ON THE APPROXIMATION PROPERTIES OF NEURAL NETWORKS

Jonathan W. Siegel
Department of Mathematics
Pennsylvania State University
State College, PA 16801
jus1949@psu.edu

Jinchao Xu
Department of Mathematics
Pennsylvania State University
State College, PA 16801
jxx1@psu.edu

April 5, 2019

ABSTRACT

We prove two new results concerning the approximation properties of neural networks. Our first result gives conditions under which the outputs of the neurons in a two layer neural network are linearly independent functions. Our second result concerns the rate of approximation of a two layer neural network as the number of neurons increases. We improve upon existing results in the literature by significantly relaxing the required assumptions on the activation function and by providing a better rate of approximation. We also provide a simplified proof that the class of functions represented by a two-layer neural network is dense in any compact set if the activation function is not a polynomial.

1 Introduction

Deep neural networks have recently revolutionized a variety of areas of machine learning, including computer vision and speech recognition [10]. A deep neural network with \( l \) layers is a statistical model which takes the following form

\[
f(x; \theta) = h_{W_l, b_l} \circ \sigma \circ h_{W_{l-1}, b_{l-1}} \circ \sigma \circ \cdots \circ \sigma \circ h_{W_1, b_1}
\]  

(1)

where \( h_{W_i, b_i}(x) = W_i x + b_i \) is an affine linear function, \( \sigma \) is a fixed activation function which is applied pointwise and \( \theta = \{W_1, \ldots, W_l, b_1, \ldots, b_l\} \) are the parameters of the model.

The approximation properties of neural networks have received a lot of attention, with many positive results. For example, in [5][12] it is shown that neural networks can approximate any function on a compact set as long as the activation function is not a polynomial, i.e. that the set

\[
\Sigma_d(\sigma) = \text{span}\left\{ \sigma(\omega \cdot x + b) : \omega \in \mathbb{R}^d, b \in \mathbb{R} \right\}
\]  

(2)

is dense in \( C(\Omega) \) for any compact \( \Omega \subset \mathbb{R}^d \). An earlier result of this form can be found in [8], and [7] shows that derivatives can be approximated arbitrarily accurately as well. An elementary and constructive proof for \( C^\infty \) functions can be found in [1].

In addition, quantitative estimates on the order of approximation are obtained for sigmoidal activation functions in [2] and for periodic activation functions in [16] and [14]. Results for general activation functions can be found in [9]. Results concerning the approximation properties of generalized translation networks (a generalization of two-layer neural networks) for smooth and analytic functions are obtained in [15] and approximation estimates for multilayer convolutional neural networks are considered in [23] and multilayer networks with rectified linear activation functions in [22]. A comparison of the effect of depth vs width on the expressive power of neural networks is presented in [13]. A review of a variety of known results, especially for networks with one hidden layer, can be found in [18].
Our work, like much of the previous work, focuses on the case of two-layer neural networks. A two layer neural network can be written in the following particularly simple way
\[ f(x; \theta) = \sum_{i=1}^{n} \beta_i \sigma(\omega_i \cdot x + b_i) \] (3)
where \( \theta = \{ \omega_1, \ldots, \omega_n, b_1, \ldots, b_n \} \) are parameters and \( n \) is the number of hidden neurons in the model.

We consider two main problems in this work. First, we consider the problem of determining when the functions output by each neuron are linearly independent, i.e. under which conditions on the parameters \( \omega_i \) and \( b_i \) is the set
\[ \{ \sigma(\omega_i \cdot x + b_i) : i = 1, \ldots, n \} \] (4)
linearly independent. We show that as long as \( \sigma \) is not a polynomial, then this set is linearly independent as long as the \( \omega_i \) are pairwise linearly independent (i.e. no two point in the same direction), which, to the best of our knowledge, is a new and surprising result concerning two-layer neural networks.

Second, we study the how the approximation properties of two-layer neural networks depends on the number of hidden neurons. In particular, we consider the class of functions where the number of hidden neurons is bounded
\[ \sum_{i=1}^{n} \sigma(\omega_i \cdot x + b_i) : \omega_i \in \mathbb{R}^d, b_i, \beta_i \in \mathbb{R} \] (5)
and prove the Theorem concerning the order of approximation as \( n \to \infty \) for activation functions with polynomial decay. Similar results appear in \[2, 9\], but we have improved their bound by a logarithmic factor for exponentially decaying activation functions and attain a significantly improved rate of approximation for polynomially decaying activation functions, in addition to providing a simplified argument.

The paper is organized as follows. In the next section, we discuss some basic results concerning the Fourier transform. We use these results to provide a simplified proof using Fourier analysis of the density result in \[12\] under the mild additional assumption of polynomial growth on \( \sigma \). In the third section, we consider the linear independence problem and prove theorem. Finally, we study the order of approximation and prove theorem. We then give concluding remarks and further research directions in the conclusion.

## 2 Preliminaries

Our arguments will make use of the theory of tempered distributions (see \[19, 21\] for an introduction) and we begin by collecting some results of independent interest, which will also be important later. We begin by noting that an activation function \( \sigma \) which satisfies a polynomial growth condition \( |\sigma(x)| \leq C(1 + |x|^n) \) for some constants \( C \) and \( n \) is a tempered distribution. As a result, we make this assumption on our activation functions in the following theorems. We briefly note that this condition is sufficient, but not necessary (for instance an integrable function need not satisfy a pointwise polynomial growth bound) for \( \sigma \) to be represent a tempered distribution.

We begin by stating a lemma which characterizes the set of polynomials in terms of their Fourier transform.

**Lemma 1.** Given a tempered distribution \( \sigma \), the following statements are equivalent:

1. \( \sigma \) is a polynomial

We begin by stating a lemma which characterizes the set of polynomials in terms of their Fourier transform.

**Lemma 1.** Given a tempered distribution \( \sigma \), the following statements are equivalent:

1. \( \sigma \) is a polynomial

2. \[ \hat{\sigma}(\xi) = \hat{\sigma} \eta_{1.2} \] (8)
2. \( \sigma_\varepsilon \) given by (7) is a polynomial for any \( \varepsilon > 0 \).
3. \( \supp(\hat{\sigma}) \subset \{0\} \).

Proof. We begin by proving that (3) and (1) are equivalent. This follows from a characterization of distributions supported at a single point (see [21], section 6.3). In particular, a distribution supported at 0 must be a finite linear combination of Dirac masses and their derivatives. In particular, if \( \hat{\sigma} \) is supported at 0, then

\[
\hat{\sigma} = \sum_{i=1}^{n} a_i \delta^{(i)}(i)
\]

Taking the inverse Fourier transform and noting that the inverse Fourier transform of \( \delta^{(i)}(i) \) is \( c_i x^i \), we see that \( \sigma \) is a polynomial. This shows that (3) implies (1), for the converse we simply take the Fourier transform of a polynomial and note that it is a finite linear combination of Dirac masses and their derivatives.

Finally, we prove the equivalence of (2) and (3). For this it suffices to show that \( \hat{\sigma} \) is supported at 0 iff \( \hat{\sigma}_\varepsilon \) is supported at 0. This follows from equation (8) and the fact that \( \eta_{\varepsilon^{-1}} \) is nowhere vanishing.

As an application of Lemma 1 let us give a simple proof of the following result. The first proof of this result can be found in [12] and is summarized in [18]. Extending this result to the case of non-smooth activation is first done in [12]. Our contribution is to provide a much simpler argument based on Fourier analysis.

**Theorem 1.** Assume that \( \sigma \) is a Riemann integrable function which satisfies a polynomial growth condition, i.e.

\[
|\sigma(t)| \leq C(1 + |t|)^p
\]

holds for some constants \( C \) and \( p \). Then if \( \sigma \) is not a polynomial, \( \Sigma_d(\sigma) \) in dense in \( C(\Omega) \) for any compact \( \Omega \subset \mathbb{R}^n \).

Proof. Let us first prove the theorem in a special case that \( \sigma \in \mathcal{C}^\infty(\mathbb{R}) \). Since \( \sigma \in \mathcal{C}^\infty(\mathbb{R}) \), it follows that for every \( \omega, b \)

\[
\frac{\partial}{\partial \omega_j} \sigma(\omega \cdot x + b) = \lim_{n \to \infty} \frac{\sigma(\omega + he_j \cdot x + b) - \sigma(\omega \cdot x + b)}{h} \in \Sigma_d(\sigma)
\]

for all \( j = 1, \ldots, d \).

By the same argument, for \( \alpha = (\alpha_1, \ldots, \alpha_d) \)

\[
D^\alpha_\omega \sigma(\omega \cdot x + b) \in \Sigma_d(\sigma)
\]

for all \( k \in \mathbb{N}, j = 1, \ldots, d, \omega \in \mathbb{R}^d \) and \( b \in \mathbb{R} \).

Now

\[
D^\alpha_\omega \sigma(\omega \cdot x + b) = x^\alpha \sigma^{(k)}(\omega \cdot x + b)
\]

where \( x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d} \). Since \( \sigma \) is not a polynomial there exists a \( \theta_k \in \mathbb{R} \) such that \( \sigma^{(k)}(\theta_k) \neq 0 \). Taking \( \omega = 0 \) and \( b = \theta_k \), we thus see that \( x_k^j \in \Sigma_\delta(\sigma) \). Thus, all polynomials of the form \( x_1^{k_1} \cdots x_d^{k_d} \) are in \( \Sigma_d(\sigma) \).

This implies that \( \Sigma_d(\sigma) \) contains all polynomials. By Weierstrass's Theorem [20] it follows that \( \Sigma_d(\sigma) \) contains \( C(K) \) for each compact \( K \subset \mathbb{R}^n \). That is \( \Sigma_d(\sigma) \) is dense in \( C(\mathbb{R}^d) \).

By the preceding lemma, it follows that \( \Sigma_d(\sigma_\varepsilon) \) is dense and so it suffices to show that \( \sigma_\varepsilon \in \Sigma_d(\sigma) \). This follows by using the Riemann integrability of \( \sigma \) and approximating the integral

\[
\int_{\mathbb{R}} \sigma(x - y) \eta_\varepsilon(y)dy
\]

by a sequence of Riemann sums, each of which is clearly in \( \Sigma_d(\sigma) \). \( \square \)

We now collect a simple lemma for which we have failed to find a reference and which we need in the next section. We note that this lemma is very similar to an observation made in the proof of Theorem 4.1 in [10] concerning the set of discontinuities of a neural network function.

**Lemma 2.** Let \( T_1, \ldots, T_n \in \mathcal{S}' \) be a finite collection of tempered distributions. Then

\[
\supp(T_1 + \cdots + T_n) \supset \left( \bigcup_{i=1}^{n} \supp(T_i) \right) - \bigcup_{i \neq j} (\supp(T_i) \cap \supp(T_j))
\]

(13)
Then we have
\[
x ∈ \left( \bigcup_{i=1}^{n} \text{supp}(T_i) \right) - \bigcup_{i \neq j} (\text{supp}(T_i) ∩ \text{supp}(T_j))
\]
This means that there exists a unique \( i \in \{1,\ldots,n\} \) such that \( x ∈ \text{supp}(T_i) \) (and \( x \notin \text{supp}(T_j) \) for all \( j \neq i \)). We wish to show that \( x ∈ \text{supp}(T_1 + \cdots + T_n) \). Recall that \( x ∈ \text{supp}(T) \) for a distribution \( T \) if for any neighborhood \( U \ni x \), there exists a \( φ ∈ C_c^∞(U) \) such that \( T(φ) \neq 0 \).

For each \( j \neq i \), since \( x \notin \text{supp}(T_j) \), there exists a neighborhood \( U_j \ni x \) such that \( T_i(φ) \neq 0 \). Moreover, since \( V ⊂ U_j \) for each \( j \neq i \), we also have \( T_j(φ) = 0 \) for each \( j \neq i \). This implies that
\[
(T_1 + \cdots + T_n)(φ) = T_i(φ) + \cdots + T_n(φ) = T_i(φ) \neq 0
\]
Since \( V \) was arbitrary, \( x ∈ \text{supp}(T_1 + \cdots + T_n) \) as desired. \( \square \)

Finally, let \( σ \) be an activation function with at most polynomial growth and \( f ∈ \mathcal{S}(\mathbb{R}^d) \) a Schwartz function. Then we have the following elementary lemma.

**Lemma 3.** Given \( ω ∈ \mathbb{R}^d \) and \( b ∈ \mathbb{R} \), define the following
\[
σ_{ω,b}(x) = σ(ω \cdot x + b)
\]
Then we have
\[
\langle σ_{ω,b}, f \rangle = \langle σ, f_{ω,b} \rangle
\]
Notice that \( σ_{ω,b} \) is the extension of a one-dimensional function to \( \mathbb{R}^d \) and \( f_{ω,b} \) is the restriction of a \( d \)-dimensional function to a one-dimensional subspace.

**Proof.** We simply calculate the Fourier tranform of \( σ_{ω,b} \). Let \( f ∈ \mathcal{S}(\mathbb{R}^d) \) be a Schwartz function. We have by definition
\[
\langle σ_{ω,b}, f \rangle = \int_{\mathbb{R}^d} σ(ω \cdot x + b) \int_{\mathbb{R}^d} e^{-iξ \cdot x} f(ξ) dξ dx
\]
We now decompose both integrals into an integral parallel and perpendicular to \( ω \), i.e. we consider the change of variables \( x = tw'_i + v \) and \( ξ = sv'_i + u \) where \( w'_i \) is the normalized vector \( ω/|ω| \) (note that \( ω \neq 0 \) by assumption) and \( u,v ⊥ v'_i \), to get
\[
\int_{\mathbb{R}} σ(t|ω| + b) \int_{\mathbb{R}} e^{iξt} \int_{u⊥ω} \int_{v⊥ω} e^{iξu} f(-sv'_i + u) dv du dt
\]
since \( x_⊥ = sv'_i + u \).

Now we use Fourier inversion on the \( d - 1 \) dimensional (Schwartz) function \( g(u) = f(-sv'_i + u) \) to see that
\[
\int_{u⊥ω} \int_{v⊥ω} e^{-iξu} f(-sv'_i + u) dv du = (2π)^{d-1} g(0) = (2π)^{d-1} f(-sv'_i)
\]
so we get
\[
\int_{\mathbb{R}} σ(t|ω| + b) \int_{\mathbb{R}} e^{iξt} f(-sv'_i) dv dt
\]
Utilizing the change of variables \( s' = -s \), and defining the function \( f_{ω'}(s) = f(sω') \), we see that
\[
\hat{σ}_{ω,b}(f) = (2π)^{d-1} \int_{\mathbb{R}} σ(t|ω| + b) f_{ω'}(t) dt
\]
or, via a final change of variables
\[
\hat{σ}_{ω,b}(f) = \frac{(2π)^{d-1}}{|ω|} \int_{\mathbb{R}} σ(t|ω| + b) \left( \frac{1 - b}{|ω|} \right) dt
\]
Now, \( \frac{1}{|ω|} f_{ω'} \left( \frac{t-b}{|ω|} \right) \) is the Fourier transform of the function
\[
f_{ω,b}(t) = e^{iωt} f(tω)
\]
and so we can rewrite this as \( \langle σ_{ω,b}, f \rangle \). \( \square \)
3 Linear Independence

In this section, we consider the linear independence of the outputs from each neuron, i.e. we ask the question: Under what conditions is the set

$$\{ \sigma(\omega_k \cdot x + b_i) : i = 1, \ldots, n \}$$

linearly independent?

First of all, note that if $\sigma$ is a polynomial of degree $k$, then each of the functions $\sigma(\omega_k \cdot x + b_i)$ is a polynomial of degree at most $k$ on $\mathbb{R}^d$. This means that at most $\binom{k+d}{d}$ of the above functions can be linearly independent.

It is known (and will follow from the results in the next section), that the number of possible linearly independent functions of the form $\sigma(\omega_k \cdot x + b_i)$ is infinite as long as $\sigma$ isn’t a polynomial. However, this result comes without a characterization of when such a set is linearly independent.

The case when $\sigma$ is a rectified linear unit was studied in (9). It was shown that the set (24) is linearly independent when the parameter vectors $(\omega_i, b_i)$ are pairwise linearly independent, i.e. no two are multiples of each other.

Our main result in this section is to generalize this result to an arbitrary activation function $\sigma$. However, we need to strengthen the condition to assuming that the $\omega_i$ are pairwise linearly independent.

**Theorem 2.** Assume that $\sigma$ is an integrable function which satisfies a polynomial growth condition, i.e. $|\sigma(x)| \leq C(1 + |x|)^n$ for some constants $C$ and $n$. Suppose in addition that $\sigma$ is not a polynomial. Then the functions $\sigma(\omega_k \cdot x + b_i)$ are linearly independent if the parameters $\omega_k$ are pairwise linearly independent, i.e. no two are multiples of each other.

**Proof.** We begin by noting that since $\sigma$ is locally integrable with at most polynomial growth at infinity, it is a tempered distribution.

Finally, we consider the support of the tempered distribution $\hat{\sigma}_{\omega_i,b_i}$. From Lemma 5 it is clear that the support is given by scaling the support of $\hat{\sigma}$ by $|\omega|$ and placing it along the line in the direction of $\omega$. In other words $x \in \text{supp}(\hat{\sigma}_{\omega_i,b_i})$ iff $x = t\omega$ for some $t \in \text{supp}(\hat{\sigma})$.

We now utilize Lemma 1 to see that if $\sigma$ is not a polynomial, then $\text{supp}(\hat{\sigma}) \not\subseteq \{0\}$, i.e. there exists a $0 \neq t \in \text{supp}(\hat{\sigma})$. Fix this element $t$ in what follows.

Assume that we have a relation of the form

$$\sum_{i=1}^n a_i \hat{\sigma}_{\omega_i,b_i} = 0$$

If $a_i \neq 0$, then $\text{supp}(a_i \hat{\sigma}_{\omega_i,b_i}) = \text{supp}(\hat{\sigma}_{\omega_i,b_i}) \ni t\omega_i$. But, by the calculation above, since the $\omega_i$ are all pairwise independent and $t \neq 0$, $t\omega_i \notin \text{supp}(a_j \hat{\sigma}_{\omega_j,b_j})$ for any other $j$. This means that $t\omega_i \notin \text{supp}(a_j \hat{\sigma}_{\omega_j,b_j}) \cap \text{supp}(a_k \hat{\sigma}_{\omega_k,b_k})$ for any pair $j, k$, which implies that

$$t\omega_i \in \left( \bigcap_{i=1}^n \text{supp}(a_i \hat{\sigma}_{\omega_i,b_i}) \right) - \bigcup_{i \neq j} \left( \text{supp}(a_i \hat{\sigma}_{\omega_i,b_i}) \cap \text{supp}(a_j \hat{\sigma}_{\omega_j,b_j}) \right)$$

and so by Lemma 2 we get

$$t\omega_i \in \text{supp} \left( \sum_{i=1}^n a_i \hat{\sigma}_{\omega_i,b_i} \right) = \text{supp}(0) = \emptyset$$

This contradiction implies that all of the $a_i$ are 0, which means that $\hat{\sigma}_{\omega_i,b_i}$ are linearly independent, as desired. \qed

The above result holds for any non-polynomial activation function $\sigma$ and only relies on the fact that the Fourier tranform of $\sigma$ is supported away from 0. However, by analyzing the Fourier transform of the activation function $\sigma$ more carefully, we can obtain stronger results for specific activation functions. As an example, we recover the result in (9) for the rectified linear unit.

**Theorem 3.** Suppose that $\sigma(t) = \max(0,t)$ and $\omega_i \neq 0$ for any $i$. Then the functions $\sigma(\omega_i \cdot x + b_i)$ are linearly independent if $(\omega_i, b_i) \neq \lambda (\omega_j, b_j)$ for all $i \neq j$.

**Proof.** The proof begins by calculating the Fourier transform of the rectified linear unit, for which we can make use of the well-known fact that the transform of $\text{sgn}(t)$ is the Hilbert transform (see (4), Chapter 3), i.e.

$$\langle \text{sgn}(t), f(t) \rangle = \text{p.v.} \int_{\mathbb{R}} \frac{1}{\pi t} f(t) dt$$

(28)
We now take $f$. This allows us to calculate the Fourier transform of $\sigma(t) = \frac{1}{2}(-i \text{sgn}(t) + 1)$ and we obtain

$$\langle \hat{\sigma}, f \rangle = \frac{f'(0)}{2} - \text{p.v.} \int_{\mathbb{R}} \frac{i}{2\pi t} f'(t) dt$$

(29)

Now suppose that the set $\{\sigma_{\omega, b}\}$ is linearly dependent. We will show that there exist $i \neq j$ such that $(\omega_i, b_i) = \lambda(\omega_j, b_j)$.

Since the $\{\sigma_{\omega, b}\}$ are linearly dependent, the proof of the previous theorem implies that there must be a dependence relation among each set of $\sigma_{\omega, b}$ for which the $\omega$ point in the same direction, i.e. there must be a dependence relation among the $\sigma_{\omega, b}$ for each occurring direction $\omega$ (with $|\omega| = 1$). Note that $\lambda_i \neq 0$ since $\omega_k \neq 0$.

So we have a dependence relation of the form

$$0 = \sum_{i=1}^{n} a_i \langle \sigma_{\lambda_i, b_i}, f \rangle = \sum_{i=1}^{n} a_i \langle \sigma_{\lambda_i, b_i}, f \rangle$$

(30)

where all $a_i \neq 0$ (discarding any indices for which $a_i$ vanishes) and which holds for every Schwartz function $f \in \mathbb{R}^d$.

Applying Lemma 3 and equation 29 and making the change of variables $t' = \lambda t$, this becomes

$$0 = \frac{1}{2} \sum_{i=1}^{n} a_i (\lambda_i f'(0) + b_i f(0)) - \text{p.v.} \int_{\mathbb{R}} \frac{i}{2\pi t} \left( \sum_{i=1}^{n} a_i \text{sgn}(\lambda_i) e^{i\pi(b_i/\lambda_i)} (\lambda_i f'(t) + b_i f(t)) \right) dt$$

(31)

for every Schwartz function $f \in \mathbb{R}$.

We now take $f$ to be a compactly supported bump function localized away from 0, so that the principal value above becomes a bona-fide integral and we get

$$0 = \int_{\text{supp}(f)} \frac{1}{t} \left( \sum_{i=1}^{n} a_i \text{sgn}(\lambda_i) e^{i\pi(b_i/\lambda_i)} (\lambda_i f'(t) + b_i f(t)) \right) dt$$

(32)

after which a simple integration by parts yields

$$0 = \int_{\text{supp}(f)} \frac{f(t)}{t} \left( \sum_{i=1}^{n} a_i \text{sgn}(\lambda_i) (i\pi b_i + b_i - t^{-1}\lambda_i) e^{i\pi(b_i/\lambda_i)} \right) dt$$

(33)

Taking $f$ to be a highly localized bump function, this means that

$$\sum_{i=1}^{n} a_i \text{sgn}(\lambda_i) (i\pi b_i + b_i - t^{-1}\lambda_i) e^{i\pi(b_i/\lambda_i)} = 0$$

(34)

for all $t \neq 0$. In fact, since each summand is a meromorphic function, this equation actually holds for all complex $t \neq 0$.

Letting $t \to \infty$ along the imaginary axis, we see that the terms $e^{i\pi(b_i/\lambda_i)}$ and $t^{-1} e^{i\pi(b_i/\lambda_i)}$ where $b_i/\lambda_i = r_{\max}$ is maximal dominate, and we consequently must have

$$\sum_{i: (b_i/\lambda_i) = r_{\max}} a_i \text{sgn}(\lambda_i) (i\pi b_i + b_i) = 0$$

(35)

and

$$\sum_{i: (b_i/\lambda_i) = r_{\max}} a_i \text{sgn}(\lambda_i) \lambda_i = 0$$

(36)

Since the $a_i, \lambda_i \neq 0$, this implies that the above sum must have at least two terms and so there exist $(\lambda_i, b_i)$ and $(\lambda_j, b_j)$ such that $b_i/\lambda_i = b_j/\lambda_j = r_{\max}$. Hence there exist $i \neq j$ such that $(\omega_i, b_i) = \lambda(\omega_j, b_j)$ as desired.

\[\square\]

### 4 Convergence Rates in Sobolev Norms

In this section, we study the order of approximation for two-layer neural networks as the number of neurons increases. In particular, we consider the space of functions represented by a two-layer neural network with $n$ neurons and activation function $\sigma$

$$\Sigma_n(\sigma) = \left\{ \sum_{i=1}^{n} \beta_i \sigma(\omega_i \cdot x + b_i) : \omega_i \in \mathbb{R}^d, b_i, \beta_i \in \mathbb{R} \right\}$$

(37)
and ask the following question: Given a function \( f \) on a bounded domain, how many neurons do we need to approximate \( f \) with a given accuracy?

We obtain the following result, which shows that the order of approximation is independent of the dimension \( d \) as long as \( f \) has a Fourier transform which decays sufficiently fast. A similar result can be found in [2], but we relax the assumptions made on the activation function \( \sigma \) and improve the rate of approximation by a logarithmic factor.

**Theorem 4.** Let \( \Omega \subset \mathbb{R}^d \) be a bounded domain with Lipschitz boundary. Assume that \( f \in H^m(\Omega) \) satisfies

\[
\gamma(f) = \int_{\mathbb{R}^d} (1 + |\omega|)^{m+1} |\hat{f}_e(\omega)| \, d\omega < \infty
\]

for some extension \( f_\varepsilon \in H^m(\mathbb{R}^d) \) with \( f_\varepsilon |_\Omega = f \). Then, if the activation function \( \sigma \in W^{m,\infty}(\mathbb{R}) \) is non-zero and satisfies the polynomial decay condition

\[
|\langle D^k \sigma(\cdot) \rangle| \leq C_m (1 + |\cdot|)^{-p}
\]

for \( 0 \leq k \leq m \) and some \( p > 1 \), we have

\[
\inf_{f_\varepsilon \in \Sigma_m(\sigma)} \|f - f_\varepsilon\|_{H^m(\Omega)} \leq C(p,m,\Omega,\sigma) \gamma(f) n^{-\frac{1}{2}}
\]

**Proof.** Observe first that translating \( \Omega \) doesn’t change the theorem and so we can assume without loss of generality that \( \Omega \) is centered at 0.

We also remark that the function \( f \in H^m(\Omega) \) can always be extended in a continuous fashion to \( H^m(\mathbb{R}^d) \) since \( \partial \Omega \) is Lipschitz (see [3, 11]), and thus the extension \( f_\varepsilon \) in the statement of the theorem certainly exists.

Note that the growth condition on \( \sigma \) implies that \( \sigma \in L^1(\mathbb{R}) \) and thus the Fourier transform of \( \sigma \) is well-defined and continuous. Since \( \sigma \) is non-zero, this implies that \( \hat{\sigma}(a) \neq 0 \) for some \( a \neq 0 \). Via a change of variables, this means that for all \( x \) and \( \omega \), we have

\[
0 \neq \hat{\sigma}(a) = \frac{1}{2\pi} \int_{\mathbb{R}} \sigma(\omega \cdot x + b) e^{-i a \omega \cdot x} db
\]

and so

\[
e e^{i a \omega \cdot x} = \frac{1}{2\pi \hat{\sigma}(a)} \int_{\mathbb{R}} \sigma(\omega \cdot x + b) e^{-i a \omega \cdot x} db
\]

Likewise, since the growth condition also implies that \( D^k \sigma \in L^1 \), we can differentiate the above expression under the integral with respect to \( x \).

This allows us to write the Fourier mode \( e^{i a \omega \cdot x} \) as an integral of neuron output functions. We substitute this into the Fourier representation of \( f \) (note that the assumption we make implies that \( f_\varepsilon \in L^1 \) so this is rigorously justified) to get

\[
f(x) = f_\varepsilon(x) = \int_{\mathbb{R}^d} e^{i a \omega \cdot x} \hat{f}_e(\omega) d\omega = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \frac{1}{2\pi \hat{\sigma}(a)} \sigma \left( \frac{a}{a} \cdot x + b \right) \hat{f}_e(\omega) e^{-i a \omega \cdot x} db d\omega
\]

The previous remark about differentiating this under the integral with respect to \( x \) implies that we also have

\[
D_x^\alpha f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \frac{1}{2\pi \hat{\sigma}(a)} D_x^\alpha \sigma \left( \frac{a}{a} \cdot x + b \right) \hat{f}_e(\omega) e^{-i a \omega \cdot x} db d\omega
\]

for all \( |\alpha| \leq m \).

Now let \( R \) be the maximum norm of an element of \( \Omega \) and consider the function

\[
h(b, \omega) = \left( 1 + \max \left( 0, |b| - \frac{|R|}{|a|} \right) \right)^{-p}
\]

A simple calculation implies that

\[
\int_{\mathbb{R}} h(b, \omega) db = 2R|a|^{-1} |\omega| + C(p) \leq C_1(p, \Omega, \sigma)(1 + |\omega|)
\]

Combined with our assumption on the Fourier transform, we get

\[
l(p, \Omega, \sigma, f) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} (1 + |\omega|)^m h(b, \omega) |\hat{f}_e(\omega)| db d\omega \leq C_1(p, \Omega, \sigma) \gamma(f)
\]
We use this to introduce a probability measure $\lambda$ on $\mathbb{R}^{d+1} = (x, \omega)$ given by

$$d\lambda = \frac{1}{I(p, \Omega, \sigma, f)}(1 + |\omega|^m h(b, \omega)|f_\omega(\omega)|dbd\omega$$

which allows us to write, for $|\alpha| \leq m$

$$D_x^\alpha f(x) = J(p, \Omega, \sigma, f)E_{d\lambda} \left((1 + |\omega|)^{-m}h(b, \omega)^{-1}e^{i\theta(\hat{f}_\omega(\omega))}D_x^\alpha \sigma \left(\frac{\omega}{a} \cdot x + b\right)\right)$$

where $J(p, \Omega, \sigma, f) = (2\pi^\frac{d}{2} \hat{\sigma}(a))^{-1}I(p, \Omega, \sigma, f)$ and $\theta(\cdot)$ denotes the phase of a complex number.

So we get

$$D_x^\alpha f(x) = J(p, \Omega, \sigma, f)E_{d\lambda} \left((1 + |\omega|)^{-m}h(b, \omega)^{-1}e^{i\theta(x)}D_x^\alpha \sigma \left(\frac{\omega}{a} \cdot x + b\right)\right)$$

for $\theta(x, \omega)$ a phase depending on $\omega$ and $b$.

We now proceed to draw $n$ samples $D_n = (x_1, b_1), ..., (x_n, b_n)$ from $d\lambda$ and note that elementary probability (the sample variance) implies that for every $x \in \Omega$

$$\mathbb{E}_{D_n} \left(D_x^\alpha f(x) - \sum_{i=1}^n a_i D_x^\alpha (a^{-1}\omega x + b_i)\right)^2 \leq \frac{1}{n} (J(p, \Omega, \sigma, f)\| (1 + |\omega|)^{-m} h(b, \omega)^{-1} D_x^\alpha \sigma (a^{-1} \omega \cdot x + b) \|_\infty^2)$$

where the coefficient $a_i = J(p, \Omega, \sigma, f)(1 + |\omega|)^{-m} h(b_i, \omega)^{-1}$, since the variance of a random variable is bounded by its maximum squared.

We proceed to consider the term

$$\| (1 + |\omega|)^{-m} h(b, \omega)^{-1} D_x^\alpha \sigma (a^{-1} \omega \cdot x + b) \|_\infty$$

Since $|\alpha| \leq m$ we immediately obtain

$$|D_x^\alpha \sigma (a^{-1} \omega \cdot x + b)| \leq |a|^{-m} (1 + |\omega|^m |D_x^\alpha \sigma (a^{-1} \omega \cdot x + b)|$$

which implies

$$\| (1 + |\omega|)^{-m} h(b, \omega)^{-1} D_x^\alpha \sigma (a^{-1} \omega \cdot x + b) \|_\infty \leq |a|^{-m} \| h(b, \omega)^{-1} D_x^\alpha \sigma (a^{-1} \omega \cdot x + b) \|_\infty$$

and for all $x \in \Omega$, $|x| \leq R$, which means that

$$|a^{-1} \omega \cdot x + b| \geq \max \left(0, |b| - \frac{R|\omega|}{|a|}\right)$$

for all $x \in \Omega$. Plugging this into equation (55) we obtain (uniformly in $x \in \Omega$)

$$h(b, \omega)^{-1} = \left(1 + \max \left(0, |b| - \frac{R|\omega|}{|a|}\right)\right)^p \leq (1 + |a^{-1} \omega \cdot x + b|)^p$$

Utilizing our assumption on the growth of $\sigma$, we get

$$\| h(b, \omega)^{-1} D_x^\alpha \sigma (a^{-1} \omega \cdot x + b) \|_\infty \leq \| (D_x^\alpha \sigma(t))(1 + |t|)^p \|_L^{-p(\alpha, \Omega)} \leq M < \infty$$

for all $|\alpha| \leq m$ and $x \in \Omega$. Plugging this into (54) and (51) we obtain

$$\mathbb{E}_{D_n} \left(D_x^\alpha f(x) - \sum_{i=1}^n a_i D_x^\alpha (a^{-1}\omega x + b_i)\right)^2 \leq \frac{1}{n} (J(p, \Omega, \sigma, f)|a|^{-m} M)^2$$

for all $|\alpha| \leq m$ and $x \in \Omega$. Integrating this over $\Omega$, summing over $|\alpha| \leq m$, taking the square root, and recalling the definition of $J$ and the bound in (47) yields

$$\mathbb{E}_{D_n} \left(\int f - \sum_{i=1}^n a_i \sigma(a^{-1}\omega x + b_i)\right)^2 \int_{\mathbb{R}^{d+1}} |\omega|^\frac{1}{2} J(p, \Omega, \sigma, f)|a|^{-m} M n^{-\frac{1}{2}} \leq C(p, m, \Omega, \sigma) \gamma(f) n^{-\frac{1}{2}}$$
This finally implies that there must exist a sample $D_n$ such that the quantity inside of the expectation is bounded by the left hand side, which means that

$$\inf_{f_n \in \Sigma_n(\sigma)} \|f - f_n\|_{H^m(\Omega)} \leq C(p,m,\Omega,\sigma)\gamma_n(f) n^{-\frac{1}{2}} \quad (60)$$

as desired. \[ \Box \]

Finally, we note that the approximation rate in this theorem holds as long as the growth condition hold for some $f \in \Sigma_d(\sigma)$, i.e. the condition need not hold for $\sigma$ itself. This includes many popular activation functions, such as the rectified linear units and sigmoidal activation functions.

5 Conclusion

We have provided two new results in the theory of approximation by neural networks. One result gives conditions under which the output from each of the neurons are linearly independent and the other improves existing results on the rate of approximation achieved by two-layer neural network as the number of neurons increases. Additionally, we have shown that a Fourier analysis argument can be used to greatly simplify an existing density result for two-layer neural networks.

One of the main questions which remain concerns our result on the rate of approximation. We believe that the assumptions on the activation function can be further relaxed, in particular, we conjecture that the growth condition can be entirely removed.

References

[1] Attali, J.G., Pagès, G.: Approximations of functions by a multilayer perceptron: a new approach. Neural networks 10(6), 1069–1081 (1997)
[2] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory 39(3), 930–945 (1993)
[3] Demengel, F., Demengel, G., Erné, R.: Functional spaces for the theory of elliptic partial differential equations. Springer (2012)
[4] Duoandikoetxea, J., Zuazo, J.D.: Fourier analysis, vol. 29. American Mathematical Soc. (2001)
[5] Ellacott, S.: Aspects of the numerical analysis of neural networks. Acta Numerica 3, 145–202 (1994)
[6] He, J., Li, L., Xu, J., Zheng, C.: Relu deep neural networks and linear finite elements. arXiv preprint arXiv:1807.03973 (2018)
[7] Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural networks 4(2), 251–257 (1991)
[8] Hornik, K.: Some new results on neural network approximation. Neural networks 6(8), 1069–1072 (1993)
[9] Hornik, K., Stinckcombe, M., White, H., Auer, P.: Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Computation 6(6), 1262–1275 (1994)
[10] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436 (2015)
[11] Leoni, G.: A first course in Sobolev spaces, vol. 181. American Mathematical Soc. (2017)
[12] Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural networks 6(6), 861–867 (1993)
[13] Lu, Z., Pu, H., Wang, F., Hu, Z., Wang, L.: The expressive power of neural networks: A view from the width. In: Advances in Neural Information Processing Systems, pp. 6231–6239 (2017)
[14] Mhaskar, H., Micchelli, C.A.: Degree of approximation by neural and translation networks with a single hidden layer. Advances in applied mathematics 16(2), 151–183 (1995)
[15] Mhaskar, H.N.: Neural networks for optimal approximation of smooth and analytic functions. Neural computation 8(1), 164–177 (1996)
[16] Mhaskar, H.N., Micchelli, C.A.: Dimension-independent bounds on the degree of approximation by neural networks. IBM Journal of Research and Development 38(3), 277–284 (1994)
[17] Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807–814 (2010)

[18] Pinkus, A.: Approximation theory of the mlp model in neural networks. Acta numerica 8, 143–195 (1999)

[19] Stein, E.M., Weiss, G.: Introduction to Fourier analysis on Euclidean spaces (PMS-32), vol. 32. Princeton university press (2016)

[20] Stone, M.H.: The generalized weierstrass approximation theorem. Mathematics Magazine 21(5), 237–254 (1948)

[21] Strichartz, R.S.: A guide to distribution theory and Fourier transforms. World Scientific Publishing Company (2003)

[22] Yarotsky, D.: Error bounds for approximations with deep relu networks. Neural Networks 94, 103–114 (2017)

[23] Zhou, D.X.: Universality of deep convolutional neural networks. arXiv preprint arXiv:1805.10769 (2018)