates for arbitrary signal classes. We next discuss the sharpness of these bounds. Motivated by Theorem 2.5, we introduce the following definition.

DEFINITION 3.1. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ bounded, and $\mathcal{C} \subset L^2(\mathbb{R}^d)$ a signal class.*

(i) Let $\mathcal{D} \subset L^2(\Omega)$ be a dictionary. Then, $\mathcal{C}$ is said to be optimally effectively representable in $\mathcal{D}$ *if*

$$\gamma^{*,eff}(\mathcal{C},\mathcal{D}) = \frac{1}{\gamma^*(\mathcal{C})}.$$

(ii) Suppose that the rectifier ρ is either globally Lipschitz-continuous, or differentiable with derivative of at most polynomial growth. Then, $\mathcal{C}$ is said to be optimally effectively representable by neural networks (with rectifier ρ) *if*

$$\gamma^{*,eff}_{\mathcal{NN}}(\mathcal{C}) = \frac{1}{\gamma^*(\mathcal{C})}.$$

Note that, by Theorem 2.5, optimal effective representability of a signal class $\mathcal{C}$ by neural networks implies the existence of a learning algorithm yielding a best M-edge approximation with asymptotically optimal memory requirements. While optimal effective representability in dictionaries is a well-studied subject, the literature is essentially void of results on optimal effective representability by neural networks. In what follows, we illustrate how to leverage known results on approximation in dictionaries to get statements on approximation via neural networks.

3.1 A Transference Principle

Given the benchmark result Theorem 2.5, we now ask the following question:

Which signal classes are optimally effectively representable by (deep) neural networks?

It turns out, as shown in [12], that the answer to this question is: A large family of signal classes. The mathematical technique developed in [12] to establish this statement is interesting in its own right as it constitutes a general framework for transferring results on function approximation through dictionaries to analogous results on approximation by neural networks. To this end, we first need the following.

DEFINITION 3.2. *Let $\Omega \subset \mathbb{R}^d$, $\rho : \mathbb{R} \to \mathbb{R}$, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ a representation system. Then, $\mathcal{D}$ is said to be* effectively representable by neural networks (with rectifier ρ) *if there exist a polynomial $\pi : \mathbb{R}^2 \to \mathbb{R}$ and $R \in \mathbb{N}$ such that for all $\eta > 0$ and every $i \in I$, there is a neural network $\Phi_{i,\eta} \in \mathcal{NN}_{\infty,R,d,\rho}$ with*

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

such that the absolute values of the weights of the $\Phi_{i,\eta}$ are bounded by $\pi(i,\eta^{-1})$.

We are now ready to state our main transference result (see [12, Theorem 3.4]).

THEOREM 3.3. *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, $\rho : \mathbb{R} \to \mathbb{R}$ a rectifier that is either globally Lipschitz-continuous, or differentiable with derivative of at most polynomial growth, let $\mathcal{D} \subset L^2(\Omega)$ be a dictionary and $\mathcal{C} \subset L^2(\Omega)$ a signal class. Suppose that $\mathcal{D}$ is effectively representable by neural networks with rectifier ρ. Then, we have*

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

In particular, optimal effective representability in $\mathcal{D}$ implies optimal effective representability by neural networks with rectifier ρ.

3.2 Affine Systems

In this section, we verify that dictionaries of *affine systems*, defined formally next, are effectively representable by neural networks. With Theorem 3.3 this will then allow us to characterize function classes $\mathcal{C}$ that are optimally M-edge approximated by neural networks.

DEFINITION 3.4. *Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ bounded, and $h \in L^2(\mathbb{R}^d)$ compactly supported. Let $\delta > 0$, $(c_i^s)_{i=1}^r \subset \mathbb{R}$, for $s = 1, \dots, S$, and $(d_i)_{i=1}^r \subset \mathbb{R}^d$. Moreover, let $(A_j)_{j\in\mathbb{N}} \subset \mathbb{R}^{d\times d}$ be expanding. Consider the compactly supported functions*

$$g_s := \sum_{i=1}^{r} c_i^s h(\cdot - d_i), \quad s = 1, \dots, S. \tag{10}$$

We define the corresponding affine system *as*

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

where χ_Ω denotes the indicator function of Ω.

Affine systems comprise a wealth of popular representation systems used in computational harmonic analysis such as wavelets, shearlets, α-molecules, and tensor products thereof. We next show that verifying whether an affine system is effectively representable requires verifying effective representability for its generator function h only.

THEOREM 3.5. *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ is an affine system according to Definition 3.4. Suppose further that for the rectifier function $\rho : \mathbb{R} \to \mathbb{R}$ there exist a constant M and a polynomial $\pi : \mathbb{R}^2 \to \mathbb{R}$ such that for all $D, \epsilon > 0$ there is $\Phi_{D,\epsilon} \in \mathcal{NN}_{L,M,d,\rho}$ with weights bounded in absolute value by $\pi(\epsilon^{-1}, D)$ and*

$$\|h - \Phi_{D,\epsilon}\|_{L^2([-D,D]^d)} \leq \epsilon, \tag{11}$$

where h is as in Definition 3.4. Assume furthermore, that there exist $a, c > 0$ such that for all $j \in \mathbb{N}$

$$\sum_{k=1}^{j-1} \det(A_k) \geq c|\det(A_j)|^a.$$

Then, $\mathcal{D}$ is effectively representable by neural networks with rectifier function ρ.

To obtain optimal effective representability of a function class $\mathcal{C}$ by a wavelet-, shearlet-, ridgelet-, or α-molecule system, the precise nature of the generator h does not play a role, rather its smoothness and vanishing moment properties only. The key idea of our construction is to build generators that are arbitrarily smooth bump functions which then, by choosing c_i^s in (10) accordingly, lead to wavelet-, shearlet-, ridgelet-, or α-molecule systems with smooth generators with vanishing moments.

Approximation of smooth bump functions, at arbitrary precision, with neural networks of fixed size is certainly not possible for all rectifiers ρ. We therefore have to restrict ourselves to networks with suitable rectifiers and recall the following definition from [3, 36].

DEFINITION 3.6. *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 $K > 0$ such that*

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

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

$$\left|\frac{1}{x^k}\rho(x)\right| \leq K|x|^{-a}, \text{ for } x < 0, \ \left|\frac{1}{x^k}\rho(x) - 1\right| \leq Kx^{-a}, \text{ for } x \geq 0, |\rho(x)| \leq K(1 + |x|)^k, \ \text{for } x \in \mathbb{R},$$

and

$$\left|\frac{d}{dx}\rho(x)\right| \leq K|x|^b, \quad \text{for } x \in \mathbb{R}.$$

For $m \in \mathbb{N}$ we denote the univariate cardinal B-spline of order $m \in \mathbb{N}$ by N_m, i.e., $N_1 = \chi_{[0,1]}$, 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.

Additionally, we define, for $d \in \mathbb{N}$, the d-dimensional B-spline of order m by N_m^d. It was established in [36] that B-splines can be well approximated by neural networks with sigmoidal rectifiers. Nonetheless, the result of [36] does not establish that dictionaries based on B-splines are *effectively* representable by neural networks. This stronger statement will be made possible by strong sigmoidality of ρ as established in [12, Theorem 4.4].

THEOREM 3.7. *Let $d, m, k \in \mathbb{N}$, and ρ strongly sigmoidal of order $k \geq 2$. Further, let $L := \lceil \log(md - d)/\log(k) \rceil + 1$. Then, there exists $M \in \mathbb{N}$, possibly dependent on s, m, k, such that for all $D, \epsilon > 0$, there is a network $\Phi_{D,\epsilon} \in \mathcal{NN}_{L,M,d,\rho}$ with*

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

Moreover, there exists a polynomial $\pi : \mathbb{R}^2 \to \mathbb{R}$ such that the absolute values of the weights of the network $\Phi_{D,\epsilon}$ are bounded by $\pi(D, \epsilon^{-1})$.

This result in combination with Theorem 3.5 implies that affine systems with generators based on linear combinations of B-splines are effectively representable by neural networks with strongly sigmoidal rectifiers. In

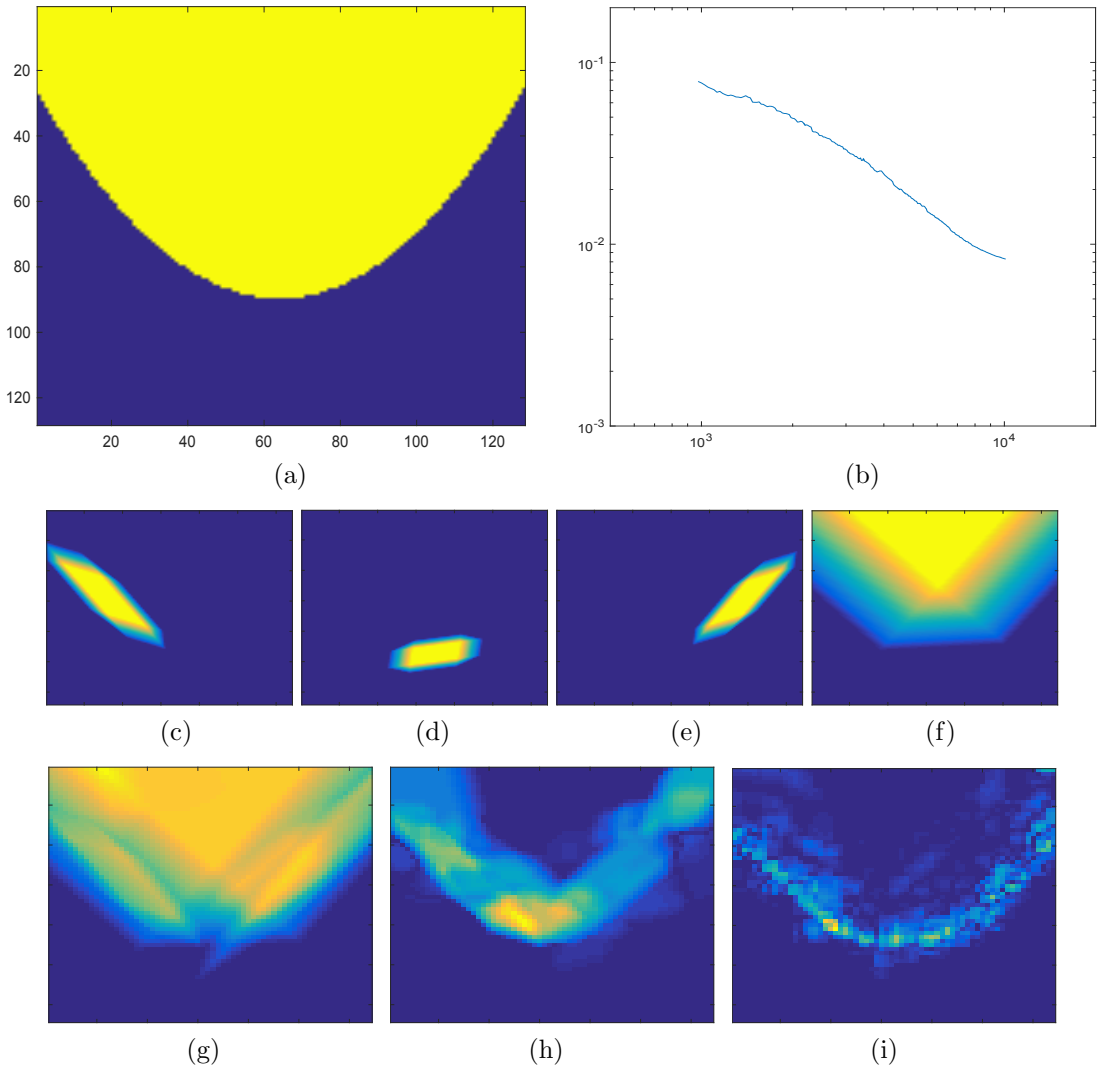


Figure 2. (a): Function with curvilinear singularity to be approximated by the network. (b): Approximation error as a function of the number of edges. (c)-(f): Shearlet-like subnetworks. (g): Reconstruction using only the 10 subnetworks with the largest supports. (h): Reconstruction using only subnetworks with medium-sized support. (i): Reconstruction using only subnetworks with very small support.

combination with Theorem 3.3 we conclude that neural networks provide optimal M-edge approximation rates for all signal classes that are optimally effectively representable in *any* affine system (wavelets, shearlets, ridgelets, or α-molecules) with generators based on linear combinations of B-splines.

In addition, we also conclude that the memory requirements for storing the corresponding optimal M-edge approximating networks are asymptotically optimal. This result serves as an explanation for the "unreasonable effectiveness" of neural networks: they combine the optimal approximation properties of all affine systems taken together and do so with minimal memory requirements.

4. NUMERICAL RESULTS

We illustrate our theoretical findings with a numerical example from [12]. The signal class $\mathcal{C}$ under consideration is given by so-called *cartoon-like images* which are piecewise C^2 functions separated by a C^2 discontinuity curve, see Figure 2(a) for an example. It is well known that $\gamma^*(\mathcal{C}) = 1$ and that $\mathcal{C}$ is optimally effectively representable by shearlets, or more generally parabolic molecules [19]. As shearlets constitute an instance of affine systems [12] we would expect an optimal M-edge approximation rate of $\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C}) = 1$. Figure 2(b) illustrates that this optimal rate can be observed empirically by training a neural network using stochastic gradient descent via backpropagation. We refer to [12, Section 7] for further details.

ACKNOWLEDGMENTS

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

REFERENCES

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

[2] Hornik, K., "Approximation capabilities of multilayer feedforward networks," *Neural Networks* **4**(2), 251 – 257 (1991).

[3] Cybenko, G., "Approximation by superpositions of a sigmoidal function," *Math. Control Signal* **2**(4), 303–314 (1989).

[4] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in Neural Information Processing Systems 25*], 1097–1105, Curran Associates, Inc. (2012).

[5] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B., "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).

[6] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D., "Mastering the game of Go with deep neural networks and tree search," *Nature* **529**(7587), 484–489 (2016).

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

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

[9] Steinwart, I. and Christmann, A., [*Support Vector Machines*], Information Science and Statistics, Springer, New York (2008).

[10] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., "Learning representations by back-propagating errors," 696–699, MIT Press, Cambridge, MA, USA (1988).

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

[12] Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P., "Optimal approximation with sparsely connected deep neural networks," arXiv:1705.01714, (2017).
[13] Donoho, D. L., "Unconditional bases are optimal bases for data compression and for statistical estimation," *Appl. Comput. Harmon. Anal.* **1**(1), 100 – 115 (1993).
[14] DeVore, R. A. and Lorentz, G. G., [*Constructive Approximation*], Springer (1993).
[15] Daubechies, I., [*Ten Lectures on Wavelets*], SIAM (1992).
[16] Candès, E. J., "Ridgelets: Theory and Applications," (1998). Ph.D. thesis, Stanford University.
[17] Candès, E. J. and Donoho, D. L., "New tight frames of curvelets and optimal representations of objects with piecewise-C2 singularities," *Comm. Pure Appl. Math.* **57**, 219–266 (2002).
[18] Guo, K., Kutyniok, G., and Labate, D., "Sparse multidimensional representations using anisotropic dilation and shear operators," *Wavelets and Splines (Athens, GA, 2005)* , 189–201 (2006).
[19] Grohs, P. and Kutyniok, G., "Parabolic molecules," *Found. Comput. Math.* **14**, 299–337 (2014).
[20] Grohs, P., Keiper, S., Kutyniok, G., and Schäfer, M., "α-molecules," *Appl. Comput. Harmon. Anal.* **41**(1), 297–336 (2016).
[21] Dauge, M. and Stevenson, R., "Sparse tensor product wavelet approximation of singular functions," *SIAM J. Math. Anal.* **42**(5), 2203–2228 (2010).
[22] Gröchenig, K., [*Foundations of time-frequency analysis*], Springer Science & Business Media (2013).
[23] Demanet, L. and Ying, L., "Wave atoms and sparsity of oscillatory patterns," *Appl. Comput. Harmon. Anal.* **23**(3), 368–387 (2007).
[24] Grohs, P., Keiper, S., Kutyniok, G., and Schäfer, M., "Cartoon approximation with α-curvelets," *J. Fourier Anal. Appl.* **22**(6), 1235–1293 (2016).
[25] Pein, A. and Voigtländer, F., "Analysis sparsity versus synthesis sparsity for α-shearlets," arXiv:1702.03559, (2017).
[26] Hornik, K., Stinchcombe, M., and White, H., "Multilayer feedforward networks are universal approximators," *Neural Networks* **2**(5), 359–366 (1989).
[27] Mhaskar, H. N., "Neural networks for optimal approximation of smooth and analytic functions," *Neural Comput.* **8**(1), 164–177 (1996).
[28] Mhaskar, H. and Micchelli, C., "Degree of approximation by neural and translation networks with a single hidden layer," *Adv. Appl. Math.* **16**(2), 151–183 (1995).
[29] DeVore, R., Oskolkov, K., and Petrushev, P., "Approximation by feed-forward neural networks," *Ann. Numer. Math.* **4**, 261–287 (1996).
[30] Barron, A. R., "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory* **39**(3), 930–945 (1993).
[31] Shaham, U., Cloninger, A., and Coifman, R. R., "Provable approximation properties for deep neural networks," *Appl. Comput. Harmon. Anal.*, in press.
[32] Grohs, P., "Optimally sparse data representations," in [*Harmonic and Applied Analysis*], 199–248, Springer (2015).
[33] Maiorov, V. and Pinkus, A., "Lower bounds for approximation by MLP neural networks," *Neurocomputing* **25**(1), 81–91 (1999).
[34] Donoho, D., Vetterli, M., DeVore, R., and Daubechies, I., "Data compression and harmonic analysis," *IEEE Trans. Inf. Theory* **44**, 2435–2476 (1998).
[35] Cohen, A., Dahmen, W., Daubechies, I., and DeVore, R., "Tree approximation and optimal encoding," *Appl. Comput. Harmon. Anal.* **11**(2), 192–226 (2001).
[36] Chui, C. K., Li, X., and Mhaskar, H. N., "Neural networks for localized approximation," *Math. Comp.* **63**(208), 607–623 (1994).