We discuss approximation properties of deep neural nets, in the case that the data concentrates near a $d$-dimensional manifold $\Gamma \in \mathbb{R}^m$. Our network essentially computes wavelet functions, which are computed from Rectified Linear Units (ReLU). Given a squared integrable function $f$ and a manifold $\Gamma$, we specify the size of a depth 4 network that approximates $f$ on $\Gamma$. We take advantage of the possibility that $d \ll m$ to construct a network in which the number of hidden units depends mostly on $d$, rather than on $m$. In addition, the network's size depends also the complexity of $f$, and the curvature of $\Gamma$. For two specific function classes, functions with sparse wavelet coefficients and $C^2$ functions we also obtain error bounds.

# 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 size 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, i.e., given a target function $f : \mathbb{R}^m \to \mathbb{R}$, we aim to specify a Multi Level Perceptron (MLP) network topology that, given optimal weights, achieves desired approximation level. We essentially construct a wavelet network, which is based on modern deep learning practice, i.e., the network has depth 4, and the wavelet functions are computed using Rectified Linear functions. In addition, We assume that the training data concentrates near a $d$-dimensional manifold in $\mathbb{R}^m$, which seems to be a reasonable assumption in many practical applications. We specify the network topology in terms of the manifold dimension $d$, the ambient dimension $m$, and the complexity of the function, which is manifested through the number of dictionary elements that are needed to approximate it. In particular, the specified number of units depends only weakly on the dimension $m$ of the ambient space and more strongly on the dimension $d$ of the manifold. 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 $C^2$. In particular, we will prove a more detailed version of the following theorem

**Theorem 1.1.** *Let $\Gamma \subset \mathbb{R}^m$ be a smooth $d$-dimensional manifold, and let $f \in L_2(\Gamma)$. Let $C_\Gamma$ be the size of a sufficiently small atlas for $\Gamma$ so that $\Gamma$ can be approximated using $C_\Gamma$ $d$-dimensional hyperplanes. Then*

- *if $f$ can be represented as a sum of $C_\Gamma$ functions with wavelet coefficients in $l_1$ then there exists a depth 4 network with $c_1 + c_2 N$ Rectifier and linear units which approximate $f$ by $f_N$ so that*

$$\|f - f_N\|_2 \leq \frac{c}{\sqrt{N}} \tag{1}$$

- *if $f$ can be represented as a sum of $C_\Gamma$ $C^2$ functions with bounded second derivative then there exists a depth 4 network with $c_1 + c_2 N$ Rectifier and linear units which approximate $f$ by $f_N$ so that for every $x \in \Gamma$*

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

*where $c_1 = mC_\Gamma + 4(m-d)+1$, $c_2 = (8d+2)C_\Gamma$ and $c$ is some constant. For general functions $f \in L_2(\Gamma)$ we get that if $f$ can be approximated using a depth 4 network of rectifier and linear units with total of $c_1 + c_2 \sum_{i=1}^{C_\Gamma} k_i$ units, where $k_i$ is the number of wavelet terms taken to approximate $f$ on the i'th chart.*

### 1.2 The structure of this manuscript

The structure of this manuscript is as follows: in Section 2 we review some of the fundamental 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 briefly concludes this manuscript.

### 1.3 Notation

$\Gamma$ denotes a $d$-dimensional manifold in $\mathbb{R}^m$. $\{(U_i, \phi_i)\}$ denotes an atlas for $\Gamma$. Tangent hyperplanes to $\Gamma$ are denoted by $H_i$. $f$ and variants of it stand for the function to be approximated. $\varphi, \psi$ are scaling and wavelet 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

A well known result by Hornik, Stinchcombe and White [11] states that an ANN with a single, possibly huge, hidden layer can approximate any Borel measurable function to an arbitrary degree of precision. A similar result was proved independently by Cybenko [12] for continuous compactly supported functions, using a single layer network of sigmoidal units. Barron [13] showed that given a function $f : \mathbb{R}^m \to \mathbb{R}$ with bounded first moment of the magnitude of the Fourier transform

$$C_f = \int_{\mathbb{R}^m} |w||\tilde{f}(w)| < \infty \tag{3}$$

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

$$\|f - f_N\|_2^2 \le \frac{c_f}{N}, \tag{4}$$

where $c_f$ is proportional to $C_f$. Approximation of continuous and $L_2$ functions through two and three layer networks containing McCulloch-Pitts (Heaviside) units (hence computing piecewise linear functions) is discussed in In [14]. Approximation results using networks containing Heaviside units are obtained also in [15].

The question of determining the appropriate number of training examples required to train a network of a given architecture can be approached via learning theory, where the expressiveness of a class of functions that can be represented using the given network architecture is measured in terms of the Vapnik-Chervonenkis (VC) dimension, in case of classification, and its extension for regression, the fat shattering dimension (see, for example, [16] and [17]).

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 [18], [19] and [20]).

Among a few recent theoretical results, in [21], 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. Arora et al. [22], provide algorithms with provable guarantees for learning in networks with sparse connectivity and random weights in $[-1, 1]$. In another interesting recent paper, Livni et al. [23] analyze of expressiveness of neural nets in terms of Turing machines and provide a provably correct algorithm for training polynomial nets of depth 2 and 3. Sedghi et.al show how the first layer weight matrix can be recovered in networks that have sparse connectivity. Patel et. al [24] give a probabilistic framework of deep learning. In particular, they describe convolutional nets in terms of message passing algorithms. Complex valued convolutional nets are proposed in [25]. Analysis of unsupervised pre-training from group-theoretic perspective is proposed in [26]; their theory also explains why higher layers tend to learn more abstract features, a well-observed phenomenon in deep learning practice. [27] provides analysis of representations which are obtained in a supervised training process, via the information bottleneck principal. An interesting connection between deep networks and theoretical physics is given in [28], where RBM-based networks are shown to correspond to block-spin re-normalization, a physical technique for compression of spin configurations.

# 3 Preliminaries

## 3.1 Harmonic analysis on spaces of homogeneous type

### 3.1.1 Construction of wavelet frames

In this section we cite several standard results, mostly from [29], showing how to construct a wavelet frame of $L_2(\mathbb{R}^d)$, and discuss some of its properties.

**Definition 3.1.** *(Definition* 1.1 *in [29])*
*A **space of homogeneous type** $(\mathcal{X}, \mu, \delta)$ is a set $\mathcal{X}$ together with a measure $\mu$ and a quasi-metric $\delta$ (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 3.2.** *(Definition* 3.14 *in [29])*
*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 [29] 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 $x, y \in \mathcal{X}$,*

$$D_k(x,y) = S_k(x,y) - S_{k-1}(x,y), \tag{5}$$

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

By standard wavelet terminology, we denote

$$\psi_{k,b}(x) \equiv 2^{-\frac{k}{2}} D_k(x,b). \tag{6}$$

**Theorem 3.3.** *(A simplified version of Theorem* 3.25 *in [29])*
*Let $\{S_k\}$ be a family of averaging kernels. Then there exist families $\{\psi_{k,b}\}, \{\widetilde{\psi}_{k,b}\}$ such that for all $f \in L_2(\mathbb{R}^d)$*

$$f(x) = \sum_{(k,b)\in\Lambda} \langle f, \widetilde{\psi}_{k,b} \rangle \psi_{k,b}(x) \tag{7}$$

*Where the functions $\psi_{k,b}$ are given by Equations* (5) *and* (6) *and $\Lambda = \{(k,b) \in \mathbb{Z} \times \mathbb{R}^d$: $b \in 2^{-\frac{k}{d}}\mathbb{Z}^d\}$.*

**Remark 3.4.** The functions $\widetilde{\psi}_{k,b}$ are called dual elements, and are also a wavelet frame of $L_2(\mathbb{R}^d)$.

### 3.2 Approximation of functions with sparse wavelet coefficients

In this section we cite a result from [15] 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 [15], 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

$$r_{k-1} := f - f_{k-1}, \tag{8}$$

finds the dictionary element that is most correlated with it

$$g_k := \arg\max_{g \in \mathcal{D}} |\langle r_{k-1}, g \rangle| \tag{9}$$

and defines a new approximation

$$f_k := P_k f, \tag{10}$$

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

**Theorem 3.5.** *(Theorem 2.1 from [15]) The error $r_N$ of the OGA satisfies*

$$\|f - f_N\| \leq \|f\|_{\mathcal{L}_1} (N+1)^{-1/2}. \tag{11}$$

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

$$\mathcal{D} = \{\psi_{k,b} : (k, b) \in \mathcal{Z} \times \mathbb{R}^d, b \in 2^{-k}\mathbb{Z}\}. \tag{12}$$

**Remark 3.6.** Let $\mathcal{D} = \{\psi_{k,b}\}$ be a wavelet frame that satisfies the regularities in conditions $3.14 - 3.19$ in [29]. 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). To see the main idea behind the proof of this claim, consider to frames $\{\psi_{k,b}\}$ and $\{\psi'_{k,b}\}$. Any element $\psi'_{k',b'}$ can be represented as

$$\psi'_{k',b'} = \sum_{k,b} \langle \psi'_{k',b'}, \widetilde{\psi}_{k,b} \rangle \psi_{k,b}. \tag{13}$$

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 [29].

**Remark 3.7.** Section 4.5 in [29] gives a way to check whether a function $f$ has sparse coefficients without actually calculating the coefficients:

$$f \in \mathcal{L}_1 \text{ iff } \sum_{k\in\mathbb{Z}} 2^{k/2}\|f * \psi_{k,0}\|_1 < \infty, \tag{14}$$

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.

### 3.3 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 $\Gamma$ is smooth, so that for every $0 < \epsilon < 1$ there exist $\delta > 0$ such that for every $x, y \in \Gamma$ with $\rho(x, y) < \delta$

$$\frac{1}{1+\epsilon}\rho(x,y) < \|x-y\|_2 < \frac{1}{1-\epsilon}\rho(x,y), \tag{15}$$

where $\rho$ is the geodesic distance on $\Gamma$.

**Definition 3.8.** *A* ***chart*** *for $\Gamma$ is a pair $(U, \phi)$ such that $U \subseteq \Gamma$ is open and*

$$\phi : U \to M, \tag{16}$$

*where $\phi$ 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 $\phi$.

**Definition 3.9.** *An* ***atlas*** *for $\Gamma$ is a collection $\{(U_i, \phi_i)\}_{i\in I}$ of charts such that $\cup_i U_i = \Gamma$.*

**Definition 3.10.** *Let $\Gamma$ be a smooth manifold. A* ***partition of unity*** *of $\Gamma$ 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 3.11.** *(Proposition 13.9 in [30]) Let $\Gamma$ be a compact manifold and $\{U_i\}_{i\in I}$ be an open cover of $\Gamma$. Then there exists a partition of unity $\{\eta_i\}_{i\in I}$ such that for each $i$, $\eta_i$ is in $C^\infty$, has compact support and $\mathrm{supp}(\eta_i) \subseteq U_i$.*

## 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 $\Gamma$ by covering it with open balls (see Section 4.2).

3. We use the open cover to obtain a partition of unity of $\Gamma$ and consequently represent any function on $\Gamma$ 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 $\Gamma$ (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 rectifier activation function is defined on $\mathbb{R}$ as

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

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

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

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

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

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 3.1, 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 A. 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

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

Figure 1 shows the construction of $\varphi$ and $\psi$ in for $d = 1, 2$.

**Remark 4.2.** We can see that

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

From Theorem 3.3 and the above construction we then get the following lemma

**Lemma 4.3.** *$\{\psi_{k,b} : k \in \mathbb{Z}, b \in 2^{-k}\mathbb{Z}\}$ 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 4.4.** *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 B.

**Remark 4.5.** With the above construction, $\varphi$ 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.

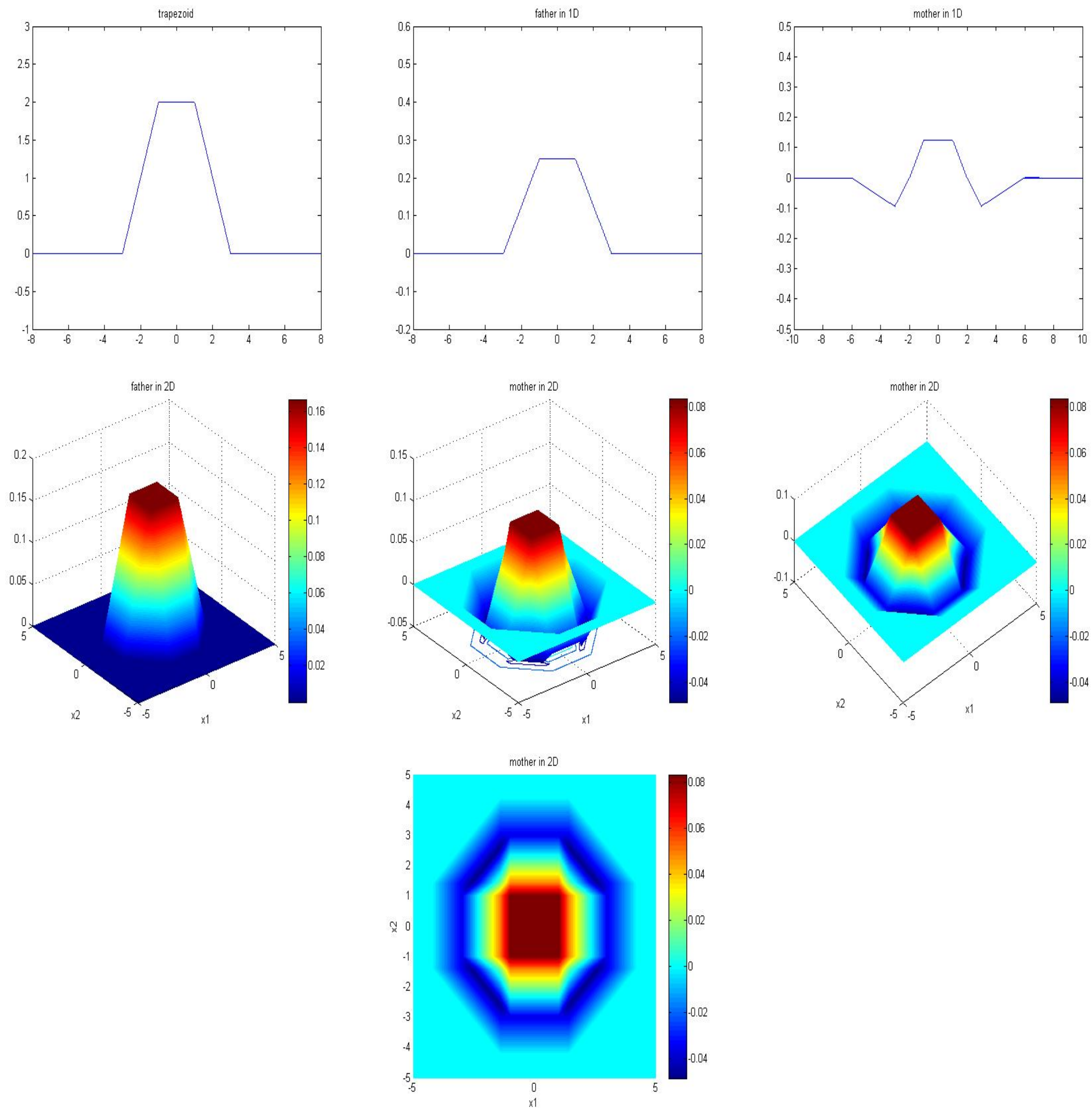


Figure 1: Top row, from left: the trapezoid function $t$, and the functions $\varphi, \psi$ on $\mathbb{R}$. Bottom rows: the functions $\varphi, \psi$ on $\mathbb{R}^2$ from several points of view.

## 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 $C_\Gamma$ of such atlas to depend on the curvature of $\Gamma$: the smoother $\Gamma$ is, the smaller is the number of charts we will need for $\Gamma$.

Following the notation of Section 3.3, let $\epsilon$ be sufficiently small so that the corresponding $\delta$ has the property that for every $m$-dimensional Euclidean ball $B(x_i,(1-\epsilon)\delta)$ centered in $x_i \in \Gamma$, $B(x_i,(1-\epsilon)\delta) \cap \Gamma$ is a local neighborhood of $x_i$. We then cover $\Gamma$ with balls of radius $\frac{(1-\epsilon)\delta}{2}$. The number of such balls that are required to cover $\Gamma$ is

$$C_\Gamma \le \left\lceil \frac{2SA(\Gamma)}{((1-\epsilon)\delta)^d} T_d \right\rceil, \tag{34}$$

where $SA(\Gamma)$ is the surface area of $\Gamma$, 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 [31], there exist covering with $T_d \le d\log d + 5d$. For example, in $d = 24$ there exist covering with thickness of 7.9.

A covering of $\Gamma$ by such a collection of balls defines an open cover of $\Gamma$ by

$$U_i \equiv B\left(x_i, \frac{(1-\epsilon)\delta}{2}\right) \cap \Gamma. \tag{35}$$

This construction is sketched in Figure 2, where $H_i$ is the tangent plane to $\Gamma$ at $x_i$ and $\phi_i$ is the orthogonal mapping from $U_i$ to $H_i$. The above construction has two important properties,

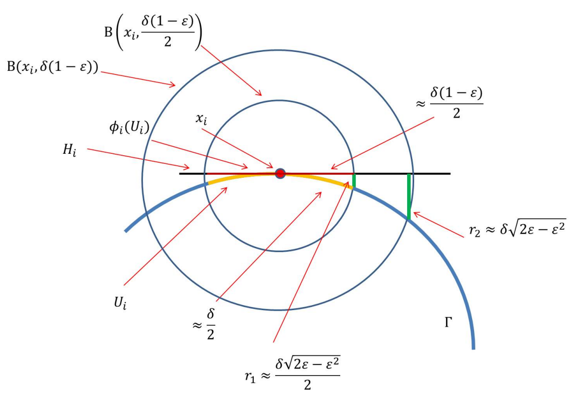


Figure 2: Construction of atlas.

summarized in Lemma 4.7

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

$$\|x - \phi_i(x)\|_2 \le r_1 \approx \frac{\delta\sqrt{2\epsilon - \epsilon^2}}{2} \tag{36}$$

*and for every $x \in \Gamma \setminus U_i$ such that $\phi_i(x) \in \phi_i(U_i)$*

$$\|x - \phi_i(x)\|_2 \ge r_2 \approx \delta\sqrt{2\epsilon - \epsilon^2}. \tag{37}$$

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

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

$\{U_i\}$ is an open cover of $\Gamma$, hence by Theorem 3.11 there exists a corresponding partition of unity, i.e., a family of compactly supported $C^\infty$ functions $\{\eta_i\}_{i=1}^{C_\Gamma}$ such that

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

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 $I_i$. Then $I_i \subset H_i$, 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 $x \in \Gamma$,*

$$\sum_i \hat{f}_i(\phi_i(x)) = f(x). \tag{40}$$

Assuming $\hat{f}_i \in L_2(\mathbb{R}^d)$, by Lemma 4.3 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 $H_i$.

By Lemma 4.7 it 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 $[-r_1, r_1]$ and will have a support which is contained in $[-r_2, r_2]$.

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 $\varphi$ 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 $[-r_2, r_2]$ and its top (small) base is between $[-r_1, r_1]$ and has height 2. Observe that if $\phi(x) > 0$ then all the summands in Equation (46) are positive as well. This definition of $\psi_{k,b}$ gives it a constant height for distance $r_1$ in each orthogonal coordinate, and then a linear decay, until it vanishes at distance $r_2$ from the manifold in this coordinate. This construction gives the following lemma

**Lemma 4.9.** *For every chart $(U_i, \phi_i)$ and every $x \in \Gamma \setminus U_i$, $x$ is outside the support of every wavelet term corresponding to chart $j \neq i$.*

**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 wavelet terms in Equation (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 $C_\Gamma$ blocks of $m$ coordinates gives the local coordinates on $\Gamma$ in the first $d$ coordinates and on the orthogonal subspace in the remaining $m-d$ coordinates.

## 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 the

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

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

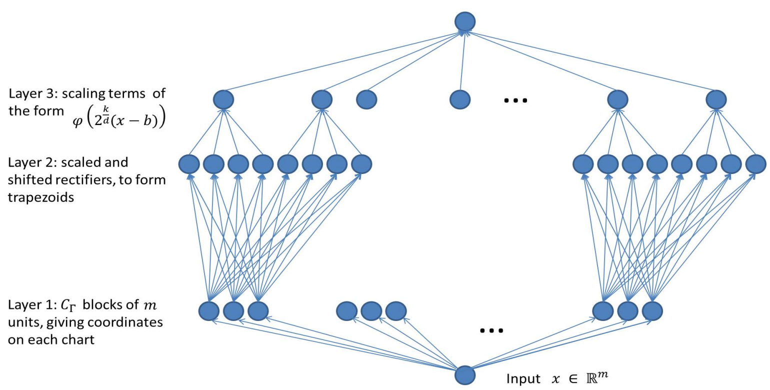


Figure 3: A sketch of the network.

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

**Theorem 5.1.** *Let $\Gamma$ 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 $C_\Gamma$ for $\Gamma$, as in Section 4.2. Then $f$ can be approximated using a 4-layer network with $mC_\Gamma$ linear units in the first hidden layer $8d\sum_{i=1}^{C_\Gamma} k_i + 4C_\Gamma(m-d)$ rectifier units in the second hidden layer, $2\sum_{i=1}^{C_\Gamma} k_i$ rectifier units in the third layer and a single linear unit in the fourth (output) layer, where $k_i$ 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 $x \in \Gamma$, $\sum_i \hat{f}_i(\phi_i(x)) = f(x)$. By Lemma 4.9 every wavelet term in the extension of the expansion of $\hat{f}_i$ is not supported on $\Gamma \setminus U_i$, so an approximation of $f$ is obtained by adding the approximations of all the $\hat{f}_i$'s.

A first layer of the network will consist $mC_\Gamma$ linear units and will compute the map as in the last paragraph of Section 4.4, i.e., linearly transform the input to $C_\Gamma$ 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 $\phi_i(x)$) and the remaining $m-d$ coordinates are with respect to directions orthogonal to $H_i$.

For each $i = 1, .., C_\Gamma$, we approximate each $\hat{f}_i$ using $k_i < \infty$ wavelet terms. By Remark 4.5, $\hat{f}_i$ can be approximated using $8dk_i$ rectifiers in the second layer, $2k_i$ 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 $mC_\Gamma$ linear units in the first hidden layer $8d\sum_{i=1}^{C_\Gamma} k_i + 4C_\Gamma(m-d)$ rectifier units in the second hidden layer, $2\sum_{i=1}^{C_\Gamma} k_i$ rectifier units in the third layer and a single linear unit in the fourth (output) layer. □

**Remark 5.2.** For sufficiently small $\delta$ in Equation (15), the desired properties of $\hat{f}_i$ (i.e., being in $L_2$ and possibly 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 $C_\Gamma$, which depends on the curvature of the manifold. The number $k_i$ 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 $k_i$ of wavelet terms needed for the approximation of $\hat{f}_i$. In particular, we consider to 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., have expansion coefficients in $l_1$), then by Theorem 3.5, $\hat{f}_i$ can be approximated by a combination $\hat{f}_{i,k_i}$ of $k_i$ wavelet terms so that*

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

*Consequently, denoting $N \equiv \max_i\{k_i\}$ and $M \equiv \max_i \|\hat{f}_i\|_2$, we obtain an $L_2$ error rate of $\frac{M}{\sqrt{N}}$ using $c_1 + c_2N$ units, where $c_1 = mC_\Gamma + 4(m-d) + 1$ 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 4.4 $\hat{f}_i$ can be approximated by $\hat{f}_{K,i}$ using terms up to scale $K$ so that for every $x \in \mathbb{R}^d$*

$$|\hat{f}_i(x) - \hat{f}_{i,K}(x)| = O\left(2^{-\frac{2K}{d}}\right). \tag{49}$$

*Observe that the grid in the $k$'th level is $2^{-\frac{k}{d}}$, hence when $f$ is compactly supported, it is approximated using $\left(2^{\frac{k}{d}}\right)^d = 2^k$ terms in the $k$'th level, and $2^{K+1}$ terms in levels $k = 0, .., K$. Writing $N \equiv 2^{K+1}$, we get a point-wise error rate of $N^{-\frac{2}{d}}$ using $c_1 + c_2N$ units, where $c_1 = mC_\Gamma + 4(m-d) + 1$ 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 $C_\Gamma \sum k_i$ 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: 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 $\Gamma$ we constructed an atlas and represented any function on $\Gamma$ 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 and a manifold $\Gamma$. 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.

In the future, we plan to obtain the partition of unity not only based on the manifold $\Gamma$ but also based on the level sets of the functions $\hat{f}_i$. This might allow us to approximate the functions $\hat{f}_i$ to a better precision using less wavelet terms. We hypothesize that a similar process, i.e., adjusting the hidden units to the level sets of the function, also occurs during standard training of neural nets. In particular, when keeping the weights of the network tied in a way that verifies that the net computes our wavelet functions, training the net will correspond to find the appropriate scales $k$ and shifts $b$ to obtain efficient representation of the functions $\hat{f}_i$'s.

Finally, observe 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

$$\lim_{x\to-\infty} \sigma(x) = 0 \text{ and } \lim_{x\to\infty} \sigma(x) = 1 \tag{50}$$

can used to construct a frame in a similar way, by computing "smooth" trapezoids. Recall also that by Remark 3.6, any two such frames are equivalent.

## Acknowledgements

The authors thank Stephan Steinerberger and Roy Lederman for their help. 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] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

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

[13] 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.

[14] E. K. Blum and L. K. Li, "Approximation theory and feedforward networks," *Neural networks*, vol. 4, no. 4, pp. 511–515, 1991.

[15] 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.

[16] P. L. Bartlett and W. Maass, "Vapnik chervonenkis dimension of neural nets," *The handbook of brain theory and neural networks*, pp. 1188–1192, 2003.

[17] E. B. Baum and D. Haussler, "What size net gives valid generalization?," *Neural computation*, vol. 1, no. 1, pp. 151–160, 1989.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

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

[25] 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.

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

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

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

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

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

[31] 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 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 [29] 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:

$$S_k(x, b) \leq C \frac{2^{-k\epsilon}}{(2^{-k} + \rho(x, b))^{1+\epsilon}}, \tag{51}$$

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

$$\varphi(x) \leq \frac{C'}{(c^{-1} + \|x\|^d)^{1+\epsilon}}. \tag{52}$$

Then

$$S_k(x, 0) = 2^k \varphi\left(2^{\frac{k}{d}} x\right) \tag{53}$$

$$\leq C' \frac{2^k}{\left(c^{-1} + 2^k \|x\|^d\right)^{1+\epsilon}} \tag{54}$$

$$= C' \frac{2^{k(1+\epsilon)} 2^{-k\epsilon}}{\left(c^{-1} + 2^k \|x\|^d\right)^{1+\epsilon}} \tag{55}$$

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

$$= C_1 \frac{2^{-k\epsilon}}{\left(2^{-k} + \rho(x, 0)\right)^{1+\epsilon}}, \tag{57}$$

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

$$|S_k(x, b) - S_k(x', b)| \leq C \left(\frac{\rho(x, x')}{2^{-k} + \rho(x, b)}\right)^\sigma \frac{2^{-k\epsilon}}{(2^{-k} + \rho(x, b))^{1+\epsilon}}. \tag{58}$$

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

$$\frac{|S_k(x,0) - S_k(x',0)|}{\rho(x,x')^\sigma} \leq \max_{z_k \text{ between } x,x'} \frac{1}{c} \|\nabla_x (S_k(z_k,0))\|. \tag{59}$$

Denote

$$F(x) \equiv \|\nabla_x (S_0(x,0))\|. \tag{60}$$

Then

$$\|\nabla_x (S_k(x,0))\| = 2^k 2^{\frac{k}{d}} F\left(2^{\frac{k}{d}} x\right). \tag{61}$$

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

$$F(x) \leq C' \frac{1}{(c^{-1} + \|x\|^d)^\sigma} \frac{1}{(c^{-1} + \|x\|^d)^{1+\epsilon}}. \tag{62}$$

We then get

$$\begin{aligned}
\frac{|S_k(x,b) - S_0(x',b)|}{\rho(x,x')^\sigma} &= \frac{1}{c} \|\nabla_x (S_k(z_k,0))\| && (63)\\
&= 2^k 2^{\frac{k}{d}} F\left(2^{\frac{k}{d}}\right) && (64)\\
&\leq C' \frac{2^{\frac{k}{d}}}{(c^{-1} + 2^k\|x\|^d)^\sigma} \frac{2^k}{(c^{-1} + 2^k\|x\|^d)^{1+\epsilon}} && (65)\\
&= C' \frac{2^{\frac{k}{d}}}{(c^{-1} + 2^k\|x\|^d)^\sigma} \frac{2^{k(1+\epsilon)} 2^{-k\epsilon}}{(c^{-1} + 2^k\|x\|^d)^{1+\epsilon}} && (66)\\
&= C' \frac{1}{(c^{-1}2^{-k} + \|x\|^d)^\sigma} \frac{2^{-k\epsilon}}{(c^{-1}2^{-k} + \|x\|^d)^{1+\epsilon}} && (67)\\
&= C_2 \frac{1}{(2^{-k} + \rho(x,0))^\sigma} \frac{2^{-k\epsilon}}{(2^{-k} + \rho(x,0))^{1+\epsilon}}, && (68)
\end{aligned}$$

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

- 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

$$\int_{\mathbb{R}^d} S_k(x,b)dx = \int_{\mathbb{R}^d} S_k(x,b)db = 1. \tag{69}$$

*Proof.* By Equation (19)

$$\int_{\mathbb{R}^d} \varphi(x)dx = 1 \tag{70}$$

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

$$\int_{\mathbb{R}^d} S_k(x,b)dx = 1. \tag{71}$$

□

- 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

$$|S_k(x,b) - S_k(x',b) - S_k(x,b') + S_k(x',b')| \tag{72}$$

$$\le C\left(\frac{\rho(x,x')}{2^{-k}+\rho(x,b)}\right)^{\sigma}\left(\frac{\rho(b,b')}{2^{-k}+\rho(x,b)}\right)^{\sigma}\frac{2^{-k\epsilon}}{(2^{-k}+\rho(x,b))^{1+\epsilon}}. \tag{73}$$

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

$$\frac{|S_k(x,b) - S_k(x',b) - S_k(x,b') + S_k(x',b')|}{\rho(x,x')^{\sigma}\rho(b,b')^{\sigma}} \tag{74}$$

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

$$\tag{76}$$

Denote

$$F(b) \equiv \frac{|S_k(x,b) - S_k(x',b)|}{\rho(x,x')^{\sigma}}. \tag{77}$$

Then by applying the mean value theorem twice we get

$$\frac{\left|\frac{|S_k(x,b)-S_k(x',b)|}{\rho(x,x')^{\sigma}} + \frac{|S_k(x,b')+S_k(x',b')|}{\rho(x,x')^{\sigma}}\right|}{\rho(b,b')^{\sigma}} \tag{78}$$

$$= \frac{|F(b) - F(b')|}{\rho(b,b')^{\sigma}} \tag{79}$$

$$\frac{1}{c} \le \max_{z \text{ between } b,b'} \nabla_b(F(z)) \tag{80}$$

$$= \frac{1}{c}\max_{z \text{ between } b,b'} \nabla_b\left(\frac{|S_k(x,z) - S_k(x',z)|}{\rho(x,x')^{\sigma}}\right) \tag{81}$$

$$\frac{1}{c^2} \le \max_{z \text{ between } b,b'}\ \max_{z' \text{ between } x,x'} \|\nabla^2_{x,b}(S_k(z',z))\| \tag{82}$$

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

$$\frac{|S_0(x,b)-S_0(x',b)-S_0(x,b')+S_0(x',b')|}{\rho(x,x')^\sigma\rho(b,b')^\sigma} \tag{83}$$

$$\leq \xi(x-b)+\xi(x-b'), \tag{84}$$

and consequently

$$\frac{|S_k(x,b)-S_k(x',b)-S_k(x,b')+S_k(x',b')|}{\rho(x,x')^\sigma\rho(b,b')^\sigma}| \tag{85}$$

$$\leq 2^k 2^{\frac{2k}{d}}\left(\xi\left(2^{\frac{k}{d}}(x-b)\right)+\xi\left(2^{\frac{k}{d}}(x-b')\right)\right). \tag{86}$$

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

$$\xi(x-b)+\xi(x-b')\leq C'\frac{1}{(c^{-2}+\|x-b\|^d)^{2\sigma}}\frac{1}{(c^{-1}+\|x-b\|^d)^{1+\epsilon}}. \tag{87}$$

We then get

$$\frac{|S_k(x,b)-S_k(x',b)-S_k(x,b')+S_k(x',b')|}{\rho(x,x')^\sigma\rho(b,b')^\sigma} \tag{88}$$

$$\leq 2^k 2^{\frac{2k}{d}}\left(\xi\left(2^{\frac{k}{d}}(x-b)\right)+\xi\left(2^{\frac{k}{d}}(x-b')\right)\right) \tag{89}$$

$$\leq C'\frac{2^{\frac{2k}{d}}}{(c^{-2}+2^k\|x-b\|^d)^{2\sigma}}\frac{2^k}{(c^{-1}+2^k\|x-b\|^d)^{1+\epsilon}} \tag{90}$$

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

$$= C_3\frac{1}{(2^{-k}+\rho(x,b))^{2\sigma}}\frac{2^{-k\epsilon}}{(2^{-k}+\rho(x,b))^{1+\epsilon}}, \tag{92}$$

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

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

# B Proof of Lemma 4.4

We first prove the following propositions.

**Proposition B.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

$$\int_{\mathbb{R}^d} xf(x)dx = 0. \tag{93}$$

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

$$\begin{aligned} 2^{-k}\int_{\mathbb{R}^d} \varphi(2^{\frac{k}{d}}(x-b))dx &= \int_{\mathbb{R}^d} \varphi(x)dx && (94)\\ &= 1, && (95)\end{aligned}$$

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

$$\begin{aligned} \int_{\mathbb{R}^d} \psi_{k,b}(x)dx &= 2^{\frac{k}{2}}\int_{\mathbb{R}^d} \varphi(2^{\frac{k}{d}}(x-b) - \varphi\left(2^{\frac{k-1}{d}}(x-b)\right)dx && (96)\\ &= 0, && (97)\end{aligned}$$

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

$$\begin{aligned} \int_{\mathbb{R}^d} x\varphi\left(2^{\frac{k}{d}}(x-b)\right)dx &= \int_{\mathbb{R}^d} (2^{-\frac{k}{d}}y+b)\varphi(y)dy && (98)\\ &= 2^{-k}b\int_{\mathbb{R}^d}\varphi(y)dy && (99)\\ &= 2^{-k}b, && (100)\end{aligned}$$

which gives

$$\begin{aligned} \int_{\mathbb{R}^d} x\psi_{k,b}(x)dx &= 2^{-\frac{k}{2}}\int_{\mathbb{R}^d} \varphi\left(2^{\frac{k}{d}}(x-b)\right) - 2^{-1}\varphi\left(2^{\frac{k-1}{d}}(x-b)\right)dx && (101)\\ &= 2^{-\frac{k}{2}}\left(2^{-k}b - 2^{-1}2^{-(k-1)}b\right) && (102)\\ &= 2^{-\frac{k}{2}}\left(2^{-k}b - 2^{-k}b\right) && (103)\\ &= 0, && (104)\end{aligned}$$

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

Finally 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 $D_k$ ([29], p. 82), which, by the above arguments, have two vanishing moments; hence they inherit this property. □

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

*Proof.* By ([29], p. 82), the dual functions are also wavelets, hence they satisfy condition 3.14 in [29] with $\epsilon' < \epsilon$. Since in the proof of Lemma 4.1, $\epsilon$ can be arbitrarily large, it implies that the duals satisfy condition 3.14 with any $\epsilon$, which proves the proposition. □

**Proposition B.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

$$\psi_{k,b}(x) = 2^{\frac{k}{2}} \psi\left(2^{\frac{k}{d}} x - b)\right) \tag{105}$$

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

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

$$|\langle \widetilde{\psi}_{k,b}, f \rangle| = O(2^{-(2\frac{k}{d}+\frac{k}{2})}) \tag{106}$$

*Proof.*

$$\langle \widetilde{\psi}_{k,b}, f \rangle = 2^{\frac{k}{2}} \int_{\mathbb{R}^d} \widetilde{\psi}\left(2^{\frac{k}{d}}(x-b)\right) f(x) dx \tag{107}$$

$$= 2^{-\frac{k}{2}} \int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) f(2^{-\frac{k}{d}} y + b) dy. \tag{108}$$

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

$$\int_{\text{supp}(\widetilde{\psi})} \widetilde{\psi}(y) f(2^{-\frac{k}{d}} y + b) dy \tag{109}$$

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

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

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

Since by Proposition B.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,

$$|\langle \widetilde{\psi}_{k,b}, f \rangle| = O(2^{-(2\frac{k}{d}+\frac{k}{2})}) \tag{112}$$

□

We will also use the following property

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

We are now ready to prove Lemma 4.4

*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

$$f = \sum_{(k,b)\in\Lambda} \langle \widetilde{\psi}_{k,b}, f\rangle \psi_{k,b}. \tag{113}$$

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

$$|f(x) - f_K(x)| \leq \sum_{k=K+1}^{\infty} \sum_{b\in 2^{-k}\mathbb{Z}} |\psi_{k,b}| \langle \widetilde{\psi}_{k,b}, f\rangle. \tag{114}$$

By Remark B.5, at most $12^d$ wavelet terms are supported on $x$ at every scale; by Proposition B.3 $|\psi_{k,b}| \leq 2^{\frac{k}{2}-2}$; by Proposition B.4 $|\langle \widetilde{\psi}_{k,b}, f\rangle| = O(2^{-(\frac{2k}{d}+\frac{k}{2})})$. Plugging these into Equation (114) gives

$$\begin{aligned} |f(x) - f_K(x)| &= O\left(\sum_{k=K+1}^{\infty} 12^d 2^{\frac{k}{2}-2} 2^{-(\frac{2k}{d}+\frac{k}{2})}\right) && (115)\\ &= O\left(\sum_{k=K+1}^{\infty} 2^{-\frac{2k}{d}}\right) && (116)\\ &= O\left(2^{-\frac{2K}{d}}\right). && (117) \end{aligned}$$

□