We discuss approximation of functions using deep neural nets. Given a function f on a d-dimensional manifold $\Gamma \subset \mathbb{R}^m$, we construct a sparsely-connected depth-4 neural network and bound its error in approximating f. The size of the network depends on dimension and curvature of the manifold Γ, the complexity of f, in terms of its wavelet description, and only weakly on the ambient dimension m. Our network essentially computes wavelet functions, which are computed from Rectified Linear Units (ReLU).

Provable approximation properties for deep neural networks

Uri Shaham[†], Alexander Cloninger[‡] and Ronald R. Coifman[‡],

Technical Report YALEU/DCS/TR-1513

[†] Department of Statistics, Yale University, New Haven CT 06511
[‡] Applied Mathematics Program, Yale University, New Haven CT 06511

Approved for public release: distribution is unlimited.
Keywords: *deep neural networks, approximation rates, manifold learning*

1 Introduction

Since 2006, deep learning algorithms achieved unprecedented success and state-of-the-art results in various machine learning and artificial intelligence tasks, most notably image recognition (for example, [1], [2], [3], [4]), Optical Character Recognition (OCR, for example, [5], [6]), speech recognition (for example, [7], [8] [9]), text analysis and Natural Language Processing (NLP, for example [10]). Deep Neural Networks (DNNs) are general in the sense of their mechanism for learning features of the data. Nevertheless, in numerous cases, results obtained with DNNs outperformed previous state-of-the-art methods, often requiring significant domain knowledge, manifested in hand-crafted features.

Despite the great success of DNNs in many practical applications, the theoretical framework of DNNs is still lacking; along with some decades-old well-known results, developing aspects of such theoretical framework are the focus of much recent academic attention. In particular, some interesting topics are (1) specification of the network topology (i.e., depth, layer sizes), given a target function, in order to obtain certain approximation properties, (2) estimating the amount of training data needed in order to generalize to test data with high accuracy, and also (3) development of training algorithms with performance guarantees.

1.1 The contribution of this work

In this manuscript we discuss the first topic. Specifically, we prove a formal version of the following result:

Theorem (informal version) 1.1. *Let $\Gamma \subset \mathbb{R}^m$ be a smooth d-dimensional manifold, $f \in L_2(\Gamma)$ and let $\delta > 0$ be an approximation level. Then there exists a depth-4 sparsely-connected neural network with N units where $N = N(\delta, \Gamma, f, m)$, computing the function f_N such that*

$$\|f - f_N\|_2^2 \leq \delta. \tag{1}$$

The number $N = N(\delta, \Gamma, f, m)$ depends on the complexity of f, in terms of its wavelet representation, the curvature and dimension of the manifold Γ and only weakly on the ambient dimension m, thus taking advantage of the possibility that $d \ll m$, which seems to be realistic in many practical applications. Moreover, we specify the exact topology of such network, and show how it depends on the curvature of Γ, the complexity of f, and the dimensions d, and m. Lastly, for two classes of functions we also provide approximation error rates: L_2 error rate for functions with sparse wavelet expansion and point-wise error rate for functions in C^2:

- if f has wavelet coefficients in l_1 then there exists a depth-4 network and a constant c so that

$$\|f - f_N\|_2^2 \leq \frac{c}{N} \tag{2}$$

- if $f \in C^2$ and has bounded Hessian, then there exists a depth-4 network so that

$$\|f - f_N\|_\infty = O\left(N^{-\frac{2}{d}}\right). \tag{3}$$

1.2 The structure of this manuscript

The structure of this manuscript is as follows: in Section 2 we review some of the fundamental theoretical results in neural network analysis, as well as some of the recent theoretical developments. In Section 3 we give quick technical review of the mathematical methods and results that are used in our construction. In Section 4 we describe our main result, namely construction of deep neural nets for approximating functions on smooth manifolds. In Section 5 we specify the size of the network needed to learn a function f, in view of the construction of the previous section. Section 6 concludes this manuscript.

1.3 Notation

Γ denotes a d-dimensional manifold in $\mathbb{R}^m$. $\{(U_i, \phi_i)\}$ denotes an atlas for Γ. Tangent hyperplanes to Γ are denoted by H_i. f and variants of it stand for the function to be approximated. φ, ψ are scaling (aka "father") and wavelet (aka "mother") functions, respectively. The wavelet terms are indexed by scale k and offset b. The support of a function f is denoted by $\mathrm{supp}(f)$.

2 Related work

There is a huge body of theoretical work in neural network research. In this section, we review some classical theoretical results on neural network theory, and discuss several recent theoretical works.

A well known result, proved independently by Cybenko [11], Hornik [12] and others states that Artificial Neural Networks (ANNs) with a single hidden layer of sigmoidal functions can approximate arbitrary closely any compactly supported continuous function. This result is known as the "Universal Approximation Property". It does not relate, however, the number of hidden units and the approximation accuracy; moreover, the hidden layer might contain a very large number of units. Several works propose extensions of the universal approximation property (see, for example[13, 14], for a regularization perspective and also using radial basis activation functions, [15] for all activation functions that achieve the universal approximation property).

The first work to discuss the approximation error rate was done by Barron [16], who showed that given a function $f : \mathbb{R}^m \to \mathbb{R}$ with bounded first moment of the magnitude of the Fourier transform

$$C_f = \int_{\mathbb{R}^m} |w||\tilde{f}(w)| < \infty \tag{4}$$

there exists a neural net with a single hidden layer of N sigmoid units, so that the output f_N of the network satisfies

$$\|f - f_N\|_2^2 \leq \frac{c_f}{N}, \tag{5}$$

where c_f is proportional to C_f. We note that the requirement (4) gets more restrictive when the ambient dimension m is large, and that the constant c_f might scale with m. The dependence on m is improved in [17], [18]. In particular, in [17] the constant is improved to be polynomial in m. For r times differentiable functions, Mahskar [19] constructs network with a single hidden layer of N sigmoid units (with weights that do not depend on the target function) that achieves an approximation error rate

$$\|f - f_N\|_2^2 = \frac{c}{N^{2r/m}}, \tag{6}$$

which is known to be optimal. This rate is also achieved (point-wise) in this manuscript, however, with respect to the dimension d of the manifold, instead of m, which might be a significant difference when $d \ll m$.

Universal approximation properties of Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs) were proved in [20] and [21] (and later improved in [22]), where it is shown that RBMs and DBNs can approximate arbitrarily well any distribution on $\{0, 1\}^m$.

During the decade of 1990s, a popular direction in neural network research was to construct neural networks in which the hidden units compute wavelets functions (see, for example [23],

[24] and [25]). These works, however, do not give any specification of network architecture to obtain desired approximation properties.

Several most interesting recent theoretical results aim to understand the experimental success of deep architectures over shallow ones, although without any theoretical guarantees. In [26], Montufar et al. show that DNNs can learn more complex functions than can learn a shallow network with same number of units, where complexity is defined as the number of linear regions of the function. Patel et. al [27] give a probabilistic framework of deep learning. In particular, they describe convolutional nets in terms of message passing algorithms. Unsupervised deep learning is recently discussed from several interesting points of view: a group-theoretic perspective is proposed in [28]; their theory also explains why higher layers tend to learn more abstract features, a well-observed phenomenon in deep learning practice. A similar conclusion is implied in [29], where the authors show that there is a mapping between RBMs and re-normalization group, a powerful coarse-graining tool in theoretical physics, which makes use of marginalization in order describe physical systems in greater length scales. [30] proposes to evaluate the representations obtained by deep networks which via the information bottleneck principal, which is a trade-off between compression of the input representation and predictive ability of the output function.

On the algorithmic side, Arora et al. [31], provide polynomial time algorithms with provable guarantees for learning in networks with sparse connectivity and random weights in $[-1, 1]$. Sedghi et.al show how the first layer weight matrix can be recovered in networks that have sparse connectivity. Livni et al. [32] analyze the expressiveness of neural nets in terms of Turing machines and provide a provably correct algorithm for training polynomial nets of depth 2 and 3. Complex valued convolutional nets are proposed in [33]. A recent work by Chui and Mhaskar brought to our attention [34] constructs a network with similar functionality to the network we construct in this manuscript. In their network the low layers map the data to local coordinates on the manifold and the upper ones approximate a target function on each chart, however using B-splines.

3 Preliminaries

3.1 Compact manifolds in $\mathbb{R}^m$

In this section we review the concepts of *smooth manifolds*, *atlases* and *partition of unity*, which will all play important roles in our construction.

Let $\Gamma \subseteq \mathbb{R}^m$ be a compact d-dimensional manifold. We further assume that Γ is smooth, and that there exists $\delta > 0$ so that for all $x \in \Gamma$, $B(x, \delta) \cap \Gamma$ is diffeomorphic to a disc, with a map that is close to the identity.

Definition 3.1. *A* ***chart*** *for Γ is a pair (U, ϕ) such that $U \subseteq \Gamma$ is open and*

$$\phi : U \to M, \tag{7}$$

where ϕ is a homeomorphism and M is an open subset of a Euclidean space.

One way to think of a chart is as a tangent plane at some point $x \in U \subseteq \Gamma$, such that the plane defines a Euclidean coordinate system on U via the map ϕ.

Definition 3.2. *An* ***atlas*** *for Γ is a collection $\{(U_i, \phi_i)\}_{i\in I}$ of charts such that $\cup_i U_i = \Gamma$.*

Definition 3.3. *Let Γ be a smooth manifold. A* ***partition of unity*** *of Γ w.r.t an open cover $\{U_i\}_{i\in I}$ is a family of nonnegative smooth functions $\{\eta_i\}_{i\in I}$ such that for every $x \in X$, $\sum_i \eta_i(x) = 1$ and for every i, $\mathrm{supp}(\eta_i) \subseteq (U_i)$.*

Theorem 3.4. *(Proposition 13.9 in [35]) Let Γ be a compact manifold and $\{U_i\}_{i\in I}$ be an open cover of Γ. Then there exists a partition of unity $\{\eta_i\}_{i\in I}$ such that for each i, η_i is in C^∞, has compact support and $\mathrm{supp}(\eta_i) \subseteq U_i$.*

3.2 Harmonic analysis on spaces of homogeneous type

3.2.1 Construction of wavelet frames

In this section we cite several standard results, mostly from [36], showing how to construct a wavelet frame of $L_2(\mathbb{R}^d)$, and discuss some of its properties.

Definition 3.5. *(Definition 1.1 in [36])*
A ***space of homogeneous type*** *$(\mathcal{X}, \mu, \delta)$ is a set $\mathcal{X}$ together with a measure μ and a quasi-metric δ (satisfies triangle inequality up to a constant A) such that for every $x \in \mathcal{X}$, $r > 0$*

- $0 < \mu(B(x, r)) < \infty$
- *There exists a constant A' such that $\mu(B(x, 2r)) \leq A'\mu(B(x, r))$*

In this manuscript, we are interested in constructing a wavelet frame on $\mathbb{R}^d$, which, equipped with Lebesgue measure and the Euclidean metric, is a space of homogeneous type.

Definition 3.6. *(Definition* 3.14 *in [36])*
Let $(\mathcal{X}, \mu, \delta)$ *be a space of homogeneous type. A family of functions* $\{S_k\}_{k\in\mathbb{Z}}$*,* $S_k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ *is said to be a family of* ***averaging kernels*** *("father functions") if conditions* 3.14 − 3.18 *and* 3.19 *with* $\sigma = \epsilon$ *in [36] are satisfied. A family* $\{D_k\}_{k\in\mathbb{Z}}$*,* $D_k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ *is said to be a family of ("mother")* ***wavelets*** *if for all* $x, y \in \mathcal{X}$*,*

$$D_k(x, y) = S_k(x, y) - S_{k-1}(x, y), \tag{8}$$

and S_k, S_{k-1} *are averaging kernels.*

By standard wavelet terminology, we denote

$$\psi_{k,b}(x) \equiv 2^{-\frac{k}{2}} D_k(x, b). \tag{9}$$

Theorem 3.7. *(A simplified version of Theorem* 3.25 *in [36])*
Let $\{S_k\}$ *be a family of averaging kernels. Then there exist families* $\{\psi_{k,b}\}, \{\widetilde{\psi}_{k,b}\}$ *such that for all* $f \in L_2(\mathbb{R}^d)$

$$f(x) = \sum_{(k,b)\in\Lambda} \langle f, \widetilde{\psi}_{k,b}\rangle \psi_{k,b}(x) \tag{10}$$

Where the functions $\psi_{k,b}$ *are given by Equations* (8) *and* (9) *and* $\Lambda = \{(k, b) \in \mathbb{Z} \times \mathbb{R}^d\colon b \in 2^{-\frac{k}{d}}\mathbb{Z}^d\}$.

Remark 3.8. The functions $\{\psi_{k,b}\}$ need to be such that for every $x \in \mathbb{R}^d$, $\sum_{(k,b)\in\Lambda} \psi_{k,b}(x)$ is sufficiently large. This is discussed in great generality in chapter 3 in [36].

Remark 3.9. The functions $\widetilde{\psi}_{k,b}$ are called dual elements, and are also a wavelet frame of $L_2(\mathbb{R}^d)$.

3.3 Approximation of functions with sparse wavelet coefficients

In this section we cite a result from [37] regarding approximating functions which have sparse representation with respect to a dictionary $\mathcal{D}$ using finite linear combinations of dictionary elements.

Let f a function in some Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$, and let $\mathcal{D} \subset \mathcal{H}$ be a dictionary, i.e., any family of functions $(g)_{g\in\mathcal{D}}$ with unit norm. Assume that f can be represented as a linear combination of elements in $\mathcal{D}$ with absolutely summable coefficients, and denote the sum of absolute values of the coefficients in the expansion of f by $\|f\|_{\mathcal{L}_1}$.

In [37], it is shown that $\mathcal{L}_1$ functions can be approximated using N dictionary terms with squared error proportional to $\frac{1}{\sqrt{N}}$. As a bonus, we also get a greedy algorithm (though not always practical) for selecting the corresponding dictionary terms. OGA is a greedy algorithm that at the k'th iteration computes the residual

$$r_{k-1} := f - f_{k-1}, \tag{11}$$

finds the dictionary element that is most correlated with it

$$g_k := \arg\max_{g\in\mathcal{D}} |\langle r_{k-1}, g\rangle| \tag{12}$$

and defines a new approximation

$$f_k := P_k f, \tag{13}$$

where P_k is the orthogonal projection operator onto span$\{g_1, ..., g_k\}$.

Theorem 3.10. *(Theorem 2.1 from [37]) The error r_N of the OGA satisfies*

$$\|f - f_N\| \leq \|f\|_{\mathcal{L}_1}(N+1)^{-1/2}. \tag{14}$$

Clearly, for $\mathcal{H} = L_2(\mathbb{R}^d)$ we can choose the dictionary to be the wavelet frame given by

$$\mathcal{D} = \{\psi_{k,b} : (k, b) \in \mathcal{Z} \times \mathbb{R}^d, b \in 2^{-k}\mathbb{Z}\}. \tag{15}$$

Remark 3.11. Let $\mathcal{D} = \{\psi_{k,b}\}$ be a wavelet frame that satisfies the regularities in conditions 3.14 − 3.19 in [36]. Then if a function f is in $\mathcal{L}_1$ with respect to $\mathcal{D}$, it is also in $\mathcal{L}_1$ with respect to any other wavelet frame that satisfies the same regularities. In other words, having expansion coefficients in l_1 does not depend on the specific choice of wavelets (as long as the regularities are satisfied). The idea behind the proof of this claim is explained in appendix A.

Remark 3.12. Section 4.5 in [36] gives a way to check whether a function f has sparse coefficients without actually calculating the coefficients:

$$f \in \mathcal{L}_1 \text{ iff } \sum_{k\in\mathbb{Z}} 2^{k/2}\|f * \psi_{k,0}\|_1 < \infty, \tag{16}$$

i.e., one can determine if $f \in \mathcal{L}_1$ without explicitly computing its wavelet coefficients; rather, by convolving f with non-shifted wavelet terms in all scales.

4 Approximating functions on manifolds using deep neural nets

In this section we describe in detail the steps in our construction of deep networks, which are designed to approximate functions on smooth manifolds. The main steps in our construction are the following:

1. We construct a frame of $L_2(\mathbb{R}^d)$ in which the frame elements can be constructed from rectified linear units (see Section 4.1).

2. Given a d-dimensional manifold $\Gamma \subset \mathbb{R}^m$, we construct an atlas for Γ by covering it with open balls (see Section 4.2).

3. We use the open cover to obtain a partition of unity of Γ and consequently represent any function on Γ as a sum of functions on $\mathbb{R}^d$ (see section 4.3).

4. We show how to extend the wavelet terms in the wavelet expansion, which are defined on $\mathbb{R}^d$, to $\mathbb{R}^m$ in a way that depends on the curvature of the manifold Γ (see Section 4.4).

4.1 Constructing a wavelet frame from rectifier units

In this section we show how Rectified Linear Units (ReLU) can be used to obtain a wavelet frame of $L_2(\mathbb{R}^d)$. The construction of wavelets from rectifiers is fairly simple, and we refer to results from Section 3.2 to show that they obtain a frame of $L_2(\mathbb{R}^d)$.

The rectifier activation function is defined on $\mathbb{R}$ as

$$\operatorname{rect}(x) = \max\{0, x\}. \tag{17}$$

we define a trapezoid-shaped function $t : \mathbb{R} \to \mathbb{R}$ by

$$t(x) = \operatorname{rect}(x+3) - \operatorname{rect}(x+1) - \operatorname{rect}(x-1) + \operatorname{rect}(x-3). \tag{18}$$

We then define the scaling function $\varphi : \mathbb{R}^d \to \mathbb{R}$ by

$$\varphi(x) = C_d \operatorname{rect}\left(\sum_{j=1}^{d} t(x_j) - 2(d-1)\right), \tag{19}$$

where the constant C_d is such that

$$\int_{\mathbb{R}^d} \varphi(x)dx = 1; \tag{20}$$

for example, $C_1 = \frac{1}{8}$. Following the construction in Section 3.2, we define

$$S_k(x, b) = 2^k \varphi(2^{\frac{k}{d}}(x-b)) \tag{21}$$

Lemma 4.1. *The family $\{S_k\}$ is a family of averaging kernels.*

The proof is given in Appendix B. Next we define the ("mother") wavelet as

$$D_k(x,y) = S_k(x,y) - S_{k-1}(x,y), \tag{22}$$

And denote

$$\psi_{k,b}(x) \equiv 2^{-\frac{k}{2}} D_k(x,b), \tag{23}$$

and

$$\begin{aligned} \psi(x) &\equiv \psi_{0,0}(x) && (24)\\ &= D_0(x,0) && (25)\\ &= S_0(x,0) - S_{-1}(x,0) && (26)\\ &= \varphi(x) - 2^{-1}\varphi(2^{-\frac{1}{d}}x)). && (27) \end{aligned}$$

Figure 1 shows the construction of φ and ψ in for $d = 1, 2$.

Remark 4.2. We can see that

$$\begin{aligned} \psi_{k,b}(x) &= 2^{-\frac{k}{2}} D_k(x,b) && (28)\\ &= 2^{-\frac{k}{2}}(S_k(x,b) - S_{k-1}(x,b)) && (29)\\ &= 2^{-\frac{k}{2}}(2^k\varphi(2^{\frac{k}{d}}(x-b)) - 2^{k-1}\varphi(2^{\frac{k-1}{d}}(x-b))) && (30)\\ &= 2^{\frac{k}{2}}\left(\varphi(2^{\frac{k}{d}}(x-b)) - 2^{-1}\varphi(2^{\frac{k-1}{d}}(x-b))\right) && (31)\\ &= 2^{\frac{k}{2}}\psi\left(2^{\frac{k}{d}}(x-b)\right). && (32) \end{aligned}$$

Remark 4.3. With the above construction, φ can be computed using a network with $4d$ rectifier units in the first layer and a single unit in the second layer. Hence every wavelet term $\psi_{k,b}$ can be computed using $8d$ rectifier units in the first layer, 2 rectifier units in the second layer and a single linear unit in the third layer. From this, the sum of k wavelet terms can be computed using a network with $8dk$ rectifiers in the first layer, $2k$ rectifiers in the second layer and a single linear unit in the third layer.

From Theorem 3.7 and the above construction we then get the following lemma:

Lemma 4.4. *$\{\psi_{k,b} : k \in \mathbb{Z}, b \in 2^{-k}\mathbb{Z}\}$ is a frame of $L_2(\mathbb{R}^d)$.*

Next, the following lemma uses properties of the above frame to obtain point-wise error bounds in approximation of compactly supported functions $f \in C^2$.

Lemma 4.5. *Let $f \in L_2(\mathbb{R}^d)$ be compactly supported, twice differentiable and let $\|\nabla_f^2\|_{op}$ be bounded. Then for every $k \in \mathbb{N} \cup \{0\}$ there exists a combination f_K of terms up to scale K so that for every $x \in \mathbb{R}^d$*

$$|f(x) - f_K(x)| = O\left(2^{-\frac{2K}{d}}\right). \tag{33}$$

The proof is given in Appendix C.

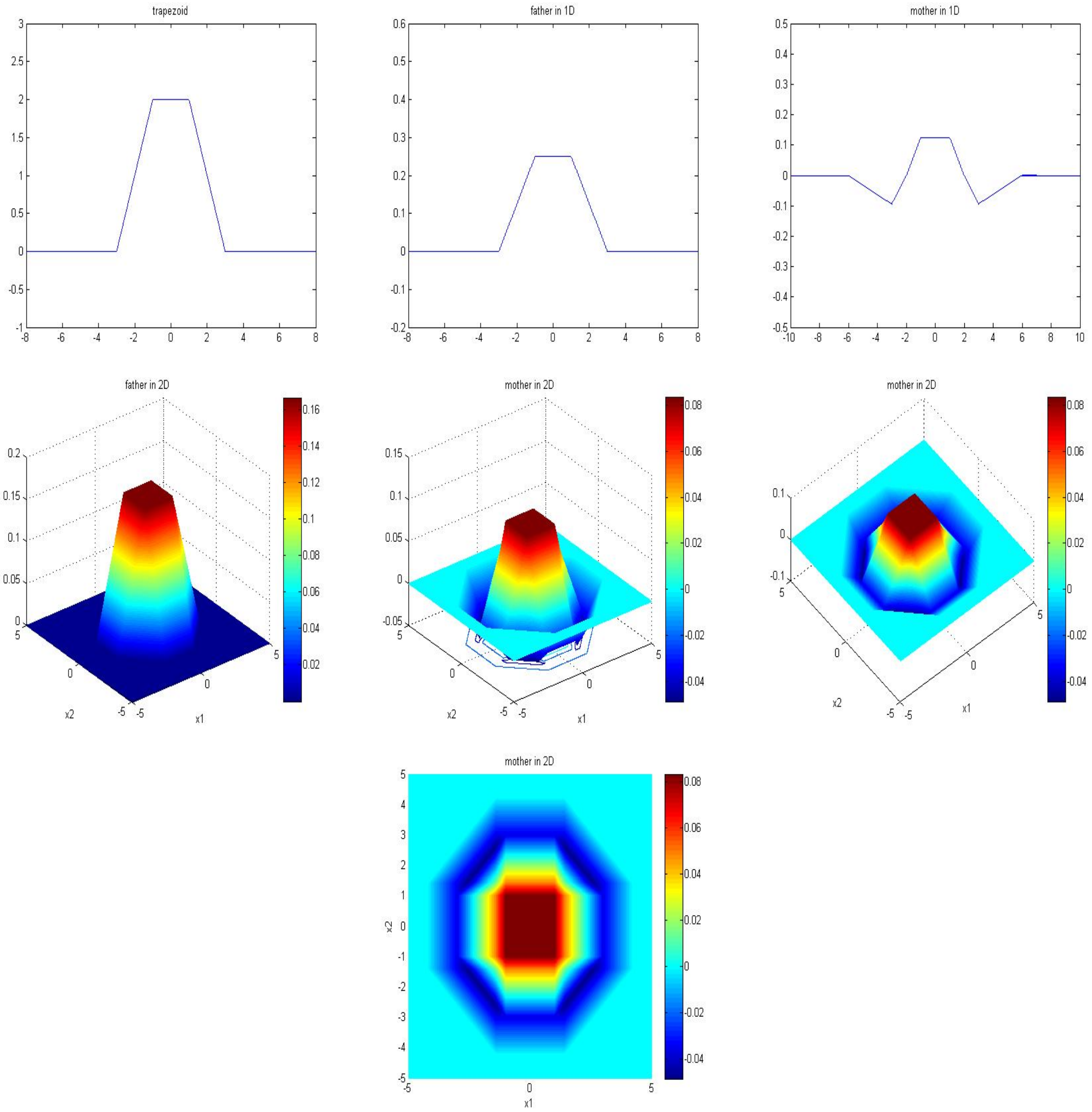

Figure 1: Top row, from left: the trapezoid function t, and the functions φ, ψ on $\mathbb{R}$. Bottom rows: the functions φ, ψ on $\mathbb{R}^2$ from several points of view.

4.2 Creating an atlas

In this section we specify the number of charts that we would like to have to obtain an atlas for a compact d -dimensional manifold $\Gamma \in \mathbb{R}^m$.

For our purpose here we are interested in a small atlas. We would like the size C_Γ of such atlas to depend on the curvature of Γ: the lower the curvature is, the smaller is the number of charts we will need for Γ.

Following the notation of Section 3.1, let $\delta > 0$ so that for all $x \in \Gamma$, $B(x, \delta) \cap \Gamma$ is diffeomorphic to a disc, with a map that is close to the identity. We then cover Γ with balls of radius $\frac{\delta}{2}$. The number of such balls that are required to cover Γ is

$$C_\Gamma \leq \left\lceil \frac{2^d SA(\Gamma)}{\delta^d} T_d \right\rceil, \tag{34}$$

where $SA(\Gamma)$ is the surface area of Γ, and T_d is the thickness of the covering (which corresponds to by how much the balls need to overlap).

Remark 4.6. The thickness T_d scales with d however rather slowly: by [38], there exist covering with $T_d \leq d \log d + 5d$. For example, in $d = 24$ there exist covering with thickness of 7.9.

A covering of Γ by such a collection of balls defines an open cover of Γ by

$$U_i \equiv B(x_i, \delta) \cap \Gamma. \tag{35}$$

Let H_i denote the tangent hyperplane tangent to Γ at x_i. We can now define an atlas by $\{(U_i, \phi_i)\}_{i=1}^{C_\Gamma}$, where ϕ_i is the orthogonal projection from U_i onto H_i.

The above construction is sketched in Figure 2. Let $\tilde{\phi}_i$ be the extension of ϕ_i to $\mathbb{R}^m$, i.e., the orthogonal projection onto H_i. The above construction has two important properties, summarized in Lemma 4.7

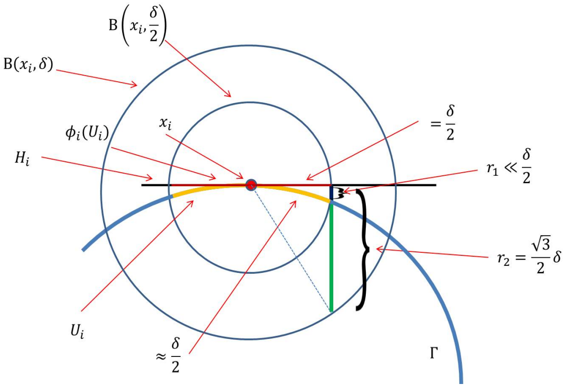
$B\left(x_i, \frac{\delta}{2}\right)$
$B(x_i, \delta)$
$\phi_i(U_i)$
x_i
H_i
$= \frac{\delta}{2}$
$r_1 \ll \frac{\delta}{2}$
$r_2 = \frac{\sqrt{3}}{2}\delta$
U_i
$\approx \frac{\delta}{2}$
Γ

Figure 2: Construction of atlas.

Lemma 4.7. *For every* $x \in U_i$,

$$\|x - \phi_i(x)\|_2 \leq r_1 \leq \frac{\delta}{2} \tag{36}$$

and for every $x \in \Gamma \setminus U_i$ *such that* $\tilde{\phi}_i(x) \in \phi_i(U_i)$

$$\|x - \tilde{\phi}_i(x)\|_2 \geq r_2 = \frac{\sqrt{3}}{2}\delta. \tag{37}$$

4.3 Representing a function on manifold as a sum of functions in $\mathbb{R}^d$

Let Γ be a compact d-dimensional manifold in $\mathbb{R}^m$, let $f : \Gamma \to \mathbb{R}$, let $A = \{(U_i, \phi_i)\}_{i=1}^{C_\Gamma}$ be an atlas obtained by the covering in Section 4.2, and let $\tilde{\phi}_i$ be the extension of ϕ_i to $\mathbb{R}^m$.

$\{U_i\}$ is an open cover of Γ, hence by Theorem 3.4 there exists a corresponding partition of unity, i.e., a family of compactly supported C^∞ functions $\{\eta_i\}_{i=1}^{C_\Gamma}$ such that

- $\eta_i : \Gamma \to [0, 1]$
- $\text{supp}(\eta_i) \subseteq (U_i)$
- $\sum_i \eta_i = 1$

Let

$$f_i \equiv f\eta_i, \tag{38}$$

and observe that $\sum_i f_i = f$. We denote the image $\phi_i(U_i)$ by I_i. Note that $I_i \subset H_i$, i.e., I_i lies in a d-dimensional hyperplane H_i which is isomorphic to $\mathbb{R}^d$. We define $\hat{f}_i$ on $\mathbb{R}^d$ as

$$\hat{f}_i(x) = \begin{cases} f_i(\phi^{-1}(x)) & x \in I_i \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

and observe that $\hat{f}_i$ is compactly supported. This construction gives the following Lemma

Lemma 4.8. *For all $x \in \Gamma$,*

$$\sum_{\{i:x\in U_i\}} \hat{f}_i(\phi_i(x)) = f(x). \tag{40}$$

Assuming $\hat{f}_i \in L_2(\mathbb{R}^d)$, by Lemma 4.4 it has a wavelet expansion using the frame that was constructed in Section 4.1.

4.4 Extending the wavelet terms in the approximation of $\hat{f}_i$ to $\mathbb{R}^m$

Assume that $\hat{f}_i \in L_2(\mathbb{R}^d)$ and let

$$\hat{f}_i = \sum_{(k,b)} \alpha_{k,b}\psi_{k,b}, \tag{41}$$

be its wavelet expansion, where $\alpha_{k,b} \in \mathbb{R}$ and $\psi_{k,b}$ is defined on $\mathbb{R}^d$.

We now show how to extend each $\psi_{k,b}$ to $\mathbb{R}^m$. Let's assume (for now) that the coordinate system is such that the first d coordinates are the local coordinates (i.e., the coordinates on H_i) and the remaining $m - d$ coordinates are of the directions which are orthogonal to H_i.

Intuitively, we would like to extend the wavelet terms on H_i to $\mathbb{R}^m$ so that they remain constant until they "hit" the manifold, and then die off before they "hit" the manifold again.

By Lemma 4.7 it therefore suffices to extend each $\psi_{k,b}$ to $\mathbb{R}^m$ so that in each of the $m-d$ orthogonal directions, $\psi_{k,b}$ will be constant in $[-\frac{r_1}{\sqrt{m-d}}, \frac{r_1}{\sqrt{m-d}}]$ and will have a support which is contained in $[-\frac{r_2}{\sqrt{m-d}}, \frac{r_2}{\sqrt{m-d}}]$.

Recall from Remark 4.2 that each of the wavelet terms $\psi_{k,b}$ in Equation (41) is defined on $\mathbb{R}^d$ by

$$\psi_{k,b}(x) = 2^{\frac{k}{2}} \left(\varphi(2^{\frac{k}{d}}(x-b)) - 2^{-1}\varphi(2^{\frac{k-1}{d}}(x-b))\right) \tag{42}$$

$$\tag{43}$$

and recall that as in Equation (19), the scaling function φ was defined on on $\mathbb{R}^d$ by

$$\varphi(x) = C_d \,\mathrm{rect}\left(\sum_{j=1}^{d} t(x_j) - 2(d-1)\right). \tag{44}$$

We extend $\psi_{k,b}$ to $\mathbb{R}^m$ by

$$\psi_{k,b}(x) \equiv 2^{\frac{k}{2}} \left(\varphi_r(2^{\frac{k}{d}}(x-b)) - 2^{-1}\varphi_r(2^{\frac{k-1}{d}}(x-b))\right), \tag{45}$$

where

$$\varphi_r(2^{\frac{k}{d}}(x-b)) \equiv C_d \,\mathrm{rect}\left(\sum_{j=1}^{d} t(2^{\frac{k}{d}}(x_j-b_j)) + \sum_{j=d+1}^{m} t_r(x_j) - 2(m-1)\right), \tag{46}$$

and t_r is a trapezoid function which is supported on $[-\frac{r_2}{\sqrt{m-d}}, \frac{r_2}{\sqrt{m-d}}]$ and its top (small) base is between $[-\frac{r_1}{\sqrt{m-d}}, \frac{r_1}{\sqrt{m-d}}]$ and has height 2. This definition of $\psi_{k,b}$ gives it a constant height for distance r_1 from H_i, and then a linear decay, until it vanishes at distance r_2. Then by construction we obtain the following lemma

Lemma 4.9. *For every chart (U_i, ϕ_i) and every $x \in \Gamma \setminus U_i$ such that $\tilde{\phi}_i(x) \in \phi_i(U_i)$, x is outside the support of every wavelet term corresponding to the i'th chart.*

Remark 4.10. Since the $m-d$ additional trapezoids in Equation (46) do not scale with k and shift with b, they can be shared across all scaling terms in Equations (45) and (41), so that the extension of the wavelet terms from $\mathbb{R}^d$ to $\mathbb{R}^m$ can be computed with $4(m-d)$ rectifiers.

Finally, in order for this construction to work for all $i = 1, ..., C_\Gamma$ the input $x \in \mathbb{R}^m$ of the network can be first mapped to $\mathbb{R}^{mC_\Gamma}$ by a linear transformation so that the each of the C_Γ blocks of m coordinates gives the local coordinates on Γ in the first d coordinates and on the orthogonal subspace in the remaining $m-d$ coordinates. These maps are essentially the orthogonal projections $\tilde{\phi}_i$.

5 Specifying the required size of the network

In the construction of Section 4, we approximate a function $f \in L_2(\Gamma)$ using a depth 4 network, where the first layer computes the local coordinates in every chart in the atlas, the second layer computes rect functions that are to form trapezoids, the third layer computes scaling functions of the form $\varphi(2^{\frac{k}{d}}(x-b))$ for various k, b and the fourth layer consists of a single node which computes

$$\hat{f} = \sum_{i=1}^{C_\Gamma} \sum_{(k,b)} \psi_{k,b}^{(i)}, \tag{47}$$

where $\psi_{k,b}^{(i)}$ is a wavelet term on the i'th chart. This network is sketched in Figure 3.

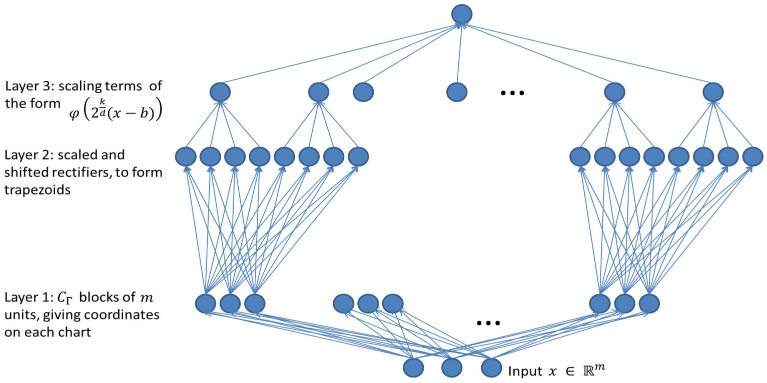
Layer 3: scaling terms of the form $\varphi\left(2^{\frac{k}{d}}(x-b)\right)$
Layer 2: scaled and shifted rectifiers, to form trapezoids
Layer 1: C_Γ blocks of m units, giving coordinates on each chart
...
...
Input $x \in \mathbb{R}^m$

Figure 3: A sketch of the network.

From this construction, we obtain the following theorem, which is the main result of this work:

Theorem 5.1. *Let Γ be a d-dimensional manifold in $\mathbb{R}^m$, and let $f \in L_2(\Gamma)$. Let $\{(U_i, \phi_i)\}$ be an atlas of size C_Γ for Γ, as in Section 4.2. Then f can be approximated using a 4-layer network with mC_Γ linear units in the first hidden layer $8d\sum_{i=1}^{C_\Gamma} N_i + 4C_\Gamma(m-d)$ rectifier units in the second hidden layer, $2\sum_{i=1}^{C_\Gamma} N_i$ rectifier units in the third layer and a single linear unit in the fourth (output) layer, where N_i is the number of wavelet terms that are used for approximating f on the i'th chart.*

Proof. As in Section 4.3, we construct functions $\hat{f}_i$ on $\mathbb{R}^d$ as in Equation (39), which, by Lemma 4.8, have the property that for every $x \in \Gamma$, $\sum_{\{i:x\in U_i\}} \hat{f}_i(\phi_i(x)) = f(x)$. The fact that $\hat{f}_i$ is compactly supported means that its wavelet approximation converges to zero outside $\phi_i(U_i)$. Together with Lemma 4.9, we then get that an approximation of f is obtained by summing up the approximations of all the $\hat{f}_i$'s.

A first layer of the network will consist mC_Γ linear units and will compute the map as in the last paragraph of Section 4.4, i.e., linearly transform the input to C_Γ blocks, each of dimension

m, so that in each block i the first d coordinates are with respect to the tangent hyperplane H_i (i.e., will give the representation $\tilde{\phi}_i(x)$) and the remaining $m-d$ coordinates are with respect to directions orthogonal to H_i.

For each $i = 1, .., C_\Gamma$, we approximate each $\hat{f}_i$ to some desired approximation level δ using $N_i < \infty$ wavelet terms. By Remark 4.3, $\hat{f}_i$ can be approximated using $8dN_i$ rectifiers in the second layer, $2N_i$ rectifiers in the third layer and a single unit in the fourth layer. By Remark 4.10, on every chart the wavelet terms in all scales and shifts can be extended to $\mathbb{R}^m$ using (the same) $4(m-d)$ rectifiers in the second layer.

Putting this together we get that to approximate f one needs a 4-layer network with mC_Γ linear units in the first hidden layer $8d\sum_{i=1}^{C_\Gamma} N_i + 4C_\Gamma(m-d)$ rectifier units in the second hidden layer, $2\sum_{i=1}^{C_\Gamma} N_i$ rectifier units in the third layer and a single linear unit in the fourth (output) layer. □

Remark 5.2. For sufficiently small radius δ in the sense of section 3.1, the desired properties of $\hat{f}_i$ (i.e., being in L_2 and possibly having sparse coefficients or being twice differentiable) imply similar properties of f.

Remark 5.3. We observe that the dependence on the dimension m of the ambient space in the first and second layers is through C_Γ, which depends on the curvature of the manifold. The number N_i of wavelet terms in the i'th chart affects the number of units in the second layer only through the dimension d of the manifold, not through m. The sizes of the third and fourth layers do not depend on m at all.

Finally, assuming regularity conditions on the $\hat{f}_i$, allows us to bound the number N_i of wavelet terms needed for the approximation of $\hat{f}_i$. In particular, we consider two specific cases: $\hat{f}_i \in \mathcal{L}_1$ and $\hat{f}_i \in C^2$, with bounded second derivative.

Corollary 5.4. *If $\hat{f}_i \in \mathcal{L}_1$ (i.e., $\hat{f}_i$ has expansion coefficients in l_1), then by Theorem 3.10, $\hat{f}_i$ can be approximated by a combination $\hat{f}_{i,N_i}$ of N_i wavelet terms so that*

$$\|\hat{f}_i - \hat{f}_{i,N_i}\|_2 \leq \frac{\|\hat{f}_i\|_{\mathcal{L}_1}}{\sqrt{N_i+1}}. \tag{48}$$

Consequently, denoting the output of the net by $\tilde{f}$, $N \equiv \max_i\{N_i\}$ and $M \equiv \max_i \|\hat{f}_i\|_{\mathcal{L}_1}$, we obtain

$$\|f - \tilde{f}\|_2^2 \leq \frac{C_\Gamma M}{N+1}, \tag{49}$$

using $c_1 + c_2 N$ units, where $c_1 = C_\Gamma(m + 4(m-d)) + 1$ and $c_2 = (8d+2)C_\Gamma$.

Corollary 5.5. *If for each i $\hat{f}_i$'s is twice differentiable and $\|\nabla^2_{\hat{f}_i}\|_{op}$ is bounded, then by Lemma 4.5 $\hat{f}_i$ can be approximated by $\hat{f}_{K,i}$ using all terms up to scale K so that for every $x \in \mathbb{R}^d$*

$$|\hat{f}_i(x) - \hat{f}_{i,K}(x)| = O\left(2^{-\frac{2K}{d}}\right). \tag{50}$$

Observe that the grid spacing in the k'th level is $2^{-\frac{k}{d}}$*. Therefore, since* f *is compactly supported, there are* $O\left(\left(2^{\frac{k}{d}}\right)^{d}\right) = O\left(2^{k}\right)$ *terms in the k'th level. Altogether, on the i'th chart there are* $O\left(2^{K+1}\right)$ *terms in levels less than* K*. Writing* $N \equiv 2^{K+1}$*, we get a point-wise error rate of* $N^{-\frac{2}{d}}$ *using* $c_1 + c_2 N$ *units, where* $c_1 = C_\Gamma(m + 4(m-d)) + 1$ *and* $c_2 = (8d+2)C_\Gamma$*.*

Remark 5.6. The unit count in Theorem 5.1 and Corollaries 5.4 and 5.5 is overly pessimistic, in the sense that we assume that the sets of wavelet terms in the expansion of $\hat{f}_i$, $\hat{f}_j$ do not intersect, where i, j are chart indices. A tighter bound can be obtained if we allow wavelet functions be shared across different charts, in which case the term $C_\Gamma \sum N_i$ in Theorem 5.1 can be replaced by the total number of distinct wavelet terms that are used on all charts, hence decreasing the constant c_2. In particular, in Corollary 5.5 we are using all terms up to the K'th scale on each chart. In this case the constant $c_2 = 8d + 2$.

Remark 5.7. The linear units in the first layer can be simulated using ReLU units with large positive biases, and adjusting the biases of the units in the second layer. Hence the first layer can contain ReLU units instead of linear units.

6 Conclusions

The construction presented in this manuscript can be divided to two main parts: analytical and topological. In the analytical part, we constructed a wavelet frame if $L_2(\mathbb{R}^d)$, where the wavelets are computed from Rectified Linear units. In the topological part, given training data on a d-dimensional manifold Γ we constructed an atlas and represented any function on Γ as sum of functions that are defined on the charts. We then used Rectifier units to extend the wavelet approximation of the functions from $\mathbb{R}^d$ to the ambient space $\mathbb{R}^m$. This construction allows us to state the size of a depth 4 neural net given a function f to be approximated on the manifold Γ. We show how the specified size depends on the complexity of the function (manifested in the number of wavelet terms in its approximation) and the curvature of the manifold (manifested in the size of the atlas). In particular, we take advantage of the fact that d can possibly be much smaller than m to construct a network with size that depends more strongly on d. In addition, we also obtained squared error rate in approximation of functions with sparse wavelet expansion and point-wise error rate for twice differentiable functions.

We are currently working on several extensions of this work. First, a more efficient wavelet representation can be obtained on each chart if one allows its wavelets to be non-isotropic and not necessarily axis aligned, but rather, to correspond to the level sets of the function being approximated. When the function is relatively constant in certain directions, the wavelet terms can be "stretched" in these directions. Such thing can be done using curvelets.

Second, we conjecture that in the representation obtained as an output of convolutional layers, the data concentrates near a collection of low dimensional manifolds embedded in a high dimensional space, which is our starting point in the current manuscript. In particular, we conjecture that the collection of manifolds corresponds to the collection of classes. We think that this is a result of the application of the same filters to all data points. Assuming our conjecture is true, one can apply our construction to the output of convolutional layers, and by that obtain a network topology which is similar to standard convolutional networks, namely fully connected layers on top of convolutional ones. This will make or arguments here applicable to cases where the data in its initial representation does not concentrate near low dimensional manifold, but its hidden representation does.

Finally, we remark that the choice of using rectifier units to construct our wavelet frame is convenient, however somewhat arbitrary. Similar wavelet frames can be constructed by any function (or combination of functions) that can be used to construct "bump" functions i.e., functions which are localized and have fast decay. For example, general sigmoid functions $\sigma : \mathbb{R} \to \mathbb{R}$, which are monotonic and have the properties

$$\lim_{x\to-\infty} \sigma(x) = 0 \text{ and } \lim_{x\to\infty} \sigma(x) = 1 \tag{51}$$

can used to construct a frame in a similar way, by computing "smooth" trapezoids. Recall also that by Remark 3.11, any two such frames are equivalent.

Acknowledgements

The authors thank Stefan Steinerberger, Roy Lederman for their help, and to Andrew Barron, Mark Tygert and Yann LeCun for their comments. Alexander Cloninger is supported by NSF Award No. DMS-1402254.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *arXiv preprint arXiv:1409.4842*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” *arXiv preprint arXiv:1502.01852*, 2015.

[4] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616, ACM, 2009.

[5] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3642–3649, IEEE, 2012.

[6] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, “The manifold tangent classifier,” in *Advances in Neural Information Processing Systems*, pp. 2294–2302, 2011.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[8] G. Dahl, A.-r. Mohamed, G. E. Hinton, *et al.*, “Phone recognition with the mean-covariance restricted boltzmann machine,” in *Advances in neural information processing systems*, pp. 469–477, 2010.

[9] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in neural information processing systems*, pp. 1096–1104, 2009.

[10] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 513–520, 2011.

[11] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[12] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[13] F. Girosi and T. Poggio, "Networks and the best approximation property," *Biological cybernetics*, vol. 63, no. 3, pp. 169–176, 1990.

[14] F. Girosi, M. B. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural computation*, vol. 7, no. 2, pp. 219–269, 1995.

[15] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural networks*, vol. 6, no. 6, pp. 861–867, 1993.

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

[17] H. N. Mhaskar, "On the tractability of multivariate integration and approximation by neural networks," *Journal of Complexity*, vol. 20, no. 4, pp. 561–590, 2004.

[18] V. Kurková and M. Sanguineti, "Comparison of worst case errors in linear and neural network approximation," *IEEE Transactions on Information Theory*, vol. 48, no. 1, pp. 264–275, 2002.

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

[20] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[21] N. Le Roux and Y. Bengio, "Deep belief networks are compact universal approximators," *Neural computation*, vol. 22, no. 8, pp. 2192–2207, 2010.

[22] G. Montufar and N. Ay, "Refinements of universal approximation results for deep belief networks and restricted boltzmann machines," *Neural Computation*, vol. 23, no. 5, pp. 1306–1319, 2011.

[23] Q. Zhang and A. Benveniste, "Wavelet networks," *Neural Networks, IEEE Transactions on*, vol. 3, no. 6, pp. 889–898, 1992.

[24] Y. C. Pati and P. S. Krishnaprasad, "Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations," *Neural Networks, IEEE Transactions on*, vol. 4, no. 1, pp. 73–85, 1993.

[25] J. Zhao, B. Chen, and J. Shen, "Multidimensional non-orthogonal wavelet-sigmoid basis function neural network for dynamic process fault diagnosis," *Computers & chemical engineering*, vol. 23, no. 1, pp. 83–92, 1998.

[26] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Advances in Neural Information Processing Systems*, pp. 2924–2932, 2014.

[27] A. B. Patel, T. Nguyen, and R. G. Baraniuk, "A probabilistic theory of deep learning," *arXiv preprint arXiv:1504.00641*, 2015.

[28] A. Paul and S. Venkatasubramanian, "Why does unsupervised deep learning work?-aperspective from group theory,"

[29] P. Mehta and D. J. Schwab, "An exact mapping between the variational renormalization group and deep learning," *arXiv preprint arXiv:1410.3831*, 2014.

[30] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *arXiv preprint arXiv:1503.02406*, 2015.

[31] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "Provable bounds for learning some deep representations," *arXiv preprint arXiv:1310.6343*, 2013.

[32] R. Livni, S. Shalev-Shwartz, and O. Shamir, "On the computational efficiency of training neural networks," in *Advances in Neural Information Processing Systems*, pp. 855–863, 2014.

[33] J. Bruna, S. Chintala, Y. LeCun, S. Piantino, A. Szlam, and M. Tygert, "A theoretical argument for complex-valued convolutional networks," *arXiv preprint arXiv:1503.03438*, 2015.

[34] C. K. Chui and H. Mhaskar, "Deep nets and manifold learning," *Personal Communication*, 2015.

[35] W. T. Loring, "An introduction to manifolds," 2008.

[36] D. Deng and Y. Han, *Harmonic analysis on spaces of homogeneous type*. No. 1966, Springer Science & Business Media, 2009.

[37] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore, "Approximation and learning by greedy algorithms," *The annals of statistics*, pp. 64–94, 2008.

[38] J. H. Conway, N. J. A. Sloane, E. Bannai, J. Leech, S. Norton, A. Odlyzko, R. Parker, L. Queen, and B. Venkov, *Sphere packings, lattices and groups*, vol. 3. Springer-Verlag New York, 1993.

A Equivalence of representations in different wavelet frames

Consider to frames $\{\psi_{k,b}\}$ and $\{\psi'_{k,b}\}$. Any element $\psi'_{k',b'}$ can be represented as

$$\psi'_{k',b'} = \sum_{k,b} \langle \psi'_{k',b'}, \widetilde{\psi}_{k,b} \rangle \psi_{k,b}. \tag{52}$$

Observe that in case $k \approx k'$, the inner product is of large magnitude only for a small number of b's. In case $k \ll k'$ or $k \gg k'$, the inner product is between peaked function which integrates to zero and a flat function, hence has small magnitude. This idea is formalized in a more general form in Section 4.7 in [36].

B Proof of Lemma 4.1

.

Proof. In order to show that the family $\{S_k\}$ in Equation (21) is a valid family of averaging kernel functions, we need to verify that conditions $3.14 - 3.19$ in [36] are satisfied. Here $\rho(x,b)$ is the volume of the smallest Euclidean ball which contains x and b, namely $\rho(x,b) = c\|x-b\|^d$, for some constant c. Our goal is to show that there exist constants $C \leq \infty$, $\sigma > 0$ and $\epsilon > 0$ such that for every $k \in \mathbb{Z}$, and $x, x', b, b' \in \mathbb{R}^d$

- 3.14:

$$S_k(x,b) \leq C \frac{2^{-k\epsilon}}{(2^{-k} + \rho(x,b))^{1+\epsilon}}, \tag{53}$$

Proof. WLOG we can assume $b = 0$, and let ϵ be arbitrary positive number. It can be easily verified that there exists a constant C' such that

$$\varphi(x) \leq \frac{C'}{\left(c^{-1} + \|x\|^d\right)^{1+\epsilon}}. \tag{54}$$

Then

$$S_k(x,0) = 2^k \varphi\left(2^{\frac{k}{d}} x\right) \tag{55}$$

$$\leq C' \frac{2^k}{\left(c^{-1} + 2^k\|x\|^d\right)^{1+\epsilon}} \tag{56}$$

$$= C' \frac{2^{k(1+\epsilon)} 2^{-k\epsilon}}{\left(c^{-1} + 2^k\|x\|^d\right)^{1+\epsilon}} \tag{57}$$

$$= C' \frac{2^{-k\epsilon}}{\left(c^{-1}2^{-k} + \|x\|^d\right)^{1+\epsilon}} \tag{58}$$

$$= C_1 \frac{2^{-k\epsilon}}{\left(2^{-k} + \rho(x,0)\right)^{1+\epsilon}}, \tag{59}$$

where $C_1 = c^{1+\epsilon}C'$. □

- 3.15, 3.16: Since $S_k(x,b)$ depends only on $x-b$ and is symmetric about the origin, it suffices to prove only 3.15. We want to show that if $\rho(x,x') \leq \frac{1}{2A}(2^{-k}+\rho(x,b))$ then

$$|S_k(x,b) - S_k(x',b)| \leq C\left(\frac{\rho(x,x')}{2^{-k}+\rho(x,b)}\right)^{\sigma}\frac{2^{-k\epsilon}}{(2^{-k}+\rho(x,b))^{1+\epsilon}}. \tag{60}$$

Proof. WLOG $b=0$; we will prove for every x,x'. Let ϵ be arbitrary positive number, and let $\sigma = \frac{1}{d}$. By the mean value theorem we get

$$\frac{|S_k(x,0) - S_k(x',0)|}{\rho(x,x')^{\sigma}} \leq \max_{z_k \text{ between } x,x'} \frac{1}{c}\|\nabla_x(S_k(z_k,0))\|. \tag{61}$$

Denote

$$F(x) \equiv \|\nabla_x(S_0(x,0))\|. \tag{62}$$

Then

$$\|\nabla_x(S_k(x,0))\| = 2^k 2^{\frac{k}{d}} F\left(2^{\frac{k}{d}}x\right). \tag{63}$$

As in the proof of condition 3.14, it can be easily verified that there exists a constant C' such that

$$F(x) \leq C'\frac{1}{(c^{-1}+\|x\|^d)^{\sigma}}\frac{1}{(c^{-1}+\|x\|^d)^{1+\epsilon}}. \tag{64}$$

We then get

$$\begin{aligned}
\frac{|S_k(x,b) - S_0(x',b)|}{\rho(x,x')^{\sigma}} &= \frac{1}{c}\|\nabla_x(S_k(z_k,0))\| && (65)\\
&= 2^k 2^{\frac{k}{d}} F\left(2^{\frac{k}{d}}\right) && (66)\\
&\leq C'\frac{2^{\frac{k}{d}}}{(c^{-1}+2^k\|x\|^d)^{\sigma}}\frac{2^k}{(c^{-1}+2^k\|x\|^d)^{1+\epsilon}} && (67)\\
&= C'\frac{2^{\frac{k}{d}}}{(c^{-1}+2^k\|x\|^d)^{\sigma}}\frac{2^{k(1+\epsilon)}2^{-k\epsilon}}{(c^{-1}+2^k\|x\|^d)^{1+\epsilon}} && (68)\\
&= C'\frac{1}{(c^{-1}2^{-k}+\|x\|^d)^{\sigma}}\frac{2^{-k\epsilon}}{(c^{-1}2^{-k}+\|x\|^d)^{1+\epsilon}} && (69)\\
&= C_2\frac{1}{(2^{-k}+\rho(x,0))^{\sigma}}\frac{2^{-k\epsilon}}{(2^{-k}+\rho(x,0))^{1+\epsilon}}, && (70)
\end{aligned}$$

where $C_2 = c^{\sigma+1+\epsilon}C'$. □

- 3.17, 3.18: Since $S_k(x,b)$ depends only on $x-b$ and is symmetric about the origin, it suffices to prove only 3.17.

Proof. By Equation (19)

$$\int_{\mathbb{R}^d} \varphi(x)dx = 1 \tag{71}$$

and consequently for every $k \in \mathbb{Z}$ and $b \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} S_k(x,b)dx = 1. \tag{72}$$

□

- 3.19: we want to show if $\rho(x,x') \le \frac{1}{2A}(2^{-k}+\rho(x,b))$ and $\rho(b,b') \le c(2^{-k}+\rho(x,b))$ then

$$|S_k(x,b) - S_k(x',b) - S_k(x,b') + S_k(x',b')| \tag{73}$$

$$\le C\left(\frac{\rho(x,x')}{2^{-k}+\rho(x,b)}\right)^{\sigma}\left(\frac{\rho(b,b')}{2^{-k}+\rho(x,b)}\right)^{\sigma}\frac{2^{-k\epsilon}}{(2^{-k}+\rho(x,b))^{1+\epsilon}}. \tag{74}$$

Proof. We will prove for all x, x', b, b'. Let $\sigma = \frac{1}{d}$. Observe that

$$\frac{|S_k(x,b) - S_k(x',b) - S_k(x,b') + S_k(x',b')|}{\rho(x,x')^{\sigma}\rho(b,b')^{\sigma}} \tag{75}$$

$$\le \frac{|\frac{|S_k(x,b)-S_k(x',b)|}{\rho(x,x')^{\sigma}} + \frac{|S_k(x,b')+S_k(x',b')|}{\rho(x,x')^{\sigma}}|}{\rho(b,b')^{\sigma}} \tag{76}$$

$$\tag{77}$$

Denote

$$F(b) \equiv \frac{|S_k(x,b) - S_k(x',b)|}{\rho(x,x')^{\sigma}}. \tag{78}$$

Then by applying the mean value theorem twice we get

$$\frac{|\frac{|S_k(x,b)-S_k(x',b)|}{\rho(x,x')^{\sigma}} + \frac{|S_k(x,b')+S_k(x',b')|}{\rho(x,x')^{\sigma}}|}{\rho(b,b')^{\sigma}} \tag{79}$$

$$= \frac{|F(b)-F(b')|}{\rho(b,b')^{\sigma}} \tag{80}$$

$$\frac{1}{c} \le \max_{z \text{ between } b,b'} \nabla_b(F(z)) \tag{81}$$

$$= \frac{1}{c}\max_{z \text{ between } b,b'} \nabla_b\left(\frac{|S_k(x,z) - S_k(x',z)|}{\rho(x,x')^{\sigma}}\right) \tag{82}$$

$$\frac{1}{c^2} \le \max_{z \text{ between } b,b'} \max_{z' \text{ between } x,x'} \|\nabla^2_{x,b}(S_k(z',z))\| \tag{83}$$

From this, we can see that Since S_k is compactly supported and bounded, there exist compactly supported function $\xi(x)$ such that

$$\frac{|S_0(x,b)-S_0(x',b)-S_0(x,b')+S_0(x',b')|}{\rho(x,x')^\sigma\rho(b,b')^\sigma} \tag{84}$$

$$\leq \xi(x-b)+\xi(x-b'), \tag{85}$$

and consequently

$$\frac{|S_k(x,b)-S_k(x',b)-S_k(x,b')+S_k(x',b')|}{\rho(x,x')^\sigma\rho(b,b')^\sigma}| \tag{86}$$

$$\leq 2^k 2^{\frac{2k}{d}}\left(\xi\left(2^{\frac{k}{d}}(x-b)\right)+\xi\left(2^{\frac{k}{d}}(x-b')\right)\right). \tag{87}$$

As in the proof of conditions $3.14, 3.15$, there exists a constant C' such that

$$\xi(x-b)+\xi(x-b') \leq C'\frac{1}{(c^{-2}+\|x-b\|^d)^{2\sigma}}\frac{1}{(c^{-1}+\|x-b\|^d)^{1+\epsilon}}. \tag{88}$$

We then get

$$\frac{|S_k(x,b)-S_k(x',b)-S_k(x,b')+S_k(x',b')|}{\rho(x,x')^\sigma\rho(b,b')^\sigma} \tag{89}$$

$$\leq 2^k 2^{\frac{2k}{d}}\left(\xi\left(2^{\frac{k}{d}}(x-b)\right)+\xi\left(2^{\frac{k}{d}}(x-b')\right)\right) \tag{90}$$

$$\leq C'\frac{2^{\frac{2k}{d}}}{(c^{-2}+2^k\|x-b\|^d)^{2\sigma}}\frac{2^k}{(c^{-1}+2^k\|x-b\|^d)^{1+\epsilon}} \tag{91}$$

$$= C'\frac{1}{(c^{-2}2^{-k}+\|x-b\|^d)^{2\sigma}}\frac{2^{-k\epsilon}}{(c^{-1}2^{-k}+\|x-b\|^d)^{1+\epsilon}} \tag{92}$$

$$= C_3\frac{1}{(2^{-k}+\rho(x,b))^{2\sigma}}\frac{2^{-k\epsilon}}{(2^{-k}+\rho(x,b))^{1+\epsilon}}, \tag{93}$$

where $C_3 = c^{2\sigma+1+\epsilon}$. □

Finally, we set $C = \max\{C_1, C_2, C_3\}$. □

C Proof of Lemma 4.5

We first prove the following propositions.

Proposition C.1. *For each k, b, $\psi_{k,b}$, $\widetilde{\psi}_{k,b}$ have two vanishing moments.*

Proof. Note that a function f on $\mathbb{R}^d$ which is symmetric about the origin satisfies

$$\int_{\mathbb{R}^d} xf(x)dx = 0. \tag{94}$$

We first show that for every $(k, b) \in \Lambda$, $\psi_{k,b}$ has two vanishing moments. For each $(k, b) \in \mathbb{Z}\times\mathbb{R}^d$

$$\begin{aligned} 2^{-k} \int_{\mathbb{R}^d} \varphi(2^{\frac{k}{d}}(x-b))dx &= \int_{\mathbb{R}^d} \varphi(x)dx \qquad (95) \\ &= 1, \qquad (96) \end{aligned}$$

by change of variables. This gives that for every $(k, b) \in \mathbb{Z} \times \mathbb{R}^d$

$$\begin{aligned} \int_{\mathbb{R}^d} \psi_{k,b}(x)dx &= 2^{\frac{k}{2}} \int_{\mathbb{R}^d} \varphi(2^{\frac{k}{d}}(x-b) - \varphi\left(2^{\frac{k-1}{d}}(x-b)\right) dx \qquad (97) \\ &= 0, \qquad (98) \end{aligned}$$

Hence the first moment of $\psi_{k,b}$ vanishes. Further, since φ is symmetric about the origin we have

$$\begin{aligned} \int_{\mathbb{R}^d} x\varphi\left(2^{\frac{k}{d}}(x-b)\right) dx &= \int_{\mathbb{R}^d} (2^{-\frac{k}{d}}y + b)\varphi(y)dy \qquad (99) \\ &= 2^{-k}b \int_{\mathbb{R}^d} \varphi(y)dy \qquad (100) \\ &= 2^{-k}b, \qquad (101) \end{aligned}$$

which gives

$$\begin{aligned} \int_{\mathbb{R}^d} x\psi_{k,b}(x)dx &= 2^{-\frac{k}{2}} \int_{\mathbb{R}^d} \varphi\left(2^{\frac{k}{d}}(x-b)\right) - 2^{-1}\varphi\left(2^{\frac{k-1}{d}}(x-b)\right) dx \qquad (102) \\ &= 2^{-\frac{k}{2}}\left(2^{-k}b - 2^{-1}2^{-(k-1)}b\right) \qquad (103) \\ &= 2^{-\frac{k}{2}}\left(2^{-k}b - 2^{-k}b\right) \qquad (104) \\ &= 0, \qquad (105) \end{aligned}$$

hence the second moment of $\psi_{k,b}$ also vanishes.

Finally, to show that the functions $\widetilde{\psi}_{k,b}$ have two vanishing moments as well, we note that the dual functions are obtained using convolution with operators D_k ([36], p. 82), which, by the above arguments, have two vanishing moments; hence they inherit this property. □

Proposition C.2. *For every (k, b), $\widehat{\psi}_{k,b}$ decays faster than any polynomial.*

Proof. By ([36], p. 82), the dual functions are also wavelets, hence they satisfy condition 3.14 in [36] with $\epsilon' < \epsilon$. Since in the proof of Lemma 4.1, ϵ can be arbitrarily large, it implies that the duals satisfy condition 3.14 with any ϵ, which proves the proposition. □

Proposition C.3. $|\psi_{k,b}| \leq 2^{\frac{k}{2}-2}$.

Proof. We note that for all $d \geq 2$, $C_d \leq \frac{1}{2 \cdot 2^d} \leq \frac{1}{8}$. Hence $\varphi(x) \leq \frac{1}{4}$, and consequently $|\psi(x)| \leq \frac{1}{4}$. Since

$$\psi_{k,b}(x) = 2^{\frac{k}{2}} \psi \left(2^{\frac{k}{d}} x - b) \right) \tag{106}$$

we get that $|\psi_{k,b}| \leq 2^{\frac{k}{2}-2}$. □

Proposition C.4. *if $f \in C^2$ and $\|\nabla_f^2\|_{op}$ is bounded, then The coefficients $\langle \widetilde{\psi}_{k,b}, f \rangle$ satisfy*

$$|\langle \widetilde{\psi}_{k,b}, f \rangle| = O(2^{-(2\frac{k}{d}+\frac{k}{2})}) \tag{107}$$

Proof.

$$\begin{aligned} \langle \widetilde{\psi}_{k,b}, f \rangle &= 2^{\frac{k}{2}} \int_{\mathbb{R}^d} \widetilde{\psi} \left(2^{\frac{k}{d}} (x-b) \right) f(x) dx & (108) \\ &= 2^{-\frac{k}{2}} \int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) f(2^{-\frac{k}{d}} y + b) dy. & (109) \end{aligned}$$

where we have used change of variables. Since that f is twice differentiable, we can replace f by its Taylor expansion near b

$$\begin{aligned} &\int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) f(2^{-\frac{k}{d}} y + b) dy & (110) \\ =&\int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) \left(f(b) + 2^{-\frac{k}{d}} \langle y, \nabla_f(b) \rangle + O(\|\nabla_f^2(b)\|_{op} (2^{-\frac{k}{d}} \|y\|_2)^2) \right) dy. & (111) \end{aligned}$$

By Proposition C.1 $\widetilde{\psi}$ has two vanishing moments; this gives

$$|\langle \widetilde{\psi}_{k,b}, f \rangle| = O\left(2^{-(2\frac{k}{d}+\frac{k}{2})} \|\nabla_f^2(b)\|_{op} \int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) \|y\|_2^2 dy \right) \tag{112}$$

Since by Proposition C.2 $\widetilde{\psi}(y)$ decays exponentially fast, the integral $\int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) \|y\|_2^2 dy$ is some finite number. As a result,

$$|\langle \widetilde{\psi}_{k,b}, f \rangle| = O(2^{-(2\frac{k}{d}+\frac{k}{2})}). \tag{113}$$

□

We will also use the following property:

Remark C.5. Every x is in the support of at most 12^d wavelet terms at every scale.

We are now ready to prove Lemma 4.5

Proof. Let $f \in L_2(\mathbb{R}^d)$, $d \leq 3$ be compactly supported, twice differentiable and with $\|\nabla_f^2\|_{op}$ bounded. f can be expressed as

$$f = \sum_{(k,b)\in\Lambda} \langle \widetilde{\psi}_{k,b}, f\rangle \psi_{k,b}. \tag{114}$$

approximating f by f_K, which only consists of the wavelet terms of scales $k \leq K$, we obtain that for every $x \in \mathbb{R}^d$

$$|f(x) - f_K(x)| \leq \sum_{k=K+1}^{\infty} \sum_{b\in 2^{-k}\mathbb{Z}} |\psi_{k,b}| \langle \widetilde{\psi}_{k,b}, f\rangle. \tag{115}$$

By Remark C.5, at most 12^d wavelet terms are supported on x at every scale; by Proposition C.3 $|\psi_{k,b}| \leq 2^{\frac{k}{2}-2}$; by Proposition C.4 $|\langle \widetilde{\psi}_{k,b}, f\rangle| = O(2^{-(\frac{2k}{d}+\frac{k}{2})})$. Plugging these into Equation (115) gives

$$\begin{aligned} |f(x) - f_K(x)| &= O\left(\sum_{k=K+1}^{\infty} 12^d 2^{\frac{k}{2}-2} 2^{-(\frac{2k}{d}+\frac{k}{2})}\right) && (116)\\ &= O\left(\sum_{k=K+1}^{\infty} 2^{-\frac{2k}{d}}\right) && (117)\\ &= O\left(2^{-\frac{2K}{d}}\right). && (118) \end{aligned}$$

□