Universal Approximation properties of Feedforward Artificial Neural Networks

A thesis submitted in partial fulfilment of the
requirements for the degree of

MASTER OF SCIENCE

of

RHODES UNIVERSITY

by

STUART FREDERICK REDPATH

October 3, 2010

Abstract

In this thesis we summarise several results in the literature which show the approximation capabilities of multilayer feedforward artificial neural networks. We show that multilayer feedforward artificial neural networks are capable of approximating continuous and measurable functions from $\mathbb{R}^n \to \mathbb{R}$ to any degree of accuracy under certain conditions.

In particular making use of the Stone-Weierstrass and Hahn-Banach theorems, we show that a multilayer feedforward artificial neural network can approximate any continuous function to any degree of accuracy, by using either an arbitrary squashing function or any continuous sigmoidal function for activation.

Making use of the Stone-Weirstrass Theorem again, we extend these approximation capabilities of multilayer feedforward artificial neural networks to the space of measurable functions under any probability measure.

Contents

List of Figures

Acknowledgements

My never-ending gratitude must go to my supervisor, Professor Mike Burton, and my co-supervisor, Professor Gunther Jäger, for their patience and encouragement through the testing times in producing this thesis.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed, and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Thanks must go to John Gillam of the Rhodes University Postgraduate Financial Aid Office for his help, and the Mellon Mentors Scholarship Fund for financial assistance.

Thank you to Andrew Craig and Zukisa Xotyeni, my academic peers, and to Joanne Barry, my Mellon Mentors Student, for their continued daily support and trusted friendship.

I would finally like to thank Professor Sizwe Mabizela for his guidance and understanding throughout my postgraduate years.

Chapter 1

Introduction

1.1 What is an Artificial Neural Network ?

Artificial Neural Networks (ANN) are non-linear mapping systems which try to simulate the structural and functional aspects of Biological Neural Networks, originating from the recognition that the brain operates in an entirely different way from that of the conventional computer. The brain is a highly complex, non-linear and parallel information processing system which is capable of performing visual recognition tasks in the order of 100-200 ms. In contrast, tasks of far less complexity would take days on a modern computer [16].

The idea is that creating groups of processing units linked together in appropriate ways can generate many complex and interesting behaviours. This stems from the connectionist approach to computation, that even though a single processing unit may not be very powerful, the system may exhibit power by virtue of the combination of such processing units [27, 31].

1.2 The Biological Model

In neuroscience, a neural network describes a collection of physically connected neurons whose inputs or signalling targets define a recognisable cir-

cuit. This means that a neuron is the basic processing unit of the central nervous system, with the communication between the various neurons involving an electrochemical process [16, 32].

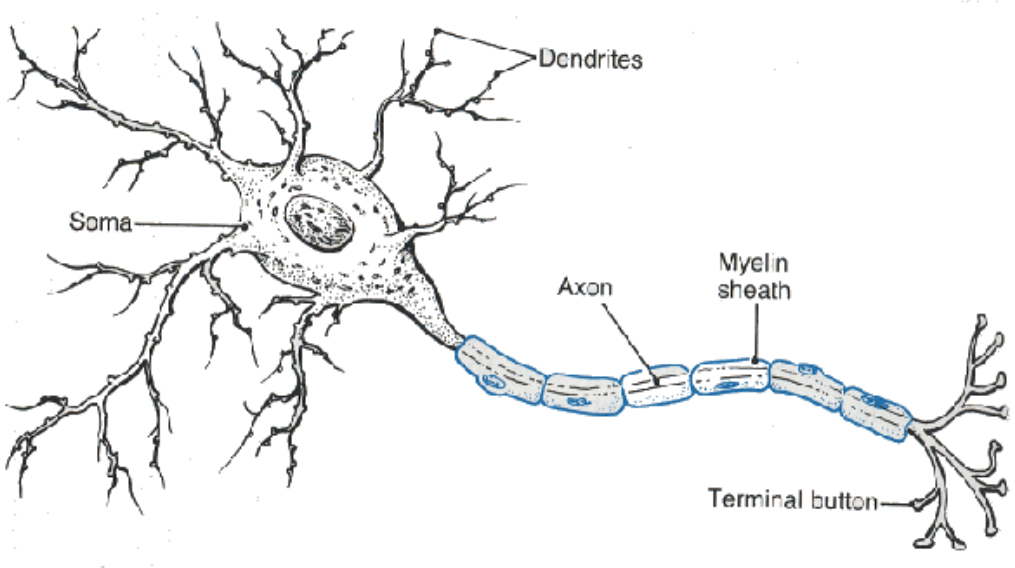

Figure 1.1: Biological Neuron [1].

The main regions to the structure of a neuron are; The cell body, or soma, out of which branch the dendrites and the axon which end in pre-synaptic terminals. The means by which the neurons interact are through several dendrites, tree-like structures that serve as input connections, and are connected via synapses to other neurons, and one axon, which grows out from a part of the cell body called an axon hillock and serves as an output connection. See Figure 1.1.

If the sum of the input signals from the dendrites to a neuron exceeds a certain threshold, the neuron sends an action potential (AP) at the axon hillock and transmits this electrochemical signal along the axon. The other end of the axon maybe split into several branches itself, all of which end in pre-synaptic terminals. Therefore action potentials are the electrochemical signals that neurons use to convey information to the brain [16].

1.3 The Mathematical Model

It is widely accepted that the neuron is the basic processing unit of a biological neural network. We therefore begin by creating a functional model of a neuron.

1.3.1 The McCulloch-Pitts Model

In order to do this we simplify the biological processes involved and identify three main elements of the neuron upon which we base the model. These are the *synapses, adder, and the activation function.*

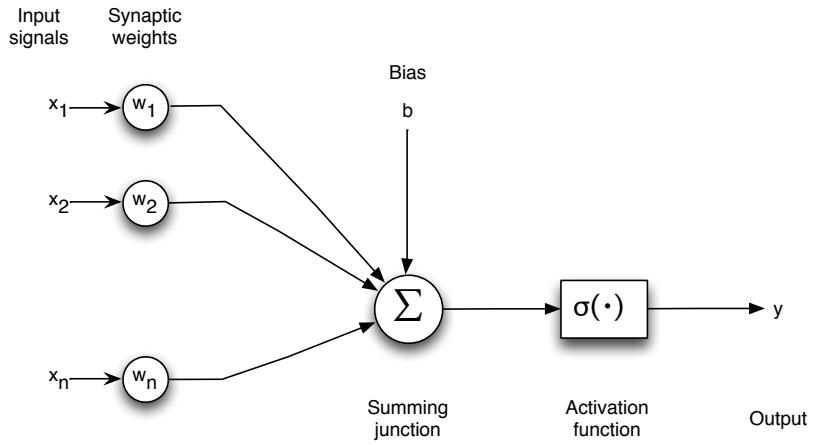

Figure 1.2: Artificial Neuron.

The *synapses* are each modelled as a weight. So that if x_j is a signal at the input of synapse j, which is connected to neuron k, then w_{kj} is the synaptic weight by which x_j is multiplied. This weight w_{kj} represents the strength of the connection along a particular synapse. Where negative weight values reflect inhibitory connections, while positive weight values designate excitatory connections [16].

The next two components model the actual activity within the cell body. The *adder* is used to sum up all the input signals, x_j, modified by their respective weights, w_{kj}, for a particular neuron k. This is simply a linear

combination of the input signals into the neuron. Finally, an *activation function*, σ_k, is used to control the amplitude of the output of the neuron k, signifying the action potential along the axon.

We also include a *bias* term, b_k, which is used to represent an externally applied threshold for a neuron k. This controls the *firing* of the neuron, by increasing or decreasing the net input into the activation function.

This means that if we let $x_1, \ldots, x_n$ be the input signals, $w_1, \ldots, w_n$ be the synaptic weights, with $b \in \mathbb{R}$ the bias, and σ the activation function for a neuron. We represent the output signal, y, of the neuron by the following equation

$$y = \sigma\left(\sum_{j=1}^{n} w_j x_j + b\right). \tag{1.1}$$

Such a neuron model is referred to as the *McCulloch-Pitts model*, after the work done by McCulloch and Pitts [24]. See Figure 1.2.

In Equation 1.1, the linear combination of the input signals $x_1, \ldots, x_n$ summed with the bias b, form what is known as an *Affine function*. We use the following notation to represent this linear combination for a particular neuron k.

Notation 1.3.1. The *Affine* function A_k may be viewed as a weighted sum of the input signals $x = (x_1, \ldots, x_n)$ added to the bias, b_k, for some neuron k with synaptic weights $W_{k1}, \ldots, W_{kn}$. Where

$$A_k(x) = \sum_{j=1}^{n} W_{kj} x_i + b_k. \tag{1.2}$$

1.3.2 Types of Activation Functions

The output of an *activation function*, $\sigma : \mathbb{R} \to \mathbb{R}$, is used to define the output signal of a neuron from a modified combination of its input signals by compressing the signal. Usually between the values $0 \leq \sigma(x) \leq 1$ or

$-1 \leq \sigma(x) \leq 1$. The three basic types of activation functions that are commonly used are the

1. Threshold function:

$$\sigma(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{1.3}$$

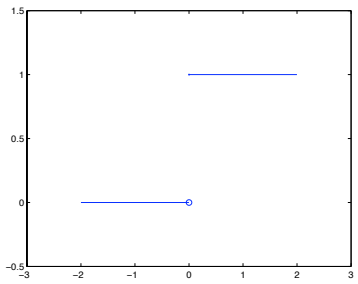

Figure 1.3: Threshold function.

2. Piecewise-linear function:

$$\sigma(x) = \begin{cases} 1 & \text{if } \frac{1}{2} \leq x, \\ x + \frac{1}{2} & \text{if } -\frac{1}{2} < x < \frac{1}{2}, \\ 0 & \text{if } x \leq -\frac{1}{2}. \end{cases} \tag{1.4}$$

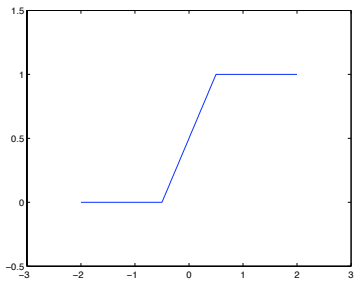

Figure 1.4: Piecewise Linear function.

3. Sigmoid function: A *sigmoidal function* is a function, σ, that is increasing, continuously differentiable, and has asymptotic properties. Such as for $a, b \in \mathbb{R}$,

$$\lim_{x \to \infty} \sigma(x) = a,$$
$$\lim_{x \to -\infty} \sigma(x) = b.$$

One of the most commonly used sigmoid functions is the *logistic function*, with slope parameter $\alpha \in \mathbb{R}$, defined by

$$\sigma(x) = \frac{1}{1 + e^{-\alpha x}}. \tag{1.5}$$

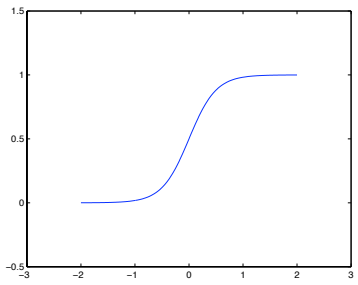

Figure 1.5: Sigmoidal function - The Logistic function.

1.4 Network Topologies

In section 1.2, we described a neural network as a collection of physically connected neurons whose inputs or signalling targets defined a recognisable circuit. Keeping this in mind, we define an artificial neural network as an interconnected structure of artificial neurons [16]. We will identify two different types of network topologies.

1.4.1 Multilayer Feedforward Networks

A *layered* neural network is a network of neurons arranged in the form of layers. For instance, we have an input layer of source neurons that connect to

an output layer of neurons, but not visa versa. Which means that this type of network is strictly *feedforward* or *acyclic*. We call this network a *single layer* network, not counting the input layer as no computation is performed at those neurons.

Another type of feedforward neural network is one which includes one or more so called *hidden layers*, whose neurons are referred to as *hidden neurons*. These are merely layers of neurons that lie between the input layer and the output layer of the network. With a hidden layer L connecting to the next hidden layer $L+1$ in the same fashion described above for the single layer network. The function of these hidden layers is to extract higher-order statistics by providing further synaptic connections [31].

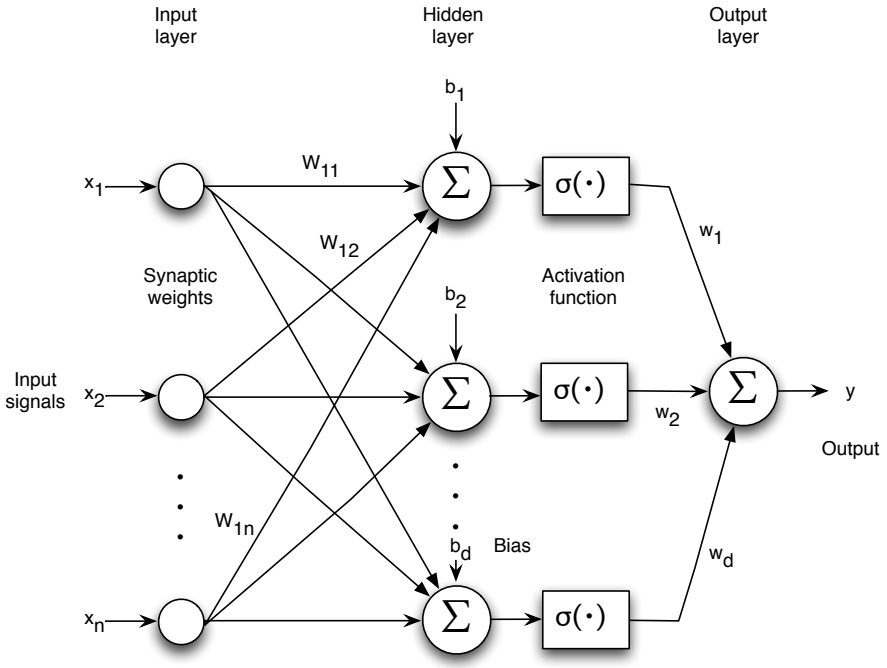

Figure 1.6: Multilayer Feedforward Artificial Neural Network.

If every neuron in each layer is connected to every other neuron in the adjacent forward layer, then the network is said to be *fully connected*. If this is not the case and some of the synaptic connections are missing, the

synaptic weights have a constant value of zero, then the network is said to be *partially connected*. Layered networks with one or more hidden layers of neurons are called *Multilayer feedforward networks*.

Notation 1.4.1. We denote by $\mathbb{R}^{\mathbb{R}^n}$, the set of all functions f from $\mathbb{R}^n$ to $\mathbb{R}$.

Definition 1.4.2. For any three layer feedforward neural network with an input layer of $n \in \mathbb{N}$ neurons, hidden layer of $d \in \mathbb{N}$ neurons and one output neuron, with input signal $x = (x_1, \ldots, x_n)$. We define an affine function A_k, see Notation 1.3.1, for each neuron k, with synaptic weights $W_{k1}, \ldots, W_{kn}$ and bias b_k, in the hidden layer to be

$$A_k(x) = \sum_{j=1}^{n} W_{kj} x_i + b_k.$$

Also for an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and synaptic weights $w_1, \ldots, w_d$ in the output layer consisting of one neuron, we define

$$\mathcal{N}_\sigma^n = \{ f \in \mathbb{R}^{\mathbb{R}^n} : f(x) = \sum_{j=1}^{d} w_j \sigma(A_j(x)) \}, \tag{1.6}$$

to denote the set of all functions from $\mathbb{R}^n \to \mathbb{R}$ with the above specified form. As standard notation we exclude the bias in the output layer, rather including another neuron in the hidden layer to act as the output layer's bias. This special hidden neuron will have constant input (eg: 1) with the weight of the neuron being the bias. Therefore $\mathcal{N}_\sigma^n$ represents a set of feedforward neural networks with n-dimensional inputs and activation function σ.

1.4.2 Recurrent Networks

A multilayer feedforward network is a network with a topology that has no *feedback* loops. Due to this such networks are limited to implementing a static mapping that depends only upon the present inputs and are independent of previous network states. If we allow for at least one feedback loop, we extend

the learning capabilities of such a neural network by introducing a non-linear dynamical behaviour which is due to the *unit delay* of particular feedback branches [31]. We call such networks *recurrent neural networks*, due to their recurrent network topology.

A difficulty that arises from these feedback loops is the training of these networks. How do we create algorithms which enable neural networks of this type to learn?

1.5 Training of Artificial Neural Networks

The primary reason for Artificial Neural Networks being so popular is due to their capability to approximate most processes found in applications, see Chapter 3, and their ability to learn from their environment. To improve a neural networks performance by learning about its environment a neural network goes through an iterative optimization procedure of adjustments to its synaptic weights and biases.

In effect this means that given some task to solve, the neural network searches for a solution f^* in a class of possible functions F which solves the task in some optimal sense. This involves defining a *cost function* $C : F \to \mathbb{R}$, such that for an optimal solution $f^* \in F$, we have that $C(f^*) \leq C(f)$, for all $f \in F$. The cost function is an important concept in learning, as it measures how far away a given solution is from the optimal solution. Learning algorithms search the solution space, F, for a function with the smallest cost.

In order to achieve this a variety of different learning paradigms have been used. We will mention three of them.

1.5.1 Supervised Learning

The idea with supervised learning is that the neural network is given a training set of pairs $\{(x, t) : x \in X, t \in T\}$ and must find a function $f^* : X \to T$, such that $f^*(x) = t$ for all the training samples. Assuming that there are no errors in the data. Here the cost function is based on the mismatch between

the network output for a training sample $f^*(x)$ and the desired target output t.

A commonly used cost function is the mean-squared error (MSE). This cost function tries to minimize the average squared error between the network output $f^*(x)$ and the target output t, over all training samples. When trying to minimise this cost function a derivation of the gradient descent algorithm is used, the so called *backpropagation algorithm* [26, 31].

1.5.2 Unsupervised Learning

In unsupervised learning a neural network is trained to respond to clusters of patterns within the input data. In this paradigm the network is supposed to discover statistically salient features of the input population. However, unlike with supervised learning there are no predefined categories into which the patterns can be classified. We therefore have to choose a cost function which is dependent on the task at hand and our priori knowledge available.

1.5.3 Reinforcement Learning

Reinforcement learning may be considered as an intermediate form of the above two paradigms. In this paradigm a neural network acts on some input from its environment and gains some form of feedback response, an output. Where by the network action is graded as good, rewarding, or bad, punishable, and based on the environments response the synaptic weights and biases are adjusted accordingly.

This can be achieved by allowing the network to generate outputs y_i, for various inputs x_i, from the environment and some instantaneous cost c_i, associated to each of those outputs. The idea is then to find a policy which minimises some measure of the long term cost, ie: the cumulative cost $\sum_{i=1}^{n} c_i$.

1.6 Applications

The benefit of artificial neural networks is that they can be used to infer a function from observations. This is particularly practical when the complexity of the task makes the design of such a function unfeasible. The possible real life application areas include:

- Classification: Pattern and Sequence Recognition,
- Data Processing: Filtering and Clustering,
- Robotics: Direction Control Manipulators,
- Regression Analysis: Time Series Prediction and Function Approximation.

In the coming chapters we will show how a multilayer feedforward artificial neural network is capable of approximating any continuous or measurable function to any degree of accuracy [3, 7, 9, 18, 19]. In particular we will determine what properties of the activation function and of the input space are required. Finally we will show that it is in fact the *multilayer feedforward architecture* itself which gives artificial neural networks the potential of being a *universal approximator* [20].

Chapter 2

Preliminaries

In this chapter we shall provide all the necessary background concepts, definitions, and theorems required for proving the various approximation properties of feedforward artificial neural networks.

2.1 Metric Spaces

We are interested in being able to approximate a certain class of functions. In order to do this we need to define what we mean by the *closeness* of two functions. We define this for an abstract class of mathematical objects.

Definition 2.1.1. A *Metric* on a set X is a non-negative real-valued function $\rho : X \times X \to \mathbb{R}$ with $x, y, z \in X$ obeying the rules

(M1) $\rho(x, y) \geq 0$,

(M2) $\rho(x, y) = 0$ if and only if $x = y$ (ρ is definite),

(M3) $\rho(y, x) = \rho(x, y)$ (ρ is symmetric),

(M4) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (the triangle inequality).

The number $\rho(x, y)$ is called the *distance* from x to y. A pair (X, ρ) where ρ is a metric on X, is called a *metric space*.

Therefore the closeness of two functions is measured by some appropriate metric.

Example 2.1.2 (Discrete Metric)**.** Let X be any non-empty set. For $x, y \in X$ define

$$\rho(x,y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{if } x = y. \end{cases}$$

This ρ is a metric, called the *discrete metric*.

Example 2.1.3 (The Finite-Dimensional Spaces $l_p^n = (\mathbb{R}^n, \rho_p)$)**.** Let $X = l_p^n$ for $(p \geq 1), n \in \mathbb{N}$. This is the finite-dimensional space of n-tuples of the form $x = (x_1, \ldots, x_n)$. Now for $x, y \in l_p^n$, let

$$\rho_p(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

Then (l_p^n, ρ_p) is a metric space.

Example 2.1.4. Let (X, ρ) and (Y, ϱ) be metric spaces. The two standard metrics for the product $X \times Y$ are

$$\hat{\rho}((x,y),(x',y')) = \rho(x,y) + \varrho(x',y'),$$
$$\hat{\varrho}((x,y),(x',y')) = \max\{\rho(x,y), \varrho(x',y')\}.$$

We will now define the basic topological concepts in an abstract metric space (X, ρ).

Definition 2.1.5 (Closed and Open Balls)**.** Let (X, ρ) be a metric space. We define a *closed ball* round an element $a \in X$ of radius $r > 0$, to be a subset of X of the form

$$B(a,r) = \{x \in X : \rho(a,x) \leq r\}.$$

Similarly we define an *open ball* round an element $a \in X$ of radius $r > 0$, to be a subset of X of the form

$$U(a,r) = \{x \in X : \rho(a,x) < r\}.$$

Definition 2.1.6 (Neighbourhoods)**.** Let (X, ρ) be a metric space. We say that a subset $N_x \subseteq X$ is a *neighbourhood* of a point $x \in X$, if there exists an $r > 0$ such that $U(x, r) \subseteq N_x$. Informally we say a subset $N_x \subseteq X$ is a *neighbourhood* of a point $x \in X$ if N_x contains some ball round x. We say that a point x is an *interior point* of a set A if and only if $A \subseteq N_x$.

Definition 2.1.7 (Closure)**.** Let (X, ρ) be a metric space. We call a point $x \in X$ a *closure point* of a subset A if given any neighbourhood N_x of x, $N_x \cap A \neq \emptyset$. We say that each neighbourhood N_x of x *meets* A. We call the set $\overline{A}$, of all closure points of A, the *closure* of A.

Definition 2.1.8 (Open and Closed Sets)**.** We say that a set is *open* if each of its points is an interior point. Similarly we say that a set is *closed* if it contains all of its closure points.

Definition 2.1.9 (Subspace)**.** A subset A of a metric space (X, ρ) is a metric space under the *relative metric* ρ_A, where ρ_A is defined by restricting the metric ρ to $A \times A$. We call (A, ρ_A) a subspace of (X, ρ).

Definition 2.1.10 (Distance between Subsets)**.** Let A, B be non-empty subsets of a metric space (X, ρ). We define the *distance* from A to B to be

$$\rho(A, B) = \inf\{\rho(a, b) : a \in A, b \in B\}.$$

Definition 2.1.11 (Bounded)**.** A subset $A \subseteq X$, of a metric space (X, ρ), is *bounded* if its *diameter*,

$$diam(A) = \sup\{\rho(x, y) : x, y \in A\},$$

is finite. Alternatively, A is bounded if A lies inside some closed or open ball.

Definition 2.1.12 (Totally Bounded)**.** A finite set of points $\{x_1, \ldots, x_n\} \subseteq X$ is an *ϵ-net* for a subset A in a metric space (X, ρ), if closed balls $B(x_i, \epsilon)$ of radius ϵ round the x_i have the property that $A \subseteq \cup_{i=1}^{n} B(x_i, \epsilon)$. We say that the closed balls $B(x_i, \epsilon)$ cover the subset A. If A possesses an ϵ-net for each $\epsilon > 0$ we say it is *totally bounded*.

In order to prove that the set of functions generated by multilayer feedforward artificial neural networks can approximate any continuous function to any degree of accuracy, we need to define what we mean by the *closeness* of a set of functions to another. This is described by the concept of *denseness*.

Definition 2.1.13 (Denseness)**.** A subset S of a metric space (X, ρ) is ρ-dense in a subset T if for every $\epsilon > 0$ and for every $t \in T$ there is an $s \in S$ such that $\rho(s, t) < \epsilon$.

Theorem 2.1.14. *A subset S is ρ-dense in a subset T if and only if the* closure $\overline{S} \supseteq T$.

Proof. Assume that $\overline{S} \supseteq T$. We have that every $t \in T$ is a closure point of S. Hence for all $t \in T$ and any neighbourhood N_t of t, we have $N_t \cap S \neq \emptyset$. Therefore there exists an $\epsilon > 0$ with corresponding open ball $U(t, \epsilon) \subseteq N_t$ such that $U(t, \epsilon) \cap S \neq \emptyset$. This implies that there exists $s \in S$ with $\rho(s, t) < \epsilon$. Hence S is ρ-dense in a T.

Conversely assume S is ρ-dense in a T. This implies that for all $\epsilon > 0$ and for all $t \in T$, there exists $s \in S$ such that $\rho(s, t) < \epsilon$. We fix $\epsilon > 0$ and let N_t be a neighbourhood of $t \in T$ corresponding to ϵ. Then there exists $s \in S$ such that $s \in N_t$. Therefore $N_t \cap S \neq \emptyset$. Hence t is a closure point of S and $T \subseteq \overline{S}$. □

We defined an artificial neural network to be a set of functions, which are themselves a linear combination of functions, the so called activation functions. Here we define what we mean by scalar functions being combined and the operations on them.

Definition 2.1.15 (Pointwise Operations)**.** We define a *scalar function* to be a real-valued function $f : X \to \mathbb{R}$. For $f, g \in X$ being scalar functions, $\lambda \in \mathbb{R}$ and $x \in X$ they are combined by the *pointwise operations* in the

following way

$$(f+g)(x) = f(x) + g(x),$$
$$(fg)(x) = f(x)g(x),$$
$$(\lambda f)(x) = \lambda(f(x)).$$

Definition 2.1.16 (Continuity)**.** Let $f : X \to Y$ be a mapping from a metric space (X, ρ) to a metric space (Y, ϱ) and let $a \in X$. We then say that f is *continuous at* a if for all $\epsilon > 0$, there exists $\delta = \delta(a, \epsilon)$ such that for any $x \in X$,

$$\varrho(f(x), f(a)) < \epsilon \text{ whenever } \rho(x, a) < \delta\,.$$

If f is continuous at each such point $a \in X$ it is called *continuous on* X.

Definition 2.1.17 (Uniform Continuity)**.** Let (X, ρ) and (Y, ϱ) be metric spaces. We say that the mapping $f : X \to Y$ is *uniformly continuous* if given any $\epsilon > 0$, there exists $\delta = \delta(\epsilon)$ such that

$$\varrho(f(x), f(y)) < \epsilon \text{ whenever } x, y \in X \text{ and } \rho(x, y) < \delta\,.$$

Theorem 2.1.18 (Continuity of Functions. See [30] pg 21)**.** *If f, g are continuous scalar functions on a metric space (X, ρ) with $\lambda \in \mathbb{R}$, then $f + g$, fg, λf are continuous.*

Definition 2.1.19 (Convergence)**.** Let (X, ρ) be a metric space. We say that a sequence (x_n) in X *converges* to a point $x \in X$, if for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon, x) \in \mathbb{N}$ such that

$$\rho(x_n, x) < \epsilon \text{ whenever } n \geq n_0.$$

We use the following notation for convergence, $x_n \to x$ as $n \to \infty$. It follows that

$$x_n \to x \iff \rho(x_n, x) \to 0\,.$$

Theorem 2.1.20 (See [33] pg 42)**.** *Let (X, ρ) be a metric space. A sequence (x_n) in X* converges *to a point $x \in X$ $\iff$ Given any neighbourhood N_x of x, there exists $n_0 \in \mathbb{N}$ such that $x_n \in N_x$ for $n \geq n_0$. We say that x_n is eventually in N_x.*

Proof. Assume $x_n \to x$ and let N_x be a neighbourhood of X. For some $\epsilon > 0$ the conditions that $\rho(q, x) < \epsilon$ with $q \in X$ imply that $q \in N_x$. From the definition of convergence there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that $n \geq n_0$ implies $\rho(x_n, x) < \epsilon$. Thus $n \geq n_0$ implies $x_n \in N_x$.

Conversely, assume that for any neighbourhood N_x of x, there exists $n_0 \in \mathbb{N}$ such that $x_n \in N_x$ for $n \geq n_0$. Fix $\epsilon > 0$ and let V_x be the set of all $q \in X$ such that $\rho(q, x) < \epsilon$. From our assumption, there exists n_0 corresponding to this V_x, such that $x_n \in V_x$ if $n \geq n_0$. Thus $\rho(x_n, x) < \epsilon$ if $n \geq n_0$, hence $x_n \to x$. □

Lemma 2.1.21 (See [33] pg 42)**.** *A sequence (x_n) in a metric space (X, ρ) can converge to at most one point x, called the* limit *of (x_n).*

Proof. Assume that $x_n \to x$ and $x_n \to x^*$. Let $\epsilon > 0$ be given. There exist $n_0 = n_0(\epsilon), n_1 = n_1(\epsilon) \in \mathbb{N}$ such that

$$n \geq n_0 \text{ implies } \rho(x_n, x) < \frac{\epsilon}{2},$$
$$n \geq n_1 \text{ implies } \rho(x_n, x^*) < \frac{\epsilon}{2}.$$

Hence if $n \geq \max(n_0, n_1)$, we have

$$\rho(x, x^*) \leq \rho(x, x_n) + \rho(x^*, x_n) < \epsilon.$$

This is for any given $\epsilon > 0$ and so we conclude that $\rho(x, x^*) = 0$. From property (M2) for a metric, this implies that $x = x^*$. □

Definition 2.1.22 (Convergence of Functions)**.** We say that a sequence of scalar functions (f_n) on X *converges pointwise* to f, if given $\epsilon > 0$ and $x \in X$, then there exists $n_0 = n_0(\epsilon, x) \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \epsilon,$$

for all $n \geq n_0$.

We say that a sequence of scalar functions (f_n) on X *converges uniformly* to f, if given $\epsilon > 0$,then there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \epsilon,$$

for all $n \geq n_0$ and all $x \in X$.

Theorem 2.1.23 (Continuity of Functions. See [30] pg 21)**.** *Let (X, ρ) be a metric space. If the sequence (f_n) of continuous scalar functions on X converges uniformly to a function f, then f is also continuous.*

Definition 2.1.24 (Cauchy Sequences)**.** We say that a sequence (x_n) in a metric space (X, ρ) is *Cauchy* if given any $\epsilon > 0$, there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that

$$\rho(x_m, x_n) < \epsilon \text{ whenever } m, n \geq n_0.$$

Theorem 2.1.25 (See [33] pg 46)**.**

1. *Every convergent sequence (x_n) in a metric space (X, ρ) is Cauchy.*

2. *Every Cauchy sequence in $\mathbb{R}^n$ converges.*

Definition 2.1.26 (Completeness)**.** We say that a metric space (X, ρ) is *complete* if every Cauchy sequence is convergent. A subset $A \subseteq X$ is complete if and only if $(A, \rho_{|A})$ is complete.

Definition 2.1.27 (Compactness)**.** A subset C of a metric space (X, ρ) is *compact* if every family of open sets which covers C (whose union contains C), has a finite subfamily which also covers C.

Theorem 2.1.28 (See [30] pg 14)**.** *The following are equivalent for a metric space (X, ρ)*

1. *X is compact,*

2. *Every sequence in X has a convergent subsequence (the Bolzano-Weierstrass Property),*

3. X is complete and totally bounded.

Lemma 2.1.29 (See [30] pg 12)**.** *Let A be a complete subset of a metric space (X, ρ). Then A is closed. Conversely if A is a closed subset of a complete metric space (X, ρ). Then A is complete.*

Corollary 2.1.30 (See [30] pg 15)**.** *A subset of a complete metric space has compact closure if and only if it is totally bounded.*

Proof. Let A be a subset of a complete metric space (X, ρ). Assume $\overline{A}$ is compact. By Theorem 2.1.28, $\overline{A}$ is complete and totally bounded. Since $A \subseteq \overline{A}$ we have that for all $\epsilon > 0$, if $\{x_1, \ldots, x_n\}$ is an ϵ-net for $\overline{A}$ then $\{x_1, \ldots, x_n\}$ is an ϵ-net for A. Thus A is totally bounded.

Conversely, assume A is totally bounded. Fix $\epsilon > 0$ and let $\{x_1, \ldots, x_n\}$ be an ϵ-net for A. Then $A \subseteq \cup_{i=1}^{n} B(x_i, \epsilon)$. Let $x^* \in \overline{A}$. If $x^* \in A$ then $x^* \in B(x_i, \epsilon)$ for some $i \in \{1, \ldots, n\}$. Alternatively if $x^* \in \overline{A} \setminus A$, then since x^* is a closure point of A we have that for any neighbourhood N_{x^*} of x^*, $N_{x^*} \cap A \neq \emptyset$. Let $B(x^*, \epsilon) \subseteq N_{x^*}$, then $B(x^*, \epsilon) \cap A \neq \emptyset$ which implies $\rho(x^*, a) < \epsilon$ for some $a \in A$. Thus $x^* \in B(x_i, \epsilon)$ for some $i \in \{1, \ldots, n\}$. Therefore $\overline{A}$ is totally bounded. Also $\overline{A}$ is a closed subset of a complete metric space (X, ρ) and is therefore complete. By Theorem 2.1.28, $\overline{A}$ is compact. □

Theorem 2.1.31. *Let A be a subset of a complete metric space (X, ρ). If A is totally bounded then it is also bounded.*

Proof. Assume A is totally bounded. Then for any $\epsilon > 0$ there exists a ϵ-net $\{x_1, \ldots, x_n\}$, such that $A \subseteq \cup_{i=1}^{n} B(x_i, \epsilon)$. Each $B(x_i, \epsilon)$ is bounded and a finite union of bounded sets is bounded. Hence A is contained in a bounded set and is thus bounded. □

In a metric space (X, ρ), a compact subset $A \subset X$ must be closed, because complete subsets of metric spaces are closed. Further A must be bounded, because totally bounded subsets are bounded.

Theorem 2.1.32 (See [30] pg 13)**.** *Let (X, ρ) and (Y, ϱ) be complete metric spaces. Then $X \times Y$ is complete under either of the two standard product metrics.*

Theorem 2.1.33 (See [30] pg 15)**.** *Let (X, ρ) and (Y, ϱ) be compact metric spaces. Then $X \times Y$ is compact under either of the two standard product metrics.*

Theorem 2.1.34 (Continuous functions on Compact Spaces. See [30] pg 19)**.** *A continuous real-valued function f on a compact metric space X is bounded and attains its bounds in the sense that there exist $a, b \in X$ such that*

$$f(a) \leq f(x) \leq f(b),$$

for all $x \in X$.

Theorem 2.1.35 (See [30] pg 21)**.** *Let f be a continuous scalar function on a compact metric space (X, ρ). Then f is uniformly continuous.*

Proof. Given $\epsilon > 0, x, y \in X$ the sets

$$E_n = \left\{ (x, y) : |f(x) - f(y)| \geq \epsilon,\, \rho(x, y) \leq \frac{1}{n} \right\},$$

form a decreasing sequence of closed sets in the compact metric space $X \times X$, having empty intersection. But if $\cap_n E_n$ were empty then $\{X \setminus E_n\}$ would be a sequence of open sets covering X. But clearly no finite subfamily of it can cover X. Hence it must follow that for some $k \in \mathbb{N}$, $E_k = \emptyset$. Which implies that

$$|f(x) - f(y)| < \epsilon,$$

whenever $\rho(x, y) \leq \frac{1}{k}$. Therefore f is uniformly continuous. □

2.2 Normed Linear Spaces

Definition 2.2.1 (Linear Space)**.** Let X be a non-empty set and F a field of scalars with the following operations

$$+ : X \times X \to X \text{ (Vector addition)} ,$$
$$\cdot : F \times X \to X \text{ (Scalar multiplication).}$$

The set X is called a *linear space* over the field F if the following properties hold for any $x, y, z \in X$ and any $k, l \in F$

(A1) $(x + y) + z = x + (y + z)$,

(A2) There exists a vector in X called the *zero vector*, denoted by 0, such that

$$x + 0 = 0 + x ,$$

(A3) For every vector $x \in X$ there exists another vector called the *negative of* x, denoted by $-x$, such that

$$x + (-x) = (-x) + x = 0 ,$$

(A4) $(x + y) = y + x$,

(M1) $k(x + y) = kx + ky$,

(M2) $(k + l)x = kx + lx$,

(M3) $(kl)x = k(lx)$,

(M4) There exists a scalar in F called the *unit scalar*, denoted by 1, such that

$$1x = x .$$

A set $Y \subseteq X$ is called a *linear subspace* of X if Y is itself a linear space with respect to the above operations.

Lemma 2.2.2 (See [34] pg 5)**.** *For a linear space X over a scalar field F with $Y \subseteq X$. Then Y is a linear subspace of X if and only if for any $x, y \in Y$ and any $k, l \in F$*

$$kx + ly \in Y\,,$$
$$0 \in Y\,.$$

Definition 2.2.3 (Normed Linear Space)**.** A *Normed Linear Space* is a linear space X over a (real) field on which there is defined a real-valued function called the *norm*, $\|\cdot\| : X \to \mathbb{R}$ having the following properties. For all $x, y \in X$, $\alpha \in \mathbb{R}$ we have

(N1) $\|x\| \geq 0$,

(N2) $\|x\| = 0$ if and only $x = 0$,

(N3) $\|\alpha x\| = |\alpha|\,\|x\|$,

(N4) $\|x + y\| \leq \|x\| + \|y\|$ (the triangle inequality).

So a normed linear space is the tuple $(X, +, \cdot, \|\cdot\|)$. We think of the number $\|x\|$ as being the *length* of the vector x.

Example 2.2.4 (Norm Induced Metric)**.** For any Normed Linear Space X, let

$$\rho(x, y) = \|x - y\|.$$

This defines a metric on X. We always assume that a normed linear space carries this metric. Thus

$$x_n \to x \iff \|x_n - x\| \to 0.$$

Definition 2.2.5 (Banach Space)**.** A *Banach Space* is a normed linear space that, regarded as a metric space, is complete.

The variety of normed spaces which appear in analysis is vast, here are some which are of greater concern to us.

Example 2.2.6 (The Little Ell-p Spaces)**.** Let $X = l_p$ for $1 \leq p < \infty$. This is the space of all *p-summable sequences* $x = (x_i)_{i=1}^{\infty}$ which satisfy the condition

$$\sum_{i=1}^{\infty} |x_i|^p < \infty \,.$$

Now for $x, y \in l_p, \lambda \in \mathbb{R}$, we define the norm to be

$$\|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}} ,$$

with corresponding norm metric

$$\begin{aligned} \rho_p(x, y) &= \|x - y\|_p \\ &= \left(\sum_{i=1}^{\infty} |x_i - y_i|^p \right)^{\frac{1}{p}} , \end{aligned}$$

and vector operations defined to be

$$\begin{aligned} x + y &= (x_i + y_i)_{i=1}^{\infty} \,, \\ \lambda x &= (\lambda x_i)_{i=1}^{\infty} \,. \end{aligned}$$

Example 2.2.7 (The Little Ell-∞ Space)**.** Let $X = l_\infty$. This is the space of all *bounded sequences* $x = (x_i)_{i=1}^{\infty}$ which satisfy the condition

$$\sup_{i \in \mathbb{N}} |x_i| < \infty \,.$$

Now for $x, y \in l_\infty$, we define the norm to be

$$\|x\|_\infty = \sup_{i \in \mathbb{N}} |x_i| \,,$$

with corresponding norm metric

$$\rho_\infty(x, y) = \sup_{i \in \mathbb{N}} |x_i - y_i| \,,$$

and vector operations defined to be

$$\begin{aligned} x + y &= (x_i + y_i)_{i=1}^{\infty} \,, \\ \lambda x &= (\lambda x_i)_{i=1}^{\infty} \,. \end{aligned}$$

Definition 2.2.8. Let (X, ρ) be a metric space, then a function $f : X \to \mathbb{R}$ is said to be *bounded* if there exists $M \in \mathbb{R}$ such that for all $x \in X$

$$|f(x)| \leq M.$$

Let $\mathcal{B}(X)$ be the set of all such functions $f : X \to \mathbb{R}$.

Example 2.2.9 (Function Spaces. See [30] pg 47)**.** Let (X, ρ) be a metric space. For any bounded functions $f, g \in \mathcal{B}(X)$ on X and scalar $\lambda \in \mathbb{R}$, we define

$$\|f\|_\infty = \sup\{|f(x)| : x \in X\},$$

to be the *supremum norm* and $\mathcal{B}(X)$ is a normed linear space with vector operations defined for any $x \in X$ by

$$(f + g)(x) = f(x) + g(x),$$
$$(\lambda f)(x) = \lambda f(x).$$

Also if T is any compact metric space, the space $\mathcal{C}_T$ of continuous real-valued functions on T, with the norm $\|\cdot\|_\infty$, is a normed linear space.

Theorem 2.2.10 (See [30] pg 49)**.** *Let (X, ρ) be a metric space. The norm is a continuous function on X, and addition and scalar multiplication are jointly continuous functions on X.*

Notation 2.2.11. Let $\mathcal{C}^n$ denote the set of continuous functions from $\mathbb{R}^n \to \mathbb{R}$ and $\mathcal{C}^n_K$ the set of continuous functions restricted to a subset $K \subseteq \mathbb{R}^n$.

Theorem 2.2.12 (See [30] pg 53)**.** *Let (T, ρ) be a compact metric space. Then $\mathcal{C}_T$, with supremum norm*

$$\|f\|_\infty = \sup\{|f(x)| : x \in T\},$$

is complete. Therefore $\mathcal{C}_T$ is a Banach Space.

Proof. Let (f_n) be a Cauchy sequence of functions in $\mathcal{C}_T$. So in the supremum norm, this means that given any $\epsilon > 0$ there exists an $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that for $m, n \geq n_0$ and $t \in T$ we have

$$|f_m(t) - f_n(t)| \leq \epsilon.$$

This means that $(f_n(t))$ is a Cauchy sequence in $\mathbb{R}$. By the completeness of $\mathbb{R}$, we have that $(f_n(t))$ is convergent. So the pointwise limit

$$f(t) = \lim_{n\to\infty} f_n(t),$$

exists for all $t \in T$.

For $m > n_0$,

$$\begin{aligned} |f_m(t) - f(t)| &= \left| f_m(t) - \lim_{n\to\infty} f_n(t) \right| \\ &= \lim_{n\to\infty} |f_m(t) - f_n(t)| \\ &\leq \epsilon\,, \end{aligned}$$

as a result of the norm being continuous. This holds for all $t \in T$ which implies that $f_n \to f$ uniformly on T. Also

$$\forall t \in T,\, \forall m > n_0,\, |f_m(t) - f(t)| \leq \epsilon$$

and so

$$\forall m > n_0,\, \|f_m(t) - f(t)\|_\infty \leq \epsilon.$$

Hence

$$\|f_m - f\|_\infty \to 0.$$

From Theorem 2.1.23 we know that the uniform limit of continuous functions is continuous.

We have shown that a Cauchy sequence of functions (f_n) in $\mathcal{C}_T$ is convergent in $\mathcal{C}_T$. Therefore $\mathcal{C}_T$ is complete and is hence a Banach Space. □

2.3 Density Theorems for Continuous Functions

We use the concept of Density to prove that the closure of the set of functions generated by feedforward artificial neural networks $\mathcal{N}_\sigma^n$, is the same as the

set of all continuous functions $\mathcal{C}^n$. This can be interpreted that every element of $\mathcal{C}^n$ can be approximated by some element of $\mathcal{N}_\sigma^n$, (or $\mathcal{A}_\sigma^n$), to any degree of accuracy. In order to do this we will need the theorem of *Stone and Weierstrass*.

Example 2.3.1. Let $K \subseteq \mathbb{R}^n$ be any compact subset and $f, g, h \in \mathcal{C}_K^n$, the space of continuous scalar functions restricted to K. Then

$$\rho_K(f,g) = \sup_{x \in K} |f(x) - g(x)|, \tag{2.1}$$

is a metric.

Proof. For any $f, g \in \mathcal{C}_K^n$ and owing to the compactness of K we are ensured that $\rho_K(f,g) < \infty$. Next we must show that ρ_K satisfies the requirements in definition 2.1.1

$$\rho_K(f,g) \geq 0, \text{ sup of non-negative numbers is non-negative.}$$

$$\begin{aligned} \rho_K(f,g) = 0 &\iff \sup_{x \in K} |f(x) - g(x)| = 0 \\ &\iff f(x) = g(x), \forall x \in K \\ &\iff f = g \text{ on } K. \end{aligned}$$

$$\begin{aligned} \rho_K(f,g) &= \sup_{x \in K} |f(x) - g(x)| \\ &= \sup_{x \in K} |-1||g(x) - f(x)| \\ &= \rho_K(g,f). \end{aligned}$$

$$\begin{aligned} \rho_K(f,h) &= \sup_{x \in K} |f(x) - g(x) + g(x) - h(x)| \\ &\leq \sup_{x \in K} |f(x) - g(x)| + \sup_{x \in K} |g(x) - h(x)| \\ &= \rho_K(f,g) + \rho_K(g,h). \end{aligned}$$

Therefore ρ_K is a metric on $\mathcal{C}_K^n$ for any compact $K \subseteq \mathbb{R}^n$. □

Definition 2.3.2. A subset S of $\mathcal{C}^n$ is said to be *uniformly dense on compacta* in $\mathcal{C}^n$ if for every compact subset $K \subseteq \mathbb{R}^n$, S is ρ_K-dense in $\mathcal{C}^n$. Where

$$\rho_K(f,g) = \sup_{x \in K} |f(x) - g(x)| \quad \text{for } f, g \in \mathcal{C}^n_K. \tag{2.2}$$

This means that when the functions $f \in S \subseteq \mathcal{C}^n$ are restricted to a compact subset $K \subseteq \mathbb{R}^n$, $f_{|_K} : K \to \mathbb{R}$, the subset S is ρ_K-dense in $\mathcal{C}^n$.

A sequence of functions (f_n) converges to a function f *uniformly on compacta* if for all compact $K \subseteq \mathbb{R}^n$ we have that $\rho_K(f_n, f) \to 0$ as $n \to \infty$.

Definition 2.3.3. A family of real valued functions $A = \mathbb{R}^E$, defined on a set E is an *algebra* if A is closed under addition, multiplication and scalar multiplication.

Definition 2.3.4. A family of real valued functions $A = \mathbb{R}^E$, is said to *separate points* on a set E if for every distinct pair $x, y \in E$ there exists $f \in A$ such that $f(x) \neq f(y)$.

Definition 2.3.5. A family of real valued functions $A = \mathbb{R}^E$, is said to *vanish at no point* on a set E if for each $x \in E$ there exists $f \in A$ such that $f(x) \neq 0$.

Theorem 2.3.6 (Weierstrass Theorem. See [33] pg 146)**.** *If f is a continuous complex-valued function on the interval $[a, b]$, there exists a sequence of polynomials P_n such that*

$$\lim_{n \to \infty} P_n(x) = f(x),$$

uniformly on $[a, b]$. If f is real, the P_n may be taken real.

Definition 2.3.7. If a family of real valued functions $A = \mathbb{R}^E$, has the property that $f \in A$ whenever $f_n \in A$, for $n = 1, 2, 3, \ldots$ and $f_n \to f$ uniformly on E. Then A is said to be *uniformly closed*.

Definition 2.3.8. Let B be the set of all functions which are limits of uniformly convergent sequences of members of the family of functions A. Then B is called the *uniform closure* of A.

Example 2.3.9. The set of all polynomials P is an algebra, separates points and vanishes at no point. The Weierstrass theorem may be stated by saying that the set of continuous functions on the interval $[a, b]$ is the uniform closure of the set of polynomials on the interval $[a, b]$.

Theorem 2.3.10 (Stone-Weierstrass Theorem. See [33] pg 150)**.** *Let A be an algebra of real-valued continuous functions on a compact subset K. If A separates points on K and if A vanishes at no point on K, then the uniform closure B of A consists of all the real-valued continuous functions on K. Alternatively A is ρ_K-dense in the space of real continuous functions on K.*

2.4 Measure and Integration

In order to extend the approximation capabilities of multilayer feedforward artificial neural networks from continuous functions to measurable functions, we will need the following basic concepts, definitions, and theorems.

Definition 2.4.1 (σ-algebra)**.** A family $\mathbb{A}$ of subsets of a set X is said to be a σ-algebra if

(S1) $\emptyset, X \in \mathbb{A}$,

(S2) If $A \in \mathbb{A}$ then the complement $A^c = X \setminus A$ belongs to $\mathbb{A}$,

(S3) If (A_n) is a sequence of sets in $\mathbb{A}$, then the union $\cup_{n=1}^{\infty} A_n$ belongs to $\mathbb{A}$.

An ordered pair $(X, \mathbb{A})$ consisting of a set X and a σ-algebra $\mathbb{A}$ of subsets of X is called a *measurable space*. The sets in $\mathbb{A}$ are called $\mathbb{A}$-*measurable sets*, but when the σ-algebra is fixed they are usually referred to as being *measurable*.

Lemma 2.4.2 (See [8] pg 3)**.** *Let X be a non-empty set. Then the intersection of any non-empty collection of σ-algebras on X is a σ-algebra on X.*

Lemma 2.4.3 (See [8] pg 3)**.** *Let X be a non-empty set, and let $\mathbf{F}$ be a collection of subsets of X. Then there exists a smallest σ- algebra on X that includes $\mathbf{F}$. This smallest σ-algebra on X that includes $\mathbf{F}$ is clearly unique and is called the σ-algebra* generated by *F.*

Proof. Let $\mathbf{C}$ be the collection of all σ-algebras on X that include $\mathbf{F}$. Then $\mathbf{C}$ is non-empty, since it contains the σ-algebra that consists of all subsets of X. Due to Lemma 2.4.2, the intersection of the σ-algebras that belong to $\mathbf{C}$ is also a σ-algebra. This σ-algebra includes $\mathbf{F}$ and is included in every other σ-algebra on X that includes $\mathbf{F}$. □

Example 2.4.4. A particularly important σ-algebra in any metric space (X, ρ) is the *Borel σ-Algebra*. This is the σ-algebra $\mathbb{B}$ generated by all open sets in that metric space. If $X = \mathbb{R}$ then $\mathbb{B}$ is the σ- algebra generated by the open intervals (a, b) in $\mathbb{R}$. Similarly if $X = \mathbb{R}^n$ then $\mathbb{B}$ is the σ-algebra generated by the open subsets of $\mathbb{R}^n$.

Lemma 2.4.5 (See [8] pg 4)**.** *The σ-algebra $\mathbb{B}$ of Borel subsets of $\mathbb{R}$ is generated by each of the following collections of sets*

(B1) *the collection of all closed subsets of $\mathbb{R}$,*

(B2) *the collection of all subintervals of $\mathbb{R}$ of the form $(-\infty, b]$,*

(B3) *the collection of all subintervals of $\mathbb{R}$ of the form $(a, b]$.*

Lemma 2.4.6 (See [8] pg 5)**.** *The σ-algebra $\mathbb{B}$ of Borel subsets of $\mathbb{R}^n$ is generated by each of the following collections of sets*

(B1) *the collection of all closed subsets of $\mathbb{R}^n$,*

(B2) *the collection of all closed half-spaces in $\mathbb{R}^n$ that have the form*

$$\{(x_1, \ldots, x_n) : x_i \leq b\}$$

for some index $i \in \{1, \ldots, n\}$ *and some* $b \in \mathbb{R}$,

(B3) *the collection of all rectangles in* $\mathbb{R}^n$ *that have the form*

$$\{(x_1, \ldots, x_n) : a_i < x_i \leq b_i\}$$

for some $i \in \{1, \ldots, n\}$.

Definition 2.4.7. Let $(X, \mathbb{A})$ be a measurable space. A function $f : X \to \mathbb{R}$ is said to be $\mathbb{A}$-measurable, or simply measurable, if for every $\alpha \in \mathbb{R}$ the set

$$\{x \in X : f(x) > \alpha\} = f^{-1}(\alpha, \infty)$$

belongs to $\mathbb{A}$ or is measurable.

Example 2.4.8. Another particularly important σ-algebra in any metric space (X, ρ) is the *Baire σ-Algebra*. This is the σ-algebra $\mathbb{B}_a$ and is defined as the smallest σ-algebra such that all continuous real-valued functions are measurable. Clearly every Baire set is a Borel set $\mathbb{B}_a \subseteq \mathbb{B}$. Next we present a theorem which states that the two σ-algebras are equal in metric spaces, but that this does not hold true in more general spaces.

Theorem 2.4.9 (See [12] pg 223)**.** *In any metric space* (X, ρ)*, every Borel set is a Baire set. So that*

$$\mathbb{B}_a = \mathbb{B}\,.$$

Example 2.4.10. Let $(X, \mathbb{A})$ be a measurable space and $f : X \to \mathbb{R}$ be a constant function. Then f is measurable since, if $f(x) = c$ for all $x \in X$, $c \in \mathbb{R}$, and if $\alpha \geq c$, then

$$\{x \in X : f(x) > \alpha\} = \emptyset,$$

whereas if $\alpha < c$, then

$$\{x \in X : f(x) > \alpha\} = X.$$

Both $\emptyset$ and X are measurable sets in the measurable space $(X, \mathbb{A})$. Hence f is measurable.

Example 2.4.11. Let $(X, \mathbb{A})$ be a measurable space. Then for any subset $E \in \mathbb{A}$ the *characteristic function* 1_E, defined by

$$1_E(x) = \begin{cases} 1, \text{ if } x \in E, \\ 0, \text{ if } x \notin E, \end{cases}$$

is measurable. In fact $\{x \in X : 1_E(x) > \alpha\}$ is either X, E, or $\emptyset$.

Example 2.4.12. Let $\mathbb{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}$. Then any continuous function $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable. In fact if f is continuous then

$$\{x \in \mathbb{R} : f(x) > \alpha\} = f^{-1}(\alpha, \infty)$$

is an open set in $\mathbb{R}$ and hence it belongs to $\mathbb{B}$.

Example 2.4.13. Let $\mathbb{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}$. Then any monotone function $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable. For suppose that f is monotone increasing in the sense that $x \leq \tilde{x}$ implies $f(x) \leq f(\tilde{x})$. Then $\{x \in \mathbb{R} : f(x) > \alpha\}$ consists of a half-line which is either of the form $\{x \in \mathbb{R} : x > \alpha\} = (\alpha, \infty)$, $\{x \in \mathbb{R} : x \geq \alpha\} = [\alpha, \infty)$, or $\mathbb{R}$ or $\emptyset$.

Theorem 2.4.14 (See [4] pg 9)**.** *Let $(\mathbb{R}, \mathbb{B})$ be a measurable space. Let $f, g : \mathbb{R} \to \mathbb{R}$ be measurable functions and $c \in \mathbb{R}$. Then the functions*

$$cf,\ f^2,\ f + g,\ fg,\ |f|$$

are also measurable.

Definition 2.4.15. For any function $f : X \to \mathbb{R}$, we define f^+ and f^- to be the nonnegative functions on X to be

$$f^+(x) = \max\{f(x), 0\},$$
$$f^-(x) = \max\{-f(x), 0\}.$$

We call f^+ the *positive part* of f and f^- the *negative part* of f. From the definitions it follows that

$$f = f^+ - f^-,$$
$$|f| = f^+ + f^-.$$

In integration theory it is frequently convenient to adjoin the two symbols $-\infty$ and ∞ to the real number system $\mathbb{R}$. The motivation behind this is that it is convenient to say that the length of the real line is $+\infty$ and that we will frequently be taking the supremum of a set of real numbers. This last reason follows on from the fact that we know a non-empty set $A \subseteq \mathbb{R}$ which has an upper bound also has a supremum in $\mathbb{R}$. If we define the supremum of a non-empty set which does not have an upper bound to be $+\infty$, then every non-empty subset of $\mathbb{R}$ will have a unique supremum. Similarly, every non-empty subset of $\mathbb{R}$ will have a unique infimum.

Definition 2.4.16. The collection of $\overline{\mathbb{R}}$ consisting of the set

$$\mathbb{R} \cup \{-\infty, +\infty\}$$

is called the *extended real number system.*

We introduce the following algebraic operations between the symbols $+\infty$ and $-\infty$ and the elements of $\mathbb{R}$.

$$\begin{aligned} x + (+\infty) &= (+\infty) + x = +\infty\,, \\ x + (-\infty) &= (-\infty) + x = -\infty\,, \end{aligned}$$

hold for each $x \in \mathbb{R}$.

$$\begin{aligned} x(+\infty) &= (+\infty)x = +\infty\,, \\ x(-\infty) &= (-\infty)x = -\infty\,, \end{aligned}$$

hold for each positive $x \in \mathbb{R}$.

$$\begin{aligned} x(+\infty) &= (+\infty)x = -\infty\,, \\ x(-\infty) &= (-\infty)x = +\infty\,, \end{aligned}$$

hold for each negative $x \in \mathbb{R}$.

We also declare that

$$\begin{aligned}(+\infty)+(+\infty)&=+\infty\,,\\(-\infty)+(-\infty)&=-\infty\,,\\(+\infty)(+\infty)=(-\infty)(-\infty)&=+\infty\,,\\(+\infty)(-\infty)=(-\infty)(+\infty)&=-\infty\,,\end{aligned}$$

and

$$0(+\infty)=(+\infty)0=0(-\infty)=(-\infty)0=0\,.$$

The sums $(+\infty)+(-\infty)$ and $(-\infty)+(+\infty)$ are left undefined.

Definition 2.4.17. An extended real-valued function f on a measurable space $(X,\mathbb{A})$ is a function which is allowed to take on the values of $\{-\infty,\infty\}$. Therefore

$$f:X\to\overline{\mathbb{R}}=\mathbb{R}\cup\{-\infty,+\infty\}\,.$$

Definition 2.4.18. An extended real-valued function f on a measurable space $(X,\mathbb{A})$ is measurable if the set $\{x\in X:f(x)>\alpha\}$ is measurable for each $\alpha\in\mathbb{R}$. Notice that this is equivalent to $f^{-1}(\alpha,\infty]$ being measurable rather than $f^{-1}(\alpha,\infty)$ for the standard real number system.

The collection of all extended real-valued measurable functions on X is denoted by $\mathcal{M}(X,\mathbb{A})$ and we define

$$\mathcal{M}^{+}(X,\mathbb{A})=\{f\in\mathcal{M}(X,\mathbb{A}):f\geq 0\},$$

to be the collection of all non-negative measurable functions on $(X,\mathbb{A})$.

Lemma 2.4.19 (See [4] pg 12)**.** *Let $(X,\mathbb{A})$ be a measurable space. If (f_n) is a sequence in $\mathcal{M}(X,\mathbb{A})$ which converges to a function f on X, then f is measurable.*

Definition 2.4.20. Let $(X,\mathbb{A})$ be a measurable space. A *simple function* φ is a measurable function on X which is a finite linear combination of

characteristic functions of measurable sets belonging to $\mathbb{A}$. For $c_i \in \mathbb{R}$, $E_i \subseteq \mathbb{A}$, $i \in \{1, \ldots, n\}$ we represent φ by

$$\varphi = \sum_{i=1}^{n} c_i 1_{E_i}.$$

By itself the above definition allows for a simple function to have many representations as a linear combination of characteristic functions. Therefore among the representations for φ there is a unique *standard representation* with the following properties. The c_i are distinct and the E_i are disjoint subsets of X such that $X = \cup_{i=1}^{n} E_i$. This means that a simple function φ has the property that $\varphi(X) < \infty$ and that it can take on only a finite number of different values.

Next we show that in a certain sense the collection of simple functions on a measurable space $(X, \mathbb{A})$ is dense in the space of measurable functions $\mathcal{M}(X, \mathbb{A})$.

Lemma 2.4.21 (See [4] pg 13)**.** *Let $(X, \mathbb{A})$ be a measurable space. For $f \in \mathcal{M}^+(X, \mathbb{A})$, then there exists a sequence (φ_n) in $\mathcal{M}(X, \mathbb{A})$ such that*

1. *$0 \leq \varphi_n(x) \leq \varphi_{n+1}(x)$ for $x \in X$, $n \in \mathbb{N}$,*
2. *$f(x) = \lim_{n\to\infty} \varphi_n(x)$ for each $x \in X$,*
3. *Each φ_n has only a finite number of real values.*

The integration of measurable functions depends on introducing some notion of the *size* of a measurable set. We will do this by defining functions called *measures* which are suggested by our idea of length, area, mass, and so forth.

Definition 2.4.22. Let $(X, \mathbb{A})$ be a measurable space. A *measure* is an extended real-valued function μ defined on the σ-algebra $\mathbb{A}$ such that

(M1) $\mu(\emptyset) = 0$,

(M2) $\mu(E) \geq 0$ for all $E \in \mathbb{A}$,

(M3) μ is *countably additive* in the sense that if (E_n) is any disjoint sequence of sets in $\mathbb{A}$, then

$$\mu(\cup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mu(E_n).$$

Since a measure μ is an extended real-valued function it can take on the value $+\infty$. If μ does not take on $+\infty$ it is said to be *finite*. More generally, if there exists a sequence (E_n) of sets in $\mathbb{A}$ with $X = \cup_{i=1}^{\infty} E_n$ such that $\mu(E_n) < +\infty$ for all $n \in \mathbb{N}$, then μ is said to be σ-finite.

Example 2.4.23. Let $(X, \mathbb{A})$ be a measurable space. Choose and fix $p \in X$. Let μ be defined for $E \in \mathbb{A}$ by

$$\mu_p(E) = \begin{cases} 0 & \text{if } p \notin E, \\ 1 & \text{if } p \in E. \end{cases}$$

Then μ_p is a finite measure on X and is called the *unit measure concentrated at* p.

Example 2.4.24. Let $\mathbb{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}$. Then there exists a unique measure λ defined on $\mathbb{B}$ which coincides with the length on open intervals, see [4] pg 104. For the interval $E = (a, b) \subseteq \mathbb{R}$, we have

$$\lambda(E) = b - a.$$

This unique measure is called the *Lebesgue measure* and it is not a finite measure, but is σ-finite.

In the above definition measures have been described as a generalisation of our notions of length, area, and mass. These quantities are all nonnegative and hence a measure has been defined as nonnegative. To extend this notion of a measure further to say that of an electric charge which has real, possibly negative values, we introduce the concept of a *Signed measure*.

Definition 2.4.25. Let $(X, \mathbb{A})$ be a measurable space. A *signed measure* is an extended real-valued function μ defined on the σ-algebra $\mathbb{A}$ such that

(SM1) $\mu(\emptyset) = 0$,

(SM2) μ is *countably additive* in the sense that if (E_n) is any disjoint sequence of sets in $\mathbb{A}$, then

$$\mu(\cup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mu(E_n).$$

Definition 2.4.26. A *measure space* is an ordered triple $(X, \mathbb{A}, \mu)$ consisting of a set X, a σ-algebra $\mathbb{A}$ of subsets of X, and a measure μ defined on X.

Example 2.4.27. Let $(X, \mathbb{A}, \mu)$ be a measure space. If for all $E \in \mathbb{A}$ we have that $0 \leq \mu(E) \leq 1$ and $\mu(X) = 1$, then $(X, \mathbb{A}, \mu)$ is called a *probability space* and μ is called a *probability measure*.

Definition 2.4.28. Let $(X, \mathbb{A}, \mu)$ be a measure space. The collection $\mathbf{N}$ of sets $N \subseteq \mathbb{A}$ such that $\mu(N) = 0$ is important in that its members are usually ignored. Such a set N is called a *null set*.

Definition 2.4.29. Let $(X, \mathbb{A}, \mu)$ be a measure space and $P(x)$ be a proposition which is defined for all $x \in X$. Then if

$$\mu(\{x \in X : \neg P(x)\}) = 0\,,$$

we say that $P(x)$ holds *μ-almost everywhere*.

Example 2.4.30. Let $(X, \mathbb{A}, \mu)$ be a measure space. We say that two functions $f, g : X \to \mathbb{R}$ are *equal μ-almost everywhere*, if $f(x) = g(x)$ when $x \notin N$, for some $N \in \mathbb{A}$ with $\mu(N) = 0$.

Definition 2.4.31 (Convergence Almost Everywhere)**.** Let $(X, \mathbb{A}, \mu)$ be a measure space. We say that a sequence (f_n) in X *converges almost everywhere* to a function f if there exists a set $M \in \mathbb{A}$ with $\mu(M) = 0$ such that for every $\epsilon > 0$ and $x \in X \setminus M$ there exists a $n_0 = n_0(\epsilon, x) \in \mathbb{N}$, such that for all $n \geq n_0$, then

$$|f_n(x) - f(x)| < \epsilon.$$

In other words, $f_n \to f$ pointwise except on a null set.

Definition 2.4.32. Let $(X, \mathbb{A}, \mu)$ be a measure space. If φ is a simple function in $\mathcal{M}^+(X, \mathbb{A})$ with $E_i \in \mathbb{A}$ for all $i \in \{1, \ldots, n\}$ having representation

$$\varphi = \sum_{i=1}^{n} c_i 1_{E_i}.$$

We define the *integral* of φ with respect to μ to be the extended real number

$$\int \varphi \, d\mu = \sum_{i=1}^{n} c_i \, \mu(E_i).$$

It should be noted that the integral, $\int \varphi \, d\mu$, is independent of the choice of the c_i and E_i in the representation of the simple function φ.

Following directly from the definition of an integral of a simple function we have the following linearity properties.

Lemma 2.4.33 (See [4] pg 28)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space. If φ and ψ are simple functions in $\mathcal{M}^+(X, \mathbb{A})$ and $c \geq 0$, then*

$$\int c\varphi \, d\mu = c \int \varphi \, d\mu,$$
$$\int (\varphi + \psi) \, d\mu = \int \varphi \, d\mu + \int \psi \, d\mu.$$

Definition 2.4.34. Let $(X, \mathbb{A}, \mu)$ be a measure space. Let f be any function belonging to $\mathcal{M}^+(X, \mathbb{A})$. We define the *integral of f with respect to μ* to be the extended real number

$$\int f \, d\mu = \sup \int \varphi \, d\mu.$$

Where the supremum is extended over all simple functions $\varphi \in \mathcal{M}^+(X, \mathbb{A})$ satisfying $0 \leq \varphi(x) \leq f(x)$ for all $x \in X$. It should be noted that this definition is unique and does not depend on the choice of φ.

Furthermore, if $E \in \mathbb{A}$, then $f \cdot 1_E$ belongs to $\mathcal{M}^+(X, \mathbb{A})$ and we define the *integral of f over E with respect to μ* to be the extended real number

$$\int_E f \, d\mu = \int f \cdot 1_E \, d\mu.$$

Theorem 2.4.35 (Monotone Convergence Theorem. See [4] pg 31)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space. If (f_n) is a monotone increasing sequence of functions in $\mathcal{M}^+(X, \mathbb{A})$ which converges to f, then*

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

As a consequence of the Monotone Convergence Theorem we have

Corollary 2.4.36 (See [4] pg 32)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space.*

1. *If $f \in \mathcal{M}^+(X, \mathbb{A})$ and $c \geq 0$, then $cf \in \mathcal{M}^+(X, \mathbb{A})$ and*

$$\int cf \, d\mu = c \int f \, d\mu.$$

2. *If $f, g \in \mathcal{M}^+(X, \mathbb{A})$, then $f + g \in \mathcal{M}^+(X, \mathbb{A})$ and*

$$\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu.$$

Lemma 2.4.37 (Fatou's Lemma. See [4] pg 33)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space. If (f_n) belongs to $\mathcal{M}^+(X, \mathbb{A})$, then*

$$\int \left(\liminf_{n\to\infty} f_n \right) d\mu \leq \liminf_{n\to\infty} \int f_n \, d\mu.$$

Corollary 2.4.38 (See [4] pg 34)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space. For any $f \in \mathcal{M}^+(X, \mathbb{A})$, then $f(x) = 0$ μ-almost everywhere on X if and only if*

$$\int f \, d\mu = 0.$$

We now present a corollary which states that the Monotone Convergence Theorem holds if convergence on X is replaced by almost everywhere convergence.

Corollary 2.4.39 (See [4] pg 35)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space. If (f_n) is a monotone increasing sequence of functions in $\mathcal{M}^+(X, \mathbb{A})$ which converges μ-almost everywhere on X to a function $f \in \mathcal{M}^+(X, \mathbb{A})$, then*

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

Definition 2.4.40. Let $(X, \mathbb{A}, \mu)$ be a measure space. We define for all measurable functions $f \in \mathcal{M}(X, \mathbb{A})$ the *integral of f with respect to μ* to be

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu,$$

where f^+ and f^-, the positive and negative parts of f, have finite integrals.

If a set E belongs to $\mathbb{A}$, we define

$$\int_E f \, d\mu = \int_E f^+ \, d\mu - \int_E f^- \, d\mu.$$

We denote the collection of all measurable functions defined on X who are *integrable*, that is all real-valued measurable functions with positive and negative parts of f having finite integrals, by $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$.

Definition 2.4.41. Let $(X, \mathbb{A}, \mu)$ be a measure space. If (E_n) is a disjoint sequence in $\mathbb{A}$ with $E = \cup_{i=1}^{\infty} E_i$, then

$$\int_E f \, d\mu = \sum_{n=1}^{\infty} \int_{E_n} f \, d\mu.$$

This integral is called the *indefinite integral of f with respect to μ* and we say that the indefinite integral of a function in $\mathcal{L}$ is *countably additive.*

Corollary 2.4.42 (See [4] pg 43)**.** *Let $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$. For functions $f \in \mathcal{M}(X, \mathbb{A})$ and $g \in \mathcal{L}$ with $|f| \leq |g|$, then $f \in \mathcal{L}$ and*

$$\int |f| \, d\mu \leq \int |g| \, d\mu \, .$$

Theorem 2.4.43 (See [4] pg 43)**.** *Let $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$. For $f, g \in \mathcal{L}$, $\alpha \in \mathbb{R}$ we have that $\alpha f \in \mathcal{L}$ and $f + g \in \mathcal{L}$ and*

$$\int \alpha f \, d\mu = \alpha \int f \, d\mu,$$
$$\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu.$$

Next we state the most important convergence theorem for integrable functions on a measure space.

Theorem 2.4.44 (Lebesgue Dominated Convergence Theorem. See [4] pg 44)**.** *Let $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$ and (f_n) be a sequence of integrable functions which converges almost everywhere to a real-valued measurable function f. If there exists an integrable function g such that $|f_n| \leq g$ almost everywhere for all $n \in \mathbb{N}$, then f is integrable and*

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

2.5 The Lebesgue Spaces

Definition 2.5.1. Let $(X, \mathbb{A}, \mu)$ be a measure space. For $f \in \mathcal{L}(X, \mathbb{A}, \mu)$, we define

$$N_\mu(f) = \int |f| \, d\mu.$$

We now present a lemma which states that N_μ is a semi-norm on the space $\mathcal{L}(X, \mathbb{A}, \mu)$.

Lemma 2.5.2 (See [4] pg 54)**.** *Let $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$. For $f, g \in \mathcal{L}$, $\alpha \in \mathbb{R}$ and $x \in X$, $\mathcal{L}$ is a linear space under the operations defined by*

$$(f + g)(x) = f(x) + g(x),$$
$$(\alpha f)(x) = \alpha f(x),$$

and N_μ is a semi-norm on $\mathcal{L}$. Further $N_\mu(f) = 0$ if and only if $f(x) = 0$ μ-almost everywhere on X.

To transform $\mathcal{L}(X, \mathbb{A}, \mu)$ into a normed linear space, we use equivalence classes of functions instead of functions, with any two functions being in the same equivalence class if they are equal μ-almost everywhere.

Definition 2.5.3. Let $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$. For functions $f, g \in \mathcal{L}$ we say that they are *μ-equivalent* if they are equal μ-almost everywhere on X. The *equivalence class determined by f* in $\mathcal{L}$ is denoted by $[f]$, which contains the set of all functions in $\mathcal{L}$ which are μ-equivalent to f. The *Lebesgue space* $\mathbf{L} = \mathbf{L}(X, \mathbb{A}, \mu)$ consists of all μ-equivalence classes in $\mathcal{L}$. That is,

$$\mathbf{L}(X, \mathbb{A}, \mu) = \mathcal{L}(X, \mathbb{A}, \mu) \diagup \mathcal{N},$$

where $\mathcal{N} = \{f \in \mathcal{L}(X, \mathbb{A}, \mu) : f = 0\,,\ \mu\text{-almost everywhere on } X\}$.

For $[f] \in \mathbf{L}$, we define its norm by

$$\|[f]\| = \int |f|\ d\mu.$$

The above norm is well defined, since if $g \in [f]$ then $g = f$ μ-almost everywhere. This implies that $|g| = |f|$ μ-almost everywhere and by Corollary 2.4.38 we have that

$$\int |g|\ d\mu = \int |f|\ d\mu\,.$$

Theorem 2.5.4 (See [4] pg 54)**.** *Let* $\mathbf{L} = \mathbf{L}(X, \mathbb{A}, \mu)$*. Then for* $f, g \in \mathbf{L}$ *and* $\alpha \in \mathbb{R}$*,* $\mathbf{L}$ *is a normed linear space under the vector operations*

$$\alpha[f] = [\alpha f],$$
$$[f] + [g] = [f + g],$$

with norm defined by

$$\|[f]\| = \int |f|\ d\mu.$$

It must be remembered that the elements of $\mathbf{L}$ are actually equivalence classes of functions in $\mathcal{L}$. Though it is convenient to regard these elements as functions and we shall make reference to the equivalence class $[f]$ by referring to *the element* f *of* $\mathbf{L}$.

Having seen that the collection of integrable functions $\mathcal{L} = \mathcal{L}(X, \mathbb{A}, \mu)$ under the norm

$$\|[f]\| = \|f\| = \int |f|\ d\mu,$$

can be transformed into a normed linear space, by creating equivalence classes of functions and identifying any two integrable functions equivalent if they are equal μ-almost everywhere on X. We will now consider a collection of related normed linear spaces of equivalence classes of measurable functions.

Definition 2.5.5. Let $(X, \mathbb{A}, \mu)$ be a measure space and for $1 \leq p < \infty$, we define the space $\mathcal{L}_p = \mathcal{L}_p(X, \mathbb{A}, \mu)$ to be the collection of all measurable functions on X such that $|f|^p$ has finite integral with respect to μ over X. Further, for functions $f, g \in \mathcal{L}_p$ we say that they are μ-equivalent if they are equal μ-almost everywhere on X. The *Lebesgue space* $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$ consists of all μ- equivalence classes in $\mathcal{L}_p$. That is,

$$\mathbf{L}_p(X, \mathbb{A}, \mu) = \mathcal{L}_p(X, \mathbb{A}, \mu) \diagup \mathcal{N},$$

where $\mathcal{N} = \{f \in \mathcal{L}_p(X, \mathbb{A}, \mu) : f = 0\,, \text{ almost everywhere on } X\}$.

For $[f] \in \mathbf{L}_p$, we define its norm by

$$\|[f]\|_p = \left\{ \int |f|^p \, d\mu \right\}^{\frac{1}{p}}.$$

For $[f], [g] \in \mathbf{L}_p$ the associated metric on $\mathbf{L}_p(X, \mathbb{A}, \mu)$ is defined by

$$\rho_p([f], [g]) = \|f - g\|_p.$$

Theorem 2.5.6 (Riesz-Fischer Completeness Theorem. See [4] pg 59)**.** *For $1 \leq p < \infty$, let $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$. If $f, g \in \mathbf{L}_p$ and $\alpha \in \mathbb{R}$, the space $\mathbf{L}_p$ is a complete normed linear space under the vector operations*

$$\alpha[f] = [\alpha f],$$
$$[f] + [g] = [f + g],$$

and norm defined by

$$\|[f]\|_p = \left\{ \int |f|^p \, d\mu \right\}^{\frac{1}{p}}.$$

Definition 2.5.7. Let $(X, \mathbb{A}, \mu)$ be a measure space, we define the space $\mathcal{L}_\infty = \mathcal{L}_\infty(X, \mathbb{A}, \mu)$ to be the collection of all measurable functions on X which are μ-almost everywhere bounded. These functions are bounded outside a set of measure zero. Further, for functions $f, g \in \mathcal{L}_\infty$ we say that they

are μ-equivalent if they are equal μ-almost everywhere on X. The *Lebesgue space* $\mathbf{L}_\infty = \mathbf{L}_\infty(X, \mathbb{A}, \mu)$ consists of all μ-equivalence classes in $\mathcal{L}_\infty$. That is

$$\mathbf{L}_\infty(X, \mathbb{A}, \mu) = \mathcal{L}_\infty(X, \mathbb{A}, \mu) \diagup \mathcal{N},$$

where $\mathcal{N} = \{f \in \mathcal{L}_\infty(X, \mathbb{A}, \mu) : f = 0\,, \text{ almost everywhere on } X\}$.

If $f \in \mathbf{L}_\infty$ and $N \in \mathbb{A}$ with $\mu(N) = 0$, we define

$$S(N) = \sup\{|f(x)| : x \notin N\}$$

and we define its norm by

$$\|[f]\|_\infty = \inf\{S(N) : N \in \mathbb{A},\ \mu(N) = 0\}\,.$$

The elements of $\mathbf{L}_\infty$ are called the *essentially bounded functions.*

Example 2.5.8. Let $f : \mathbb{R} \to \mathbb{R}$ be the Lebesgue-measurable function in the measure space $(\mathbb{R}, \mathbb{B}, \lambda)$, defined by

$$f(t) = \begin{cases} t & \text{if } t \in \mathbb{N}\,, \\ 0 & \text{otherwise.} \end{cases}$$

The function f is unbounded on the set of Natural Numbers $\mathbb{N}$, which has Lebesgue measure of zero as it is countable. However outside the set $\mathbb{N}$ the function is bounded by zero. Hence the function f has an *essential upper bound* of 0.

Theorem 2.5.9 (Completeness Theorem for the Space $\mathbf{L}_\infty$. See [4] pg 61)**.** *Let $\mathbf{L}_\infty = \mathbf{L}_\infty(X, \mathbb{A}, \mu)$. If $f, g \in \mathbf{L}_\infty$ and $\alpha \in \mathbb{R}$, the space $\mathbf{L}_\infty$ is a complete normed linear space under the vector operations*

$$\alpha[f] = [\alpha f],$$
$$[f] + [g] = [f + g],$$

with

$$S(N) = \sup\{|f(x)| : x \notin N\}\,,$$

and norm defined by

$$\|[f]\|_\infty = \inf\{S(N) : N \in \mathbb{A},\ \mu(N) = 0\}.$$

In dealing with collections of measurable functions in a measure space $(X, \mathbb{A}, \mu)$ it is convenient to consider sequences of convergent measurable functions. It is known that uniform convergence implies point-wise convergence and that point-wise convergence implies μ-almost everywhere convergence and that in general the converse implications do not hold. Except however, if X consists of only a finite number of points, then point-wise convergence implies uniform convergence; if the only set of measure zero is the empty set, then μ-almost everywhere convergence implies point-wise convergence. We will consider convergence in the Lebesgue spaces $\mathbf{L}_p$ for $1 \leq p < \infty$.

Definition 2.5.10. A sequence (f_n) in $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$ *converges in* $\mathbf{L}_p$ to $f \in \mathbf{L}_p$, if for every $\epsilon > 0$ there exists a $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that for all $n \geq n_0$, then

$$\|f_n - f\|_p = \left\{\int |f_n - f|\ d\mu\right\}^{\frac{1}{p}} < \epsilon.$$

Theorem 2.5.11 (See [4] pg 67)**.** *Let* $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$*. Suppose the* $\mu(X) < +\infty$ *and* (f_n) *is a sequence in* $\mathbf{L}_p$ *which converges uniformly on* X *to* f*. Then* f *belongs to* $\mathbf{L}_p$ *and the sequence* (f_n) *converges in* $\mathbf{L}_p$ *to* f*.*

Theorem 2.5.12 (See [4] pg 67)**.** *Let* $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$*. Suppose* (f_n) *is a sequence in* $\mathbf{L}_p$ *which converges* μ*-almost everywhere to a measurable function* f*. If there exists a* $g \in \mathbf{L}_p$ *such that for all* $x \in X$ *and* $n \in \mathbb{N}$*,*

$$|f_n(x)| \leq g(x),$$

then $f \in \mathbf{L}_p$ *and the sequence* (f_n) *converges in* $\mathbf{L}_p$ *to* f*.*

Example 2.5.13. In the measure space $([0,1], \mathbb{B}, \lambda)$, consider the intervals of the form $[0,1]$, $\left[0, \frac{1}{2}\right]$, $\left[\frac{1}{2}, 1\right]$, $\left[0, \frac{1}{3}\right]$, $\left[\frac{1}{3}, \frac{2}{3}\right]$, $\left[\frac{2}{3}, 1\right]$, $\left[0, \frac{1}{4}\right]$, $\left[\frac{1}{4}, \frac{1}{2}\right]$, $\left[\frac{1}{2}, \frac{3}{4}\right]$, $\left[\frac{3}{4}, 1\right]$, $\left[0, \frac{1}{5}\right]$, ...

Let f_n be the characteristic function on the nth interval on the above list and let $f(x) = 0$ for all $x \in [0,1]$. If $n \geq \frac{m(m+1)}{2} = (1+2+\cdots+m)$, then f_n is a characteristic function of an interval whose measure is at most $\frac{1}{m}$. Hence

$$\begin{aligned}(\|f_n - f\|_p)^p &= \int |f_n - f|^p \, d\lambda \\ &= \int f_n \, d\lambda \\ &\leq \frac{1}{m}.\end{aligned}$$

Therefore (f_n) converges in $\mathbf{L}_p$ to f. Now if we choose any point $x^* \in [0,1]$, then the sequence $(f_n(x^*))$ has a subsequence consisting only of ones and another subsequence consisting only of zeros. Therefore, the sequence (f_n) *does not converge at any point of* $[0,1]$.

Although convergence in $\mathbf{L}_p$ does not imply convergence μ-almost everywhere, we take note that convergence in $\mathbf{L}_p$ is related to another type of convergence that is of interest.

Definition 2.5.14. Let $(X, \mathbb{A}, \mu)$ be a measure space. A sequence (f_n) of measurable real-valued functions is said to *converge in measure* to a measurable real-valued function f, if for all $\alpha > 0$,

$$\lim_{n\to\infty} \mu\left(\{x \in X : |f_n(x) - f(x)| \geq \alpha\}\right) = 0.$$

Theorem 2.5.15 (See [4] pg 71)**.** *Let $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$. Let (f_n) be a sequence in $\mathbf{L}_p$ which converges in measure to f and let $g \in \mathbf{L}_p$ be such that*

$$|f_n(x)| \leq g(x),$$

μ-almost everywhere. Then $f \in \mathbf{L}_p$ and the sequence (f_n) converges in $\mathbf{L}_p$ to f.

Theorem 2.5.16 (Tchebyshev's Inequality. See [8] pg 67)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space, and let $f : X \to [0, +\infty]$ be an extended real-valued measurable function on X. If $t > 0$ and for $A_t = \{x \in X : f(t) \geq t\}$, then*

$$\mu\left(\{x \in X : f(t) \geq t\}\right) \leq \frac{1}{t}\int_{A_t} f \, d\mu \leq \frac{1}{t}\int f \, d\mu\,.$$

Proof. For $t > 0$ and $f : X \to [0, +\infty]$ we have the following relation

$$0 \leq t\, 1_{A_t} \leq f\, 1_{A_t} \leq f\,,$$

and by Corollary 2.4.42 this implies

$$\int t\, 1_{A_t}\, d\mu \leq \int f\, 1_{A_t}\, d\mu \leq \int_{A_t} f\, d\mu \leq \int f\, d\mu\,.$$

Now

$$\int t\, 1_{A_t}\, d\mu = t\, \mu(A_t)$$

and this implies

$$t\, \mu(A_t) \leq \int f\, 1_{A_t}\, d\mu \leq \int f\, d\mu\,.$$

Therefore

$$\mu(A_t) \leq \frac{1}{t} \int_{A_t} f\, d\mu \leq \frac{1}{t} \int f\, d\mu\,.$$

□

2.6 Density Theorems for Measurable Functions

In extending the approximation capabilities of the set of feedforward artificial neural networks $\mathcal{N}_\sigma^n$, and $\mathcal{A}_\sigma^n$, to measurable functions we again make use of the concept of density and the properties of measures on locally compact spaces.

Definition 2.6.1. A metric space (X, ρ) is said to be *locally compact* if for every point $x \in X$, there is an open neighbourhood N_x of x which has compact closure.

Example 2.6.2. The Euclidean spaces $\mathbb{R}^n$ are locally compact Hausdorff spaces.

Certain measures have the property that every measurable set, in the σ-algebra on which they are defined, is *approximately open* and *approximately closed*. Measures with this property are called *Regular measures*.

Definition 2.6.3. Let (X, ρ) be a metric space with $\mathbb{A}$ being a σ-algebra on X. A measure μ on $\mathbb{A}$ is said to be *regular* if

(R1) each compact subset $K \subseteq X$ satisfies $\mu(K) < +\infty$,

(R2) each set $A \in \mathbb{A}$ satisfies

$$\mu(A) = \inf \{\mu(U) : A \subseteq U \text{ and } U \text{ is open}\},$$

(R3) each open subset $U \subseteq X$ satisfies

$$\mu(U) = \sup \{\mu(K) : K \subseteq U \text{ and } K \text{ is compact}\}.$$

Theorem 2.6.4 (See [8] pg 26)**.** *The Lebesgue measure λ on the measurable space $(\mathbb{R}^n, \mathbb{B})$ is regular.*

Definition 2.6.5. Let $\mathbb{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}$. A measure μ on $\mathbb{A}$ is said to be a *Borel measure* if its domain is $\mathbb{B}$.

Theorem 2.6.6 (See [8] pg 40)**.** *Let μ be a finite measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. Then μ is regular.*

Definition 2.6.7. Let (X, ρ) be a metric space and $f : X \to \mathbb{R}$ a continuous function on X. The *support* of f, denoted by $supp(f)$, is the closure of

$$\{x \in X : f(x) \neq 0\}.$$

If X is a locally compact Hausdorff space (metric spaces are Hausdorff) we denote by $\mathcal{K}(X)$ the set of continuous functions $f : X \to \mathbb{R}$ for which $supp(f)$ is compact.

Example 2.6.8. If $X = \mathbb{R}$ the functions f on X with compact support are the bounded functions and therefore vanish at infinity.

Definition 2.6.9 (Support of a Measure)**.** Let (X, ρ) be a metric space and $\mathbb{B}$ be the Borel σ-algebra on X with measure μ defined on $\mathbb{B}$. Then the *support of the measure* μ is defined to be the set of all points $x \in X$ for which every open neighbourhood N_x of x has positive measure

$$supp(\mu) = \{x \in X : x \in N_x \Rightarrow \mu(N_x) > 0\}.$$

Example 2.6.10. Let $\mathbb{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}$ with the Lebesgue measure λ. Then for an arbitrary point $x \in \mathbb{R}$, any open neighbourhood N_x of x will contain some open interval

$$(x - \epsilon, x + \epsilon)\,,$$

for some $\epsilon > 0$. This interval has Lebesgue measure $2\epsilon > 0$ and so $\lambda(N_x) \geq 2\epsilon > 0$. Since $x \in \mathbb{R}$ was arbitrary, the $supp(\lambda) = \mathbb{R}$.

Example 2.6.11. Let $\mathbb{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}$ with the unit point measure μ_p concentrated at $p \in \mathbb{R}$ and pick any $x \in \mathbb{R}$. If $x = p$, then every open neighbourhood N_x of x will contain p and so $\mu_p(N_x) = 1 > 0$. On the other hand if $x \neq p$, then there exists a sufficiently small open ball $B(x, \epsilon)$ that does not contain p and so $\mu_p(B(x, \epsilon)) = 0$. We conclude that the $supp(\mu_p)$ is the closure of the singleton set $\{p\}$, which is $\{p\}$ itself.

Theorem 2.6.12 (See [8] pg 108)**.** *Let $\mathcal{L}_p = \mathcal{L}_p(X, \mathbb{A}, \mu)$ with $1 \leq p \leq +\infty$. Then the simple functions in $\mathcal{L}_p$ form a dense subspace of $\mathcal{L}_p$ and so determine a dense subspace of $\mathbf{L}_p$.*

Theorem 2.6.13 (See [8] pg 109)**.** *Let $[a, b] \subseteq \mathbb{R}$ be a closed and bounded interval in the measure space $([a, b], \mathbb{A}, \mu)$ and let $1 \leq p < +\infty$. Then the subspace of $\mathbf{L}_p([a, b], \mathbb{A}, \mu)$ determined by the continuous functions on $[a, b]$ is dense in $\mathbf{L}_p([a, b], \mathbb{A}, \mu)$.*

Theorem 2.6.14 (See [8] pg 227)**.** *Let X be a locally compact Hausdorff space, $\mathbb{A}$ be a σ-algebra on X that includes the Borel σ-algebra $\mathbb{B}$ of X and let μ be a regular measure on the measurable space $(X, \mathbb{A})$. For $1 \leq p < +\infty$, the set of all continuous functions on X with compact support $\mathcal{K}(X)$, is dense in the space of $\mathcal{L}_p(X, \mathbb{A}, \mu)$ and so determines a dense subspace of $\mathbf{L}_p(X, \mathbb{A}, \mu)$.*

Theorem 2.6.15 (See [2] pg 90)**.** *Let $\mathcal{L}_\infty = \mathcal{L}_\infty(X, \mathbb{A}, \mu)$ and let $f \in \mathcal{L}_\infty$. For all $\epsilon > 0$, there is a simple function $\varphi \in \mathcal{L}_\infty$ such that*

$$\|f - \varphi\|_\infty < \epsilon\,.$$

Thus the simple functions are dense in $\mathcal{L}_\infty$ and so determines a dense subspace of $\mathbf{L}_\infty(X, \mathbb{A}, \mu)$.

Theorem 2.6.16 (Lusin's Theorem. See [8] pg 227)**.** *Let X be a locally compact Hausdorff space, $\mathbb{A}$ be a σ-algebra on X that includes the Borel σ-algebra $\mathbb{B}$ of X, μ be a regular measure on the measurable space $(X, \mathbb{A})$ and let $f : X \to \mathbb{R}$ be a measurable function. If $A \in \mathbb{A}$ and satisfies $\mu(A) < +\infty$ and if $\epsilon > 0$, then there is a compact subset $K \subseteq A$ such that $\mu(A \setminus K) < \epsilon$ and such that the restriction of f to K is continuous. Moreover, there is a function $g \in \mathcal{K}(X)$ that agrees with f at each point in K. If $A \neq \emptyset$, then the function g can be chosen so that*

$$\sup\{|g(x)| : x \in X\} \leq \sup\{|f(x)| : x \in A\}.$$

Theorem 2.6.17 (See [15] pg 241)**.** *Let $(X, \mathbb{B}_a, \mu)$ be the measure space with Baire σ-algebra $\mathbb{B}_a$ and Baire measure μ defined on $\mathbb{B}_a$. Then for any $\epsilon > 0$ and any integrable simple Baire function φ, there exists an integrable simple function*

$$\phi = \sum_{i=1}^{n} c_i \, 1_{E_i},$$

such that every E_i is a compact Baire set and

$$\int |\varphi - \phi| \, d\mu \leq \epsilon.$$

Theorem 2.6.18 (See [15] pg 242)**.** *Let $(X, \mathbb{B}_a, \mu)$ be the measure space with Baire σ-algebra $\mathbb{B}_a$ and Baire measure μ defined on $\mathbb{B}_a$. Then for any $\epsilon > 0$ and simple function*

$$\phi = \sum_{i=1}^{n} c_i \, 1_{E_i},$$

such that every E_i is a compact Baire set. Then there exists a function with compact support $h \in \mathcal{K}(X)$ such that

$$\int |\phi - h| \, d\mu \leq \epsilon.$$

In our endeavour to approximate measurable functions by multilayer feed-forward artificial neural networks, we are only concerned with distinguishing between classes of μ-equivalent functions [20]. We now present the definitions necessary to make sense of the relevant metric used to achieve this.

In the same way that we transformed the space of all integrable functions $\mathcal{L}(X, \mathbb{A}, \mu)$ into the Lebesgue Space $\mathbf{L}(X, \mathbb{A}, \mu)$ of equivalence classes of integrable functions, see Definition 2.5.3, we transform the space of all measurable functions $\mathcal{M}$.

Note that the requirement for μ to be a probability measure is merely one of convenience born out of practical considerations. The following concepts and ideas will hold for any finite measure μ [20].

Definition 2.6.19. Let μ be a probability measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. For functions $f, g \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$, we say they are μ-equivalent if

$$\mu\{x \in \mathbb{R}^n : f(x) = g(x)\} = 1\,.$$

The *equivalence class determined by* f in $\mathcal{M}$ is denoted by $[f]$, which contains the set of all functions in $\mathcal{M}$ which are μ-equivalent to f.

The space $\mathbf{M} = \mathbf{M}(\mathbb{R}^n, \mathbb{B}, \mu)$ consists of all μ-equivalence classes in $\mathcal{M}$. That is

$$\mathbf{M}(\mathbb{R}^n, \mathbb{B}, \mu) = \mathcal{M}(\mathbb{R}^n, \mathbb{B}) \diagup \mathcal{N},$$

where $\mathcal{N} = \{f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B}) : f = 0\,,\ \mu\text{-almost everywhere on } \mathbb{R}^n\}$.

Example 2.6.20. Let μ be a probability measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. For $f, g \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ we define

$$\rho_\mu : \mathcal{M}(\mathbb{R}^n, \mathbb{B}) \times \mathcal{M}(\mathbb{R}^n, \mathbb{B}) \to \mathbb{R}^+$$

by

$$\rho_\mu(f, g) = \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} < \epsilon\}\,.$$

Lemma 2.6.21. *Let μ be a probability measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. Then ρ_μ, see Example 2.6.20, is a metric on $\mathbf{M}(\mathbb{R}^n, \mathbb{B})$.*

Proof. For any $f, g, h \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$, the infimum exists as the set

$$\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} < \epsilon\}$$

is a subset of $\mathbb{R}$ which is bounded below by 0. Next we must show that ρ_μ satisfies the requirements in Definition 2.1.1

(M1) $\rho_\mu(f,g) \geq 0$, since the infimum of non-negative numbers is non-negative.

(M2)

$$\begin{aligned}
\rho_\mu(f,g) = 0 &\iff \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} < \epsilon\} = 0 \\
&\iff \forall \epsilon > 0,\ \mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} < \epsilon \\
&\iff |f(x) - g(x)| < \epsilon,\ \mu\text{-almost everywhere} \\
&\iff f = g,\ \mu\text{-almost everywhere} \\
&\iff [f] = [g]\,.
\end{aligned}$$

(M3)

$$\begin{aligned}
\rho_\mu(f,g) &= \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} < \epsilon\} \\
&= \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |-1|\,|g(x) - f(x)| > \epsilon\} < \epsilon\} \\
&= \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |g(x) - f(x)| > \epsilon\} < \epsilon\} \\
&= \rho_\mu(g,f)\,.
\end{aligned}$$

(M4) Let

$$\begin{aligned}
\rho_\mu(f,h) &= \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |f(x) - h(x)| > \epsilon\} < \epsilon\} \\
&= \epsilon^*\,.
\end{aligned}$$

Then ϵ^* is the greatest lower bound for $|f(x) - h(x)|$. Now by the triangle inequality

$$|f(x) - h(x)| \leq |f(x) - g(x)| + |g(x) - h(x)|$$

and it follows then that

$$\epsilon^* < |f(x) - g(x)| + |g(x) - h(x)|\,.$$

Therefore ϵ^* is a lower bound for $|f(x) - g(x)| + |g(x) - h(x)|$.

If

$$\begin{aligned}\rho_\mu(f,g) &= \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} < \epsilon\} \\ &= \widetilde{\epsilon}\,,\end{aligned}$$

and

$$\begin{aligned}\rho_\mu(g,h) &= \inf\{\epsilon > 0 : \mu\{x \in \mathbb{R}^n : |g(x) - h(x)| > \epsilon\} < \epsilon\} \\ &= \overline{\epsilon}\,,\end{aligned}$$

then since $\widetilde{\epsilon}+\overline{\epsilon}$ is the greatest lower bound for $|f(x) - g(x)|+|g(x) - h(x)|$, we have $\epsilon^* < \widetilde{\epsilon} + \overline{\epsilon}$. Hence

$$\rho_\mu(f,h) \leq \rho_\mu(f,g) + \rho_\mu(g,h)\,.$$

□

We take notice that two functions are close in this metric if and only if there is small probability that they differ significantly.

Lemma 2.6.22. *Let μ be a measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. For $f, g \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$, if f is μ-equivalent to g then $\rho_\mu(f,g) = 0$.*

Proof. Since f is μ-equivalent to g, we know that

$$\mu\{x \in \mathbb{R}^n : f(x) = g(x)\} = 1.$$

Which implies

$$\mu\{x \in \mathbb{R}^n : |f(x) - g(x)| = 0\} = 1.$$

Therefore for all $\epsilon > 0$,

$$\mu\{x \in \mathbb{R}^n : |f(x) - g(x)| > \epsilon\} = 0.$$

Hence $\rho_\mu(f,g) = 0$. □

Lemma 2.6.23. *Let μ be a probability measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. For a sequence (f_n) in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ and $f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$, the following are equivalent*

1. *$\rho_\mu(f_n, f) \to 0$,*
2. *For every $\epsilon > 0$, $\mu\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\} \to 0$,*
3. *$\int \min\{|f_n(x) - f(x)|, 1\}\, d\mu \to 0$.*

Proof.

(1) $\iff$ (2) :

Suppose $\rho_\mu(f_n, f) \to 0$. Then for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$,

$$|\rho_\mu(f_n, f) - 0| < \epsilon.$$

This implies that for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$,

$$\inf\{\tilde{\epsilon} > 0 : \mu\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \tilde{\epsilon}\} < \tilde{\epsilon}\} < \epsilon.$$

So the greatest lower bound such that

$$\mu\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \tilde{\epsilon}\} < \tilde{\epsilon}$$

whenever $n \geq n_0$, is less than ϵ.

Therefore for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$,

$$\mu\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\} < \epsilon.$$

Hence for all $\epsilon > 0$,

$$\mu\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\} \to 0.$$

(2) $\implies$ (3) :

Suppose that for every $\epsilon > 0$,

$$\mu\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\} \to 0.$$

This implies that for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$,

$$\mu\left\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\right\} < \epsilon.$$

Then choose $\frac{\epsilon}{2} > 0$ and $n_0 = n_0(\frac{\epsilon}{2}) \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\mu\left\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \frac{\epsilon}{2}\right\} < \frac{\epsilon}{2},$$

and define $E_{\frac{\epsilon}{2}} \subseteq \mathbb{R}^n$ to be

$$E_n = \left\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \frac{\epsilon}{2}\right\}.$$

This implies that

$$\begin{aligned}\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu &= \int_{\mathbb{R}^n \setminus E_{\frac{\epsilon}{2}}} \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \\ &\quad + \int_{E_{\frac{\epsilon}{2}}} \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,.\end{aligned}$$

Hence

$$\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \to 0.$$

$(3) \Longrightarrow (2)$:

Assume that

$$\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \to 0\,.$$

Then for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$,

$$\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu < \epsilon\,.$$

By Theorem 2.5.16,

$$\mu\left\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\right\} \leq \frac{1}{\epsilon}\int |f_n - f|\, d\mu\,.$$

Therefore, by assumption we have that

$$\mu\left\{x \in \mathbb{R}^n : |f_n(x) - f(x)| > \epsilon\right\} < \epsilon\,.$$

□

Knowing that ρ_μ-convergence is equivalent to convergence in measure (probability), see Lemma 2.6.23, we now show how uniform convergence on compacta is related to ρ_μ-convergence.

Lemma 2.6.24. *Let μ be a probability measure on $\mathbb{B}$, the Borel σ-algebra of subsets of $\mathbb{R}^n$. If a sequence (f_n) in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ converges uniformly on compacta to $f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$, then*

$$\rho_\mu(f_n, f) \to 0.$$

Proof. By Lemma 2.6.23, we have that

$$\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \to 0$$

implies

$$\rho_\mu(f_n, f) \to 0.$$

Therefore for any chosen $\epsilon > 0$ it is sufficient to find an $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$, we have

$$\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu < \epsilon.$$

Without loss of generality and since μ is a probability measure on the Borel σ-algebra $\mathbb{B}$ of $\mathbb{R}^n$, we suppose that

$$\mu(\mathbb{R}^n) = 1.$$

Since $\mathbb{R}^n$ is a locally compact metric space and μ is a finite measure (probability measures are finite) we have by Theorem 2.6.6 that μ is a regular measure. Seeing that μ is a finite measure by Theorem 2.6.16, Lusin's Theorem, there exists a compact subset $K \subseteq \mathbb{R}^n$ such that

$$\mu(\mathbb{R}^n \setminus K) < \frac{\epsilon}{2}.$$

Which is equivalent to saying that

$$\mu(K) > 1 - \frac{\epsilon}{2}.$$

By our assumption the sequence (f_n) converges uniformly on compacta to f. Therefore choose an $n_0 = n_0(\epsilon) \in \mathbb{N}$, such that for all $n \geq n_0$,

$$\begin{aligned} \|f_n - f\|_K &= \sup_{x \in K} |f_n(x) - f(x)| \\ &< \frac{\epsilon}{2}. \end{aligned}$$

Now for any $\epsilon > 0$ and for all $n \geq n_0$, we have that

$$\begin{aligned} \int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu &= \int_{\mathbb{R}^n \setminus K} \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \\ &\quad + \int_K \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,. \end{aligned}$$

Therefore

$$\int \min\left\{|f_n(x) - f(x)|, 1\right\} d\mu \to 0\,,$$

and by Lemma 2.6.23 it follows that

$$\rho_\mu(f_n, f) \to 0.$$

□

2.7 Linear Functionals

In dealing with feedforward artificial neural networks we find ourselves in the setting of normed linear spaces. The mappings from such normed linear spaces into themselves or other normed linear spaces are of great interest, especially those that are themselves both linear and continuous. Here we define the simplest of such mappings, the *linear functionals*.

Definition 2.7.1. When X and Y are sets, the symbol

$$f : X \to Y$$

will mean that f is a *mapping* of X into Y. If $A \subseteq X$ and $B \subseteq Y$, the *image* $f(A)$ of A and the *inverse image* $f^{-1}(B)$ of B are defined by

$$f(A) = \{f(x) \in B : x \in A\} \ ,$$
$$f^{-1}(B) = \{x \in A : f(x) \in B\} \ .$$

Suppose now that the sets X and Y are linear spaces over the same scalar field F. A mapping $T : X \to Y$ is said to be *linear* if for all $x, y \in X$ and $\alpha, \beta \in F$

$$T(\alpha x + \beta y) = \alpha Tx + \beta Ty \, .$$

Definition 2.7.2. Let X be a normed linear space with norm $\| \cdot \| : X \to \mathbb{R}$ and scalar field F. A *linear functional* f on the normed linear space X is a linear map from X into the associated scalar field F.

Definition 2.7.3. Let X and Y be normed linear spaces with associated norms $\| \cdot \|_X$ and $\| \cdot \|_Y$. A linear map $T : X \to Y$ is *continuous at* $a \in X$ if for all $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that

$$\|Tx - Ta\|_Y < \epsilon \text{ whenever } \|x - a\|_X < \delta \, .$$

The linear map T is said to be *continuous* if it is continuous at each point in X.

Definition 2.7.4. Let X and Y be normed linear spaces with associated norms $\| \cdot \|_X$ and $\| \cdot \|_Y$. We say that a linear map $T : X \to Y$ is *bounded* if there is a $k \in \mathbb{R}$ such that

$$\|Tx\|_Y \leq k\|x\|_X,$$

for all $x \in X$ where $k \in \mathbb{R}$ is a *bound* for the linear map T.

Theorem 2.7.5 (See [30] pg 100)**.** *Let X and Y be normed linear spaces with associated norms $\| \cdot \|_X$ and $\| \cdot \|_Y$. A linear map $T : X \to Y$ is continuous if and only if it is bounded.*

Definition 2.7.6. Let X and Y be normed linear spaces with associated norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. We define the *norm* of a bounded linear map $T : X \to Y$ to be

$$\|T\| = \sup\{\|T(x)\|_Y : \|x\|_X = 1\}.$$

Definition 2.7.7. Let X be a normed linear space with norm $\|\cdot\|$ and associated scalar field F. The *dual space* of the normed linear space X, denoted by X^*, is the space of all bounded, and therefore continuous, linear functionals on X.

Theorem 2.7.8 (See [30] pg 110)**.** *Let X be a normed linear space with norm $\|\cdot\|$ and associated scalar field F. Then the dual space X^* is a Banach space, with dual norm defined for $f, g \in X^*$ and $\lambda \in F$ by*

$$\|f\| = \sup\{|f(x)| : \|x\| = 1\},$$

and linear operations defined for $x \in X$ by

$$(f+g)(x) = f(x) + g(x),$$
$$(\lambda f)(x) = \lambda(f(x)).$$

Next we will state the *Hahn-Banach theorems*. The Extension Theorem which states that if f is a bounded linear functional defined on a subspace M of a normed linear space X we can extend f to all of the normed linear space X without changing the norm of f. As a direct consequence of the Extension Theorem, the theorems which state that in a normed linear space X there are enough bounded linear functionals to distinguish points.

Theorem 2.7.9 (The Hahn-Banach Extension Theorem. See [30] pg 116)**.** *Let X be a normed linear space with norm $\|\cdot\|$ and associated scalar field F. Also let M be a linear subspace of X. Now if f_0 is a bounded linear functional defined on M, then f_0 has an extension to a bounded linear functional f on X such that*

$$\|f\| = \|f_0\| = \sup\{|f_0(x)| : x \in M, \|x\| \leq 1\}.$$

Theorem 2.7.10 (See [2] pg 141)**.** *Let X be a normed linear space with norm $\|\cdot\|$ and associated scalar field F. Also let M be a linear subspace of X and X^* be the dual space of X.*

1. *If $x_0 \notin \overline{M}$, then there is an $f \in X^*$ such that $f = 0$ on M , $f(x_0) = 1$, and $\|f\| = \frac{1}{d}$ where d is the distance from x_0 to M,*

2. *$x_0 \in \overline{M}$ if and only if for every $f \in X^*$ that vanishes on M also vanishes at x_0,*

3. *If $x_0 \neq 0$, then there is an $f \in X^*$ such that $\|f\| = 1$ amd $f(x_0) = \|x_0\|$ Thus the maximum value of*

$$\frac{|f(x)|}{\|x\|},$$

for $x \neq 0$ is achieved at x_0. In particular, if $x \neq y$ then there is an $f \in X^*$ such that $f(x) \neq f(y)$.

We now define the concept of an *annihilator* which is derived from the geometric property of perpendicularity between vectors in inner product spaces.

Definition 2.7.11. Let X be a normed linear space and X^* the corresponding dual space. Given any $x \in X$ and $f \in X^*$ we say that x *annihilates* f, denoted by $x \perp f$, if $f(x) = 0$.

Further for any $x \in X$ we say that x annihilates a subset $B \subseteq X^*$, denoted by $x \perp B$, if $x \perp f$ for all $f \in B$. Conversely for any $f \in X^*$ we say that a subset $A \subseteq X$ annihilates f, denoted by $A \perp f$, if $x \perp f$ for all $x \in A$.

Definition 2.7.12. Let X be a normed linear space and X^* the corresponding dual space. The *annihilator* of $B \subseteq X^*$ is the subset of X defined by

$$B^{\perp} = \{x \in X : x \perp B\}.$$

Similarly, the annihilator of $A \subseteq X$ is the subset of X^* defined by

$$A^{\perp} = \{f \in X^* : A \perp f\}.$$

In the work of Cybenko [11], the Hahn-Banach Theorem and the Riesz Representation Theorem in conjunction with the notion of a measure being annihilated are used to prove that feedforward artificial neural networks are dense in the space of continuous functions. In order to do this we will need the concept of linear functionals on the $\mathbf{L}_p$ and $\mathcal{C}^n$ spaces, the Fourier Transform of a measure, and we state the Riesz Representation theorems.

Definition 2.7.13. A *linear functional* on $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$ is a mapping G of $\mathbf{L}_p$ into $\mathbb{R}$ such that for all $\alpha, \beta \in \mathbb{R}$ and $f, g \in \mathbf{L}_p$

$$G(\alpha f + \beta g) = \alpha G(f) + \beta G(g)\,.$$

The linear functional G is *bounded* if there exists a $k \in \mathbb{R}$ such that for all $f \in \mathbf{L}_p$

$$|G(f)| \leq k\|f\|_p\,.$$

Further, the *norm* of G is defined to be

$$\|G\| = \sup\left\{|G(f)| : f \in \mathbf{L}_p\,, \|f\|_p = 1\right\}\,.$$

It is a consequence of the linearity of the integral and Holder's Inequality that if $g \in \mathbf{L}_q$, where $q = \infty$ when $p = 1$ and $q = \frac{p}{p-1}$ otherwise, and we define G on $\mathbf{L}_p$ by

$$G(f) = \int f g \, d\mu, \tag{2.3}$$

then G is a linear functional with norm at most equal to $\|g\|_q$. See [4] pg 89. The Riesz Representation theorems yield a converse to this observation.

Theorem 2.7.14 (The Riesz Representation Theorem on $\mathbf{L}_1$. See [4] pg 90)**.** *Let $(X, \mathbb{A}, \mu)$ be a σ-finite measure space and G a bounded linear functional on $\mathbf{L}_1 = \mathbf{L}_1(X, \mathbb{A}, \mu)$. Then there exists a $g \in \mathbf{L}_\infty = \mathbf{L}_\infty(X, \mathbb{A}, \mu)$ such that for all $f \in \mathbf{L}_1$,*

$$G(f) = \int f g \, d\mu\,.$$

Moreover, $\|G\| = \|g\|_\infty$ and $g \geq 0$ if G is a positive linear functional.

Theorem 2.7.15 (The Riesz Representation Theorem on $\mathbf{L}_p$. See [4] pg 91)**.** *Let $(X, \mathbb{A}, \mu)$ be a measure space and G a bounded linear functional on $\mathbf{L}_p = \mathbf{L}_p(X, \mathbb{A}, \mu)$ for $1 < p < \infty$. Then there exists a $g \in \mathbf{L}_q = \mathbf{L}_q(X, \mathbb{A}, \mu)$ where $q = \frac{p}{p-1}$, such that for all $f \in \mathbf{L}_p$,*

$$G(f) = \int f g \, d\mu \, .$$

Moreover, $\|G\| = \|g\|_q$

Theorem 2.7.16 (The Riesz Representation Theorem on $\mathcal{C}$. See [4] pg 106)**.** *Let K be a compact subset of $\mathbb{R}$ and if G is a bounded linear functional on $\mathcal{C}_K$, then there exists a measure μ defined on the Borel subsets $\mathbb{B}$ of $\mathbb{R}$ such that for all $f \in \mathcal{C}_K$,*

$$G(f) = \int_K f \, d\mu \, .$$

Moreover, the norm $\|G\|$ of G equals $\mu(K)$.

Theorem 2.7.17 (The Riesz Representation Theorem on $\mathcal{C}_K$. See [2] pg 184)**.** *Let K be a compact Hausdorff space and L a bounded linear functional on $\mathcal{C}_K$.*

1. *Then there is a unique finite signed measure μ on the σ-algebra of K such that for all $f \in \mathcal{C}_K$, we have*

   $$L(f) = \int_K f \, d\mu \, .$$

   Moreover, the norm $\|L\|$ of L equals $\mu(K)$.

2. *Then there is a unique regular finite signed measure λ on the Borel subsets $\mathbb{B}$ of K such that for all $f \in \mathcal{C}_K$, we have*

   $$L(f) = \int_K f \, d\lambda \, .$$

   Moreover, the norm $\|L\|$ of L equals $\lambda(K)$.

Definition 2.7.18 (See [12])**.** Let $(\mathbb{R}^n, \mathbb{B}, \mu)$ be the Borel measurable space on $\mathbb{R}^n$ with finite measure μ. The *Fourier Transform* of the measure μ with $t, x \in \mathbb{R}^n$ is

$$\hat{\mu}(t) = \int_{\mathbb{R}^n} \exp(i\, t \cdot x)\, d\mu\,.$$

Theorem 2.7.19 (See [34] pg 176)**.** *The Fourier Transform is a continuous, linear, one-to-one mapping.*

Chapter 3

Various Approximation Results

In this chapter we will present the different methods used by Hornik et al. [18, 19, 20] and Cybenko [11] to prove the various approximation results for the set of multilayer feedforward artificial neural networks $\mathcal{N}_\sigma^n$.

3.1 The Method of Stone-Weierstrass for Continuous Functions

In this section we will show using the ideas of Hornik et al. [20], how a multilayer feedforward artificial neural network is capable of approximating a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ under certain conditions.

We defined, see Definition 1.4.2, the set of feedforward artificial neural networks with $d \in \mathbb{N}$ representing the neurons in the hidden layer, w_j the weights in the linear output layer, σ the continuous activation function, and A_j the affine function acting on the input $x \in \mathbb{R}^n$ to be

$$\mathcal{N}_\sigma^n = \{f : f(x) = \sum_{j=1}^{d} w_j \sigma(A_j(x))\}. \tag{3.1}$$

In [18, 19, 20], Hornik et al. define another set of functions, which represents a more general system of networks. These more general networks are

the algebra generated by $\mathcal{N}_\sigma^n$.

Definition 3.1.1. For any continuous function $\sigma\colon \mathbb{R} \to \mathbb{R}$, $d \in \mathbb{N}$ the number of neurons in the hidden layer, w_j the weights in the linear output layer, $l_j \in \mathbb{N}$, and A_{jk} an *affine* function acting on the input $x \in \mathbb{R}^n$, we define the set

$$\mathcal{A}_\sigma^n = \{f : f(x) = \sum_{j=1}^{d} w_j \cdot \prod_{k=1}^{l_j} \sigma(A_{jk}(x))\}, \tag{3.2}$$

to be the algebra generated by $\mathcal{N}_\sigma^n$.

We will firstly prove the general results for the set $\mathcal{A}_\sigma^n$ and subsequently extend them to the set of multilayer feedforward artificial neural networks $\mathcal{N}_\sigma^n$. We notice that the multilayer feedforward artificial neural networks in $\mathcal{N}_\sigma^n$ are a special case of the networks in $\mathcal{A}_\sigma^n$ with the $l_j = 1$, $\forall j \in \mathbb{N}$.

Proposition 3.1.2. *When σ is continuous both $\mathcal{N}_\sigma^n$ and $\mathcal{A}_\sigma^n$ are subsets of $\mathcal{C}^n$.*

Proof. We firstly show that $\mathcal{N}_\sigma^n \subseteq \mathcal{C}^n$. Let $f \in \mathcal{N}_\sigma^n$ then

$$f(x) = \sum_{j=1}^{d} w_j \sigma(A_j(x)).$$

Since σ is continuous then for any given $\epsilon > 0$, there is a $\delta > 0$ such that

$$\|\sigma(A_j(x)) - \sigma(A_j(a))\| < \frac{\epsilon}{\sum_{j=1}^{d} |w_j|},$$

whenever $\|x - a\| < \delta$ with $x, a \in \mathbb{R}^n$.

Then

$$\begin{aligned}\|f(x)-f(a)\| &= \|\sum_{j=1}^{d} w_j\sigma(A_j(x)) - \sum_{j=1}^{d} w_j\sigma(A_j(a))\| \\ &= \|\sum_{j=1}^{d} w_j\left[\sigma(A_j(x)) - \sigma(A_j(a))\right]\| \\ &\leq \sum_{j=1}^{d} |w_j|\|\sigma(A_j(x)) - \sigma(A_j(a))\| \\ &< \sum_{j=1}^{d} |w_j|\frac{\epsilon}{\sum_{j=1}^{d}|w_j|} \\ &= \epsilon.\end{aligned}$$

We now show that $\mathcal{A}_\sigma^n \subseteq \mathcal{C}^n$. Let $f \in \mathcal{A}_\sigma^n$ then

$$f(x) = \sum_{j=1}^{d} w_j \cdot \prod_{k=1}^{l_j} \sigma(A_{jk}(x)).$$

Since σ is continuous and a product of continuous functions is continuous then for any given $\epsilon > 0$, there is a $\delta > 0$ such that

$$\|\prod_{k=1}^{l_j}\sigma(A_{jk}(x)) - \prod_{k=1}^{l_j}\sigma(A_{jk}(a))\| < \frac{\epsilon}{\sum_{j=1}^{d}|w_j|},$$

whenever $\|x-a\| < \delta$ with $x, a \in \mathbb{R}^n$. Then

$$\begin{aligned}\|f(x)-f(a)\| &= \|\sum_{j=1}^{d} w_j \cdot \prod_{k=1}^{l_j}\sigma(A_{jk}(x)) - \sum_{j=1}^{d} w_j \cdot \prod_{k=1}^{l_j}\sigma(A_{jk}(a))\| \\ &= \|\sum_{j=1}^{d} w_j \left[\prod_{k=1}^{l_j}\sigma(A_{jk}(x)) - \prod_{k=1}^{l_j}\sigma(A_{jk}(a))\right]\| \\ &\leq \sum_{j=1}^{d}|w_j|\|\prod_{k=1}^{l_j}\sigma(A_{jk}(x)) - \prod_{k=1}^{l_j}\sigma(A_{jk}(a))\| \\ &< \sum_{j=1}^{d}|w_j|\frac{\epsilon}{\sum_{j=1}^{d}|w_j|} \\ &= \epsilon.\end{aligned}$$

□

In order to apply Theorem 2.3.10 to the set of functions $\mathcal{A}_\sigma^n$, we must show that for any activation function $\sigma : \mathbb{R} \to \mathbb{R}$ that is continuous and non-constant, that $\mathcal{A}_\sigma^n$ is indeed an algebra, that it separates points and vanishes at no point on a compact subset $K \subseteq \mathbb{R}^n$.

Lemma 3.1.3. *For any continuous non-constant function $\sigma : \mathbb{R} \to \mathbb{R}$, $\mathcal{A}_\sigma^n$ is an algebra.*

Proof. Let $f, g \in \mathcal{A}_\sigma^n$ and $x \in \mathbb{R}^n$, then

$$f(x) = \sum_{j=1}^{q} \beta_j \prod_{k=1}^{l_j} (\sigma(A_{jk}(x))),$$
$$g(x) = \sum_{j=1}^{s} \delta_j \prod_{k=1}^{m_j} (\sigma(A'_{jk}(x))).$$

To show that $\mathcal{A}_\sigma^n$ is an algebra, we must show that it is closed under addition, multiplication, and scalar multiplication.

For $x \in \mathbb{R}^n$

$$\begin{aligned}(f+g)(x) &= f(x) + g(x) \\ &= \sum_{j=1}^{q} \beta_j \prod_{k=1}^{l_j} (\sigma(A_{jk}(x))) + \sum_{j=1}^{s} \delta_j \prod_{k=1}^{m_j} (\sigma(A'_{jk}(x))) \\ &= \sum_{j=1}^{t} \alpha_j \prod_{k=1}^{n_j} (\sigma(\overline{A}_{jk}(x)))\,,\end{aligned}$$

where

$$\begin{aligned}\alpha_j &= \beta_j && \text{for } j = 1, 2, \ldots, q; \\ \alpha_j &= \delta_j && \text{for } j = q+1, \ldots, q+s; \\ \overline{A}_{jk}(x) &= A_{jk}(x) && \text{for } k = 1, 2, \ldots, l_j \text{ and } j = 1, 2, \ldots, q; \\ \overline{A}_{jk}(x) &= A'_{jk}(x) && \text{for } k = l_j + 1, \ldots, l_j + n_j \text{ and } j = q+1, \ldots, q+s.\end{aligned}$$

Hence $f + g \in \mathcal{A}_\sigma^n$.

Next for $x \in \mathbb{R}^n$

$$\begin{aligned}(f \cdot g)(x) &= f(x) \cdot g(x) \\ &= \left[\sum_{j=1}^{q} \beta_j \prod_{k=1}^{l_j} (\sigma(A_{jk}(x)))\right] \left[\sum_{i=1}^{s} \delta_j \prod_{p=1}^{m_i} (\sigma(A'_{ip}(x)))\right] \\ &= \sum_{j=1}^{q} \sum_{i=1}^{s} \beta_j \delta_i \prod_{k=1}^{l_j} \prod_{p=1}^{m_i} \sigma(\overline{A}_{jk}(x)) \sigma(\overline{A}_{ip}(x)) \,.\end{aligned}$$

Hence $f \cdot g \in \mathcal{A}_\sigma^n$.

Finally, for $x \in \mathbb{R}^n$ and $a \in \mathbb{R}$ then

$$\begin{aligned}(af)(x) &= af(x) \\ &= a \left[\sum_{j=1}^{q} \beta_j \prod_{k=1}^{l_j} (\sigma(A_{jk}(x)))\right] \\ &= \sum_{j=1}^{q} (a\beta_j) \prod_{k=1}^{l_j} (\sigma(A_{jk}(x))) \,.\end{aligned}$$

Hence $af \in \mathcal{A}_\sigma^n$ and therefore $\mathcal{A}_\sigma^n$ is an algebra. □

Lemma 3.1.4. *For any continuous non-constant function $\sigma : \mathbb{R} \to \mathbb{R}$, $\mathcal{A}_\sigma^n$ separates points on any compact set $K \subseteq \mathbb{R}^n$.*

Proof. If $x, y \in K$ such that $x \neq y$ then there is an affine function $A : \mathbb{R}^n \to \mathbb{R}$, such that $\sigma(A(x)) \neq \sigma(A(y))$. To show this we choose any $a, b \in \mathbb{R}$ with $a \neq b$ such that $\sigma(a) \neq \sigma(b)$. This is possible because σ is a continuous non-constant function on $\mathbb{R}$. We choose the affine function $A(\cdot)$ so that $A(x) = a, A(y) = b$. To see this remember that an affine function $A : \mathbb{R}^n \to \mathbb{R}$ with $x, y \in \mathbb{R}^n$ and $\alpha_i, \beta \in \mathbb{R}$, $\forall i \in \{1, \ldots, n\}$, can be represented as

$$A(x) = \sum_{i=1}^{n} \alpha_i x_i + \beta.$$

This means that we seek the unknowns $(\alpha_1, \ldots, \alpha_n)$ and β which are a solu-

tion to the following equations

$$\sum_{i=1}^{n} \alpha_i x_i + \beta = a,$$
$$\sum_{i=1}^{n} \alpha_i y_i + \beta = b.$$

Since $x \neq y$ this implies that without loss of generality that $x_1 \neq y_1$. So if we set

$$\alpha_2 = \alpha_3 = \cdots = \alpha_n = 0,$$

we must solve

$$\alpha_1 x_1 + \beta = a,$$
$$\alpha_1 y_1 + \beta = b.$$

This leads us to the following definition for the affine function $A : \mathbb{R}^n \to \mathbb{R}$ with $z \in \mathbb{R}^n$

$$A(z) = \frac{(a-b)}{(x_1 - y_1)} \cdot z_1 + 0 \cdot z_2 + \cdots + 0 \cdot z_n + \frac{(bx_1 - ay_1)}{(x_1 - y_1)}.$$

Then $\sigma(A(x)) \neq \sigma(A(y))$. This ensures that $\mathcal{A}_\sigma^n$ separates points on any compact $K \subseteq \mathbb{R}^n$. □

Lemma 3.1.5. *For any continuous non-constant function $\sigma : \mathbb{R} \to \mathbb{R}$, $\mathcal{A}_\sigma^n$ vanishes at no point on a compact subset $K \subseteq \mathbb{R}^n$.*

Proof. To show this we choose any $b \in \mathbb{R}$ such that $\sigma(b) \neq 0$ and set $A(x) = b$. Then $A(\cdot)$ is an affine function on $\mathbb{R}^n$. Then for any $x \in K$, $\sigma(A(x)) = \sigma(b) \neq 0$. This ensures that for any $x \in K$ there exists an $f \in \mathcal{A}_\sigma^n$ with $f(x) \neq 0$. Hence $\mathcal{A}_\sigma^n$ vanishes at no point on any compact set K. □

Theorem 3.1.6. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any continuous non-constant function. Then $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$.*

Proof. Taking note of Lemmas 3.1.3, 3.1.4 and 3.1.5 we have that $\mathcal{A}_\sigma^n$ satisfies the requirements for Theorem 2.3.10 and as such implies that $\mathcal{A}_\sigma^n$ is ρ_K-dense

in the space of continuous real-valued functions on compact K. Since K is any compact subset of $\mathbb{R}^n$, we have that $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$. □

This means that the set of networks $\mathcal{A}_\sigma^n$ is capable of approximating any real-valued function over a compact subset to any degree of accuracy. The compact set requirement holds whenever the possible values for the set of inputs are closed and bounded ($x \in K$), which is generally the case in applications when $K \subseteq \mathbb{R}^n$. We take notice that this result works for any *activation function* that is continuous and non-constant.

In an attempt to strengthen the approximation capabilities of the set of networks $\mathcal{A}_\sigma^n$, we relax the restrictions on the activation functions σ. This seems natural as the requirement that σ be continuous precludes many interesting discontinuous functions. One such function for example, is the *characteristic function* in Example 2.4.11.

In particular the function $1_{[0,\infty)}$, which is not continuous, is commonly used in applications of artificial neural networks. In order to strengthen the approximation capabilities Hornik et al. [20] look at a particular type of activation function.

Definition 3.1.7. A function $\sigma : \mathbb{R} \to [0, 1]$ is said to be a *squashing function* if it is non- decreasing and

$$\lim_{x \to \infty} \sigma(x) = 1\,,$$
$$\lim_{x \to -\infty} \sigma(x) = 0\,.$$

Therefore a sigmoidal function is a special case of a squashing function that is increasing and continuously differentiable.

The argument below will show that for squashing functions the requirement of continuity is not needed, but rather the property that squashing functions are bounded. In order to verify that $\mathcal{A}_\sigma^n$ are uniformly dense on

compacta in $\mathcal{C}^n$ for any squashing function $\sigma : \mathbb{R} \to [0,1]$ we will need the following lemma.

Lemma 3.1.8. *Let $\sigma : \mathbb{R} \to [0,1]$ be a continuous squashing function and $\upsilon : \mathbb{R} \to [0,1]$ be an arbitrary squashing function. Then for every $\epsilon > 0$ there is an element $\tau_\epsilon \in \mathcal{N}_\upsilon^n$ such that*

$$\sup_{x \in \mathbb{R}} |\sigma(x) - \tau_\epsilon(x)| < \epsilon.$$

Proof. Pick an arbitrary $\epsilon > 0$. Without loss of generality we take $\epsilon < 1$. We must now find a finite collection of constants $\beta_j \in \mathbb{R}$, and affine functions A_j , $j = \{1, 2, \ldots, Q-1\}$ such that

$$\sup_{x \in \mathbb{R}} \left| \sigma(x) - \sum_{j=1}^{Q-1} \beta_j \upsilon(A_j(x)) \right| < \epsilon.$$

Now we choose a Q such that $\frac{1}{Q} < \frac{\epsilon}{2}$. We are able to do this as a consequence of the Archimedean Property as the set of natural numbers $\mathbb{N}$ is not bounded above.

For the $j \in \{1, \ldots, Q-1\}$ we set the

$$\beta_j = \frac{1}{Q}.$$

Next choose an $M > 0$ so that

$$\upsilon(-M) < \frac{\epsilon}{2Q},$$
$$\upsilon(M) < 1 - \frac{\epsilon}{2Q}.$$

Such an $M \in \mathbb{R}$ exists because υ is a squashing function.

Subsequently for $j \in \{1, \ldots, Q-1\}$ we also set

$$r_j = \sup\{x : \sigma(x) = \frac{j}{Q}\},$$
$$r_Q = \sup\{x : \sigma(x) = 1 - \frac{1}{2Q}\}.$$

Owing to the fact that σ is a continuous squashing function we are assured that the r_j exist for $j \in \{1, \ldots, Q\}$. This is a consequence of the Intermediate Value Theorem. Also $r_j \leq r_{j+1}$ since σ is non-decreasing.

Finally for any $r < s$ we define $A_{r,s} : \mathbb{R} \to \mathbb{R}$ to be the affine function satisfying

$$A_{r,s}(r) = M,$$
$$A_{r,s}(s) = -M.$$

Seeing as $A_{r,s}$ is an affine function from $\mathbb{R}$ it has the form

$$A_{r,s}(x) = ax + b.$$

Solving for the unknowns $a, b \in \mathbb{R}$ we have

$$A_{r,s}(x) = \frac{M}{s-r}[-2x + (s+r)].$$

The desired approximation is then

$$\tau_\epsilon(x) = \sum_{j=1}^{Q-1} \beta_j \upsilon(A_{r_j, r_{j+1}}(x)).$$

As a result, on each of the intervals

$$(-\infty, r_1]\,,\ (r_1, r_2]\,,\ \ldots\,,\ (r_{Q-1}, r_Q]\,,\ (r_Q, \infty).$$

we have that

$$|\sigma(x) - \tau_\epsilon(x)| < \epsilon.$$

□

Theorem 3.1.9. *For any squashing function $\sigma : \mathbb{R} \to [0, 1]$, the set of networks $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$.*

Proof. In order to prove this, it is sufficient to show that $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{A}_\upsilon^n$, where υ is some continuous squashing function.

For any $\epsilon > 0$ and $h \in \mathcal{C}^n$ we will show that there exist functions $f \in \mathcal{A}_\sigma^n$ and $g \in \mathcal{A}_\upsilon^n$ such that for any compact $K \subseteq \mathbb{R}^n$, making use of the triangle inequality property of the metric ρ_K, we have

$$\begin{aligned}\rho_K(f,h) &\leq \rho_K(f,g) + \rho_K(g,h) \\ &< \epsilon.\end{aligned}$$

For $\mathcal{A}_\sigma^n$ to be uniformly dense on compacta in $\mathcal{A}_\upsilon^n$ we must show that for every compact subset $K \subseteq \mathbb{R}^n$, $\mathcal{A}_\sigma^n$ is ρ_K-dense in $\mathcal{A}_\upsilon^n$. This can be achieved by showing that every function of the form

$$\prod_{k=1}^{l} \upsilon(A_k(\cdot))$$

can be uniformly approximated by members of $\mathcal{A}_\sigma^n$.

We pick an arbitrary $\epsilon > 0$. Because multiplication is continuous and $[0,1]^l$ is compact there exists a $\delta > 0$ such that

$$|a_k - b_k| < \delta,$$

for $0 \leq a_k, b_k \leq 1$ with $k \in \{1, \ldots, l\}$, which implies

$$\left| \prod_{k=1}^{l} a_k - \prod_{k=1}^{l} b_k \right| < \epsilon.$$

Since $[0,1]^l$ is compact, multiplication is uniformly continuous and attains its supremum and infimum on $[0,1]^l$.

Now by Lemma 3.1.8, we know that there exists a function $\tau_\delta \in \mathcal{N}_\sigma^n$ such that

$$\sup_{x \in \mathbb{R}} |\upsilon(x) - \tau_\delta(x)| < \delta,$$

where

$$\tau_\delta(\cdot) = \sum_{t=1}^{T} \beta_t \sigma(A_t(\cdot)) \,.$$

Following on from the fact that

$$|\upsilon(x) - \tau_\delta(x)| < \delta,$$

for all $x \in \mathbb{R}$ and that multiplication is continuous we can conclude that

$$\sup_{x\in\mathbb{R}^n} \left| \prod_{k=1}^{l} \upsilon(A_k(x)) - \prod_{k=1}^{l} \tau_\delta(A_k(x)) \right| < \epsilon.$$

We take notice of the fact that $\overline{A_k} = A_t(A_k)(\cdot)$ is an affine function from $\mathbb{R}^n$ to $\mathbb{R}$. This arises from $A_k : \mathbb{R}^n \to \mathbb{R}$ being an affine function and that

$$\begin{aligned}
\overline{A_k}(x) &= A_t(A_k(x)) \\
&= A_t(\sum_{i=1}^{n} A_{ki}x_i + b_k) \\
&= \beta_t(\sum_{i=1}^{n} A_{ki}x_i + b_k) + b_t \\
&= \sum_{i=1}^{n} (\beta_t A_{ki})x_i + (\beta b_k + b_t) \\
&= \sum_{i=1}^{n} \overline{A_{ki}}x_i + \overline{b_k} \\
&= \overline{A_k}(x)\,.
\end{aligned}$$

Which leads us to realize that

$$\begin{aligned}
\prod_{k=1}^{l} \tau_\delta(A_k(\cdot)) &= \prod_{k=1}^{l} \sum_{t=1}^{T} \beta_t \sigma(A_t(A_k(\cdot))) \\
&\in \mathcal{A}_\sigma^n\,.
\end{aligned}$$

Thus $\prod_{k=1}^{l} \upsilon(A_k(\cdot))$ can be uniformly approximated by elements of $\mathcal{A}_\sigma^n$. This means that $\mathcal{A}_\sigma^n$ is ρ_K-dense in $\mathcal{A}_\upsilon^n$ for any compact subset $K \subseteq \mathbb{R}^n$. Therefore $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{A}_\upsilon^n$.

Now pick any $\epsilon > 0$ and let $h \in \mathcal{C}^n$ be any continuous function. We need to find an $f \in \mathcal{A}_\sigma^n$ such that

$$\rho_K(f, h) < \epsilon,$$

for any compact $K \subseteq \mathbb{R}^n$.

From Theorem 3.1.6, we know that $\mathcal{A}_\upsilon^n$ is uniformly dense on compacta in $\mathcal{C}^n$, since υ is a continuous squashing function which is non-constant. This implies that for all $\epsilon_1 > 0$ there exists $g \in \mathcal{A}_\upsilon^n$ such that

$$\rho_K(g, h) < \epsilon_1,$$

for any compact $K \subseteq \mathbb{R}^n$.

From above we know that $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{A}_\upsilon^n$. This means that by definition for all $\epsilon_2 > 0$ there exists $f \in \mathcal{A}_\sigma^n$ such that

$$\rho_K(f, g) < \epsilon_2,$$

for any compact $K \subseteq \mathbb{R}^n$.

Hence if we let

$$\overline{\epsilon_1} = \frac{\epsilon_1 \epsilon}{\epsilon_1 + \epsilon_2},$$
$$\overline{\epsilon_2} = \frac{\epsilon_2 \epsilon}{\epsilon_1 + \epsilon_2},$$

this implies as a consequence of the triangle inequality for metrics, that

$$\begin{aligned} \rho_K(f, h) &\leq \rho_K(f, g) + \rho_K(g, h) \\ &< \overline{\epsilon_1} + \overline{\epsilon_2} \\ &= \epsilon. \end{aligned}$$

Now this is possible for any $\epsilon > 0$ and $h \in \mathcal{C}^n$ with $K \subseteq \mathbb{R}^n$ compact.

Therefore $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$. □

Having shown that the set of networks $\mathcal{A}_\sigma^n$ are able to approximate any continuous real-valued function for an arbitrary squashing function over any compact subset, we now extend this result to the set of multilayer feedforward artificial neural networks $\mathcal{N}_\sigma^n$. To achieve this we will need the following lemmas.

Lemma 3.1.10. *For every squashing function $\sigma : \mathbb{R} \to [0, 1]$, $\epsilon > 0$, $M > 0$, there is a function $C_{M,\epsilon} \in \mathcal{N}_\sigma^1$ such that*

$$\sup_{x \in [-M,M]} \left| C_{M,\epsilon}(x) - \cos_{[-M,M]}(x) \right| < \epsilon.$$

Proof. Let $F : \mathbb{R} \to \mathbb{R}$ be the *cosine squasher* [14, 20], see Figure 3.1,

$$F(x) = \left[1 + \cos\left(x + \frac{3\pi}{2}\right)\right]\left(\frac{1}{2}\right) 1_{\{\frac{-\pi}{2} \leq x \leq \frac{\pi}{2}\}} + 1_{\{x > \frac{\pi}{2}\}}. \tag{3.3}$$

Now by adding, subtracting, and scaling a finite number of affinely shifted versions of equation (3.3), we can construct the cosine function on any interval $[-M, M]$. This means that the constructed cosine function, $\cos_{[-M,M]}(\cdot)$, has period of $2M$ and is defined by the following equation

$$\cos_{[-M,M]}(x) = \left[1 + \cos\left(x\,\frac{\pi}{2M} + \frac{3\pi}{2}\right)\right]\left(\frac{1}{2}\right) 1_{\{-M \leq x \leq M\}} + 1_{\{x > M\}}. \tag{3.4}$$

Since our constructed cosine function is a continuous squashing function, by Lemma 3.1.8, we know that there exists a function $C_{M,\epsilon} \in \mathcal{N}_\sigma^1$ and by applying the triangle inequality we have that

$$\sup_{x \in [-M,M]} \left| C_{M,\epsilon}(x) - \cos_{[-M,M]}(x) \right| < \epsilon.$$

□

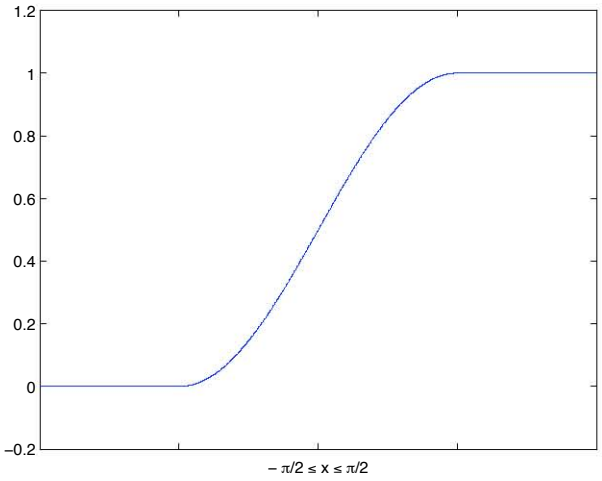

Figure 3.1: Cosine Squasher on the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

Lemma 3.1.11. *Let $g(\cdot) = \sum_{j=1}^{Q} \beta_j \cos(A_j(\cdot))$, where $A_j : \mathbb{R}^n \to \mathbb{R}$ is an affine function. Then for any squashing function $\sigma : \mathbb{R} \to [0, 1]$, any compact subset $K \subseteq \mathbb{R}^n$, and for any $\epsilon > 0$ there exists an $f \in \mathcal{N}_\sigma^n$ such that*

$$\sup_{x \in K} |g(x) - f(x)| < \epsilon.$$

Proof. We pick $M > 0$ such that for $j \in \{1, \ldots, Q\}$ we have $A_j(K) \subseteq [-M, M]$. This is possible due to the fact that Q is finite, K is compact, and the $A_j(\cdot)$ are continuous. By Theorem 2.1.34, every A_j attains its bounds.

For instance, allow

$$M = \max\{\sup A_j(K) : j \in \{1, \ldots, Q\}\},$$
$$-M = \min\{\inf A_j(K) : j \in \{1, \ldots, Q\}\}.$$

Let $Q' = Q\sum_{j=1}^{Q}|\beta_j|$. Now by Lemma 3.1.10, for all $x \in K$ we have that

$$\left|\sum_{j=1}^{Q}\beta_j C_{M,\frac{\epsilon}{Q'}}(A_j(x)) - g(x)\right| < \epsilon.$$

Since $C_{M,\frac{\epsilon}{Q'}} \in \mathcal{N}_\sigma^1$, we see that

$$f(\cdot) = \sum_{j=1}^{Q}\beta_j C_{M,\frac{\epsilon}{Q'}}(A_j(x)) \in \mathcal{N}_\sigma^n\,.$$

□

Theorem 3.1.12. *For any squashing function $\sigma : \mathbb{R} \to [0,1]$, the set $\mathcal{N}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$.*

Proof. Since $\cos(\cdot)$ is a continuous non-constant function, by Theorem 3.1.6 and for $Q, l_j \in \mathbb{N}$, $\beta_j \in \mathbb{R}$, with $A_{jk} : \mathbb{R}^n \to \mathbb{R}$ being an affine function, the trigonometric polynomials

$$\left\{\sum_{j=1}^{Q}\beta_j\prod_{k=1}^{l_j}\cos(A_{jk}(\cdot))\right\}$$

are uniformly dense on compacta in $\mathcal{C}^n$.

Repeatedly applying the trigonometric identity

$$(\cos a)\cdot(\cos b) = \cos(a+b) - \cos(a-b),$$

allows us to rewrite every trigonometric polynomial in the form

$$\sum_{t=1}^{T}\alpha_t\cos(A_t(\cdot))\,,$$

where $a_t \in \mathbb{R}$ and $A_t : \mathbb{R}^n \to \mathbb{R}$ is an affine function.

Now from Lemma 3.1.11, we have that for any squashing function $\sigma : \mathbb{R} \to [0,1]$, any compact subset $K \subseteq \mathbb{R}^n$, and any $\epsilon > 0$ there exists $f \in \mathcal{N}_\sigma^n$ such that

$$\sup_{x \in K} \left| \sum_{t=1}^{T} \alpha_t \cos(A_t(x)) - f(x) \right| < \epsilon.$$

Hence $\mathcal{N}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$. □

We have shown using the ideas of Hornik et al. [20], that for any continuous function there is a multilayer feedforward artificial neural network in the set $\mathcal{N}_\sigma^n$, with $\sigma : \mathbb{R} \to [0,1]$ being an arbitrary squashing function, that can approximate the continuous function to any degree of accuracy on any compact subset $K \subseteq \mathbb{R}^n$ in the ρ_K metric.

3.2 The Method of Hahn-Banch for Continuous Functions

In this section we will show using the ideas of Cybenko [11], how the set of multilayer feedforward artificial neural networks $\mathcal{N}_\sigma^n$, is capable of approximating a continuous function $f : [0,1]^n \to \mathbb{R}$ on the unit hypercube of $\mathbb{R}^n$ under certain conditions placed on the activation function σ.

The main condition that an activation function σ must satisfy, is that it must possess a property that Cybenko [11] defines as *discriminatory*.

Definition 3.2.1. Let μ be a finite, signed, regular Borel measure on the measurable space $([0,1]^n, \mathbb{B})$ and let $A : \mathbb{R}^n \to \mathbb{R}$ be an affine function as defined in Notation 1.3.1. We say that any function $\sigma : \mathbb{R} \to \mathbb{R}$, is *discriminatory* if for any measure μ and any affine function A we have that

$$\int_{[0,1]^n} \sigma(A)\, d\mu = 0\,, \tag{3.5}$$

implies $\mu = 0$.

In order to show, for any continuous activation function $\sigma : \mathbb{R} \to \mathbb{R}$ which is discriminatory, that $\mathcal{N}_\sigma^n$ is dense in the space of continuous functions $\mathcal{C}_{[0,1]^n}$ on the unit hypercube, we need to show that $\mathcal{N}_\sigma^n$ is a linear subspace of $\mathcal{C}_{[0,1]^n}$.

Lemma 3.2.2. *For any continuous function $\sigma : \mathbb{R} \to \mathbb{R}$, $\mathcal{N}_\sigma^n$ is a linear subspace of the space of continuous functions $\mathcal{C}_{[0,1]^n}$ on the unit hypercube.*

Proof. We need to show that for any $f, g \in \mathcal{N}_\sigma^n$ and $\alpha, \beta \in \mathbb{R}$ that

$$\alpha f + \beta g \in \mathcal{N}_\sigma^n .$$

Let $f, g \in \mathcal{N}_\sigma^n$, then for all $x \in [0,1]^n$

$$f(x) = \sum_{j=1}^{q} a_j(\sigma(A_j(x))),$$
$$g(x) = \sum_{j=1}^{s} b_j(\sigma(A'_j(x))).$$

This means that for $x \in [0,1]^n$

$$\begin{aligned}(\alpha f + \beta g)(x) &= \alpha f(x) + \beta g(x) \\ &= \alpha \left[\sum_{j=1}^{q} a_j\, \sigma(A_j(x))\right] + \beta \left[\sum_{j=1}^{s} b_j\, \sigma(A'_j(x))\right] \\ &= \sum_{j=1}^{q} \alpha\, a_j\, \sigma(A_j(x)) + \sum_{j=1}^{s} \beta\, b_j\, \sigma(A'_j(x)) \\ &= \sum_{j=1}^{t} \delta_j\, \sigma(\overline{A}_j(x))\,,\end{aligned}$$

where

$$\begin{aligned} \delta_j &= \alpha\, a_j && \text{for } j = 1, 2, \ldots, q; \\ \delta_j &= \beta\, b_j && \text{for } j = q+1, \ldots, q+s; \\ \overline{A}_j(x) &= A_j(x) && \text{for } j = 1, 2, \ldots, q; \\ \overline{A}_j(x) &= A'_j(x) && \text{for } j = q+1, \ldots, q+s. \end{aligned}$$

Hence $(\alpha f + \beta g) \in \mathcal{N}_\sigma^n$. □

Theorem 3.2.3. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any continuous discriminatory function. Then $\mathcal{N}_\sigma^n$ is dense in $\mathcal{C}_{[0,1]^n}$.*

Proof. By Lemma 3.2.2, we know that $\mathcal{N}_\sigma^n$ determines a linear subspace of $\mathcal{C}_{[0,1]^n}$. We will show that the closure of $\mathcal{N}_\sigma^n$, $\overline{\mathcal{N}_\sigma^n}$, is the same as the space of continuous functions $\mathcal{C}_{[0,1]^n}$ on the unit hypercube .

We show this by the method of contradiction and assume that the closure of $\mathcal{N}_\sigma^n$ is not equal to $\mathcal{C}_{[0,1]^n}$, but is rather a closed proper subspace of $\mathcal{C}_{[0,1]^n}$.

By the Hahn-Banach Extension Theorem 2.7.9 and Theorem 2.7.10, there exists a bounded linear functional L on $\mathcal{C}_{[0,1]^n}$, such that $L \neq 0$ but

$$L(\mathcal{N}_\sigma^n) = L(\overline{\mathcal{N}_\sigma^n}) = 0\,.$$

By the Riesz Representation Theorem 2.7.17, we have that for a bounded linear functional L on $\mathcal{C}_{[0,1]^n}$, there exists a unique regular finite signed measure μ defined on the unit hypercube $[0,1]^n \subseteq \mathbb{R}^n$, such that for all $h \in \mathcal{C}_{[0,1]^n}$

$$L(h) = \int_{[0,1]^n} h(x)\, d\mu\,.$$

Moreover, the norm $\|L\|$ of L equals $\mu([0,1]^n)$.

Now for any affine function A as defined in Notation 1.3.1, $\sigma(A)$ is in the closure of $\mathcal{N}_\sigma^n$. Also since $\overline{\mathcal{N}_\sigma^n} \subseteq \mathcal{C}_{[0,1]^n}$ we must have that

$$L(\sigma(A)) = \int_{[0,1]^n} \sigma(A(x))\, d\mu = 0\,.$$

Seeing as we assumed that σ was discriminatory, this means that the measure $\mu = 0$, the zero measure. This implies that for all $h \in \mathcal{C}_{[0,1]^n}$,

$$L(h) = \int_{[0,1]^n} h(x)\, d\mu = 0\,.$$

But this contradicts our assumption, as the bounded linear functional $L \neq 0$. Hence, the closure of the subspace determined by $\mathcal{N}_\sigma^n$ must equal $\mathcal{C}_{[0,1]^n}$. Which means that $\mathcal{N}_\sigma^n$ is dense in $\mathcal{C}_{[0,1]^n}$. □

We have shown that for any function σ which is continuous and discriminatory, that $\mathcal{N}_\sigma^n$ is dense in the space of continuous functions $\mathcal{C}_{[0,1]^n}$ on the unit hypercube. However in most feedforward artificial neural network applications the activation functions are generally required to be continuous and sigmoidal. This raises the question of which types of functions are discriminatory, as the requirement of the activation function to be discriminatory could preclude many other interesting activation functions. We will show using the ideas of Cybenko [11] that any bounded, measurable sigmoidal function is discriminatory.

Theorem 3.2.4. *Any bounded, measurable sigmoidal function $\sigma : \mathbb{R} \to \mathbb{R}$ is discriminatory. In particular, any continuous sigmoidal function is discriminatory.*

Proof. Let $(\mathbb{R}^n, \mathbb{B}, \mu)$ be a measure space. In order to show the claim we note that for any $x \in \mathbb{R}^n$, $b \in \mathbb{R}$, $\kappa \in \mathbb{N}$, and any affine function $A : \mathbb{R}^n \to \mathbb{R}$ as defined in Notation 1.3.1, we have that

$$\sigma(\kappa(A(x)) + b) \to \begin{cases} 1 & \text{for } A(x) > 0 \quad \text{as } \kappa \to \infty, \\ 0 & \text{for } A(x) < 0 \quad \text{as } \kappa \to \infty, \\ \sigma(b) & \text{for } A(x) = 0 \quad \text{for all } \kappa\,. \end{cases}$$

This is as a result of σ being a sigmoidal function as defined in Section 1.3.2.

This means that as $\kappa \to \infty$, the functions

$$\sigma_\kappa(x) = \sigma(\kappa(A(x)) + b)\,,$$

converge pointwise to the bounded function

$$\gamma(x) = \begin{cases} 1 & \text{for } A(x) > 0, \\ 0 & \text{for } A(x) < 0, \\ \sigma(b) & \text{for } A(x) = 0\,. \end{cases}$$

Then for all $\epsilon > 0$ there exists a $\kappa_0 = \kappa_0(\epsilon) \in \mathbb{N}$ such that for all $\kappa > \kappa_0$, we have

$$|\sigma_\kappa(x) - \gamma(x)| < \epsilon\,.$$

So if $A(x) = 0$, then

$$|\sigma_\kappa(x) - \gamma(x)| = |\sigma(b) - \sigma(b)| < \epsilon\,,$$

also if $A(x) > 0$, then

$$|\sigma_\kappa(x) - \gamma(x)| = |\sigma(\kappa(A(x)) + b) - 1|\ .$$

Noting that as $\kappa \to \infty$,

$$\kappa(A(x)) + b) \to \infty$$

would imply, as a result of σ being a bounded, measurable sigmoidal function, that

$$\sigma(\kappa(A(x)) + b) \to 1\,.$$

So for $\kappa \to \infty$ we have

$$|\sigma(\kappa(A(x)) + b) - 1| \to 0\,.$$

This means that for all $\epsilon > 0$ there exists $\kappa_0 = \kappa_0(\epsilon) \in \mathbb{N}$, such that for all $k \geq \kappa_0$ we have

$$|\sigma(\kappa(A(x)) + b) - 1| < \epsilon\,.$$

Similarly if $A(x) < 0$, then

$$|\sigma_\kappa(x) - \gamma(x)| = |\sigma(\kappa(A(x)) + b) - 0|\ .$$

Noting that as $\kappa \to \infty$,

$$\kappa(A(x)) + b) \to -\infty$$

would imply, as a result of σ being a bounded, measurable sigmoidal function, that

$$\sigma(\kappa(A(x)) + b) \to 0\,.$$

So for $\kappa \to \infty$ we have

$$|\sigma(\kappa(A(x)) + b) - 0| \to 0\,.$$

This means that for all $\epsilon > 0$ there exists $\kappa_0 = \kappa_0(\epsilon) \in \mathbb{N}$, such that for all $k \geq \kappa_0$ we have

$$|\sigma(\kappa(A(x)) + b) - 0| < \epsilon\,.$$

Using the above we will show that if the integral of the functions σ_κ equal zero, that this implies that the measure μ must be the zero measure.

Therefore for an affine function A we define the hyperplane

$$\mathbf{H}_A = \{x \in \mathbb{R}^n : A(x) = 0\}\,,$$

and the open half-spaces

$$\mathbf{P}_A = \{x \in \mathbb{R}^n : A(x) > 0\}\,,$$
$$\mathbf{N}_A = \{x \in \mathbb{R}^n : A(x) < 0\}\,.$$

Then since for all $\kappa > 0$ we have that σ_κ are measurable functions and $|\sigma_\kappa| \leq 1$ as σ is a bounded, measurable sigmoidal function, the Lesbegue Dominated Convergence Theorem 2.4.44 implies that

$$\begin{aligned} 0 &= \lim_{\kappa\to\infty} \int_{[0,1]^n} \sigma_\kappa(x)\, d\mu \\ &= \int_{[0,1]^n} \gamma(x)\, d\mu \\ &= \sigma(b).\mu(\mathbf{H}_A) + 1.\mu(\mathbf{P}_A) + 0.\mu(\mathbf{N}_A)\,. \end{aligned}$$

Remembering that we are trying to approximate continuous functions on the unit hypercube of $\mathbb{R}^n$, we have that $[0,1]^n = \mathbf{H}_A \cup \mathbf{P}_A \cup \mathbf{N}_A$.

Next we must show that the measure of the hyperplane $\mathbf{H}_A$ and half-spaces $\mathbf{P}_A$ and $\mathbf{N}_A$ equalling zero, implies that the measure μ must be the

zero measure. Remembering that the measure μ is a signed measure and can take on negative values, means that this is not trivial. We shall show that the bounded linear functional defined by the measure μ annihilates the unit hypercube $[0,1]^n$.

For a fixed affine function $A : \mathbb{R}^n \to \mathbb{R}$ and any bounded measurable function h, we define the linear functional

$$F(h) = \int_{[0,1]^n} h(A(x))\, d\mu\,.$$

Noting that μ is a finite signed measure, we have that F is a bounded linear functional on $\mathbf{L}_\infty$ as a consequence of Equation 2.3, with the function $g = 1 \in \mathbf{L}_1$.

Let $1_{[0,\infty)}$ be the characteristic function of the interval $[0,\infty)$ as defined in Example 2.4.11, so that

$$\begin{aligned} F(1_{[0,\infty)}) &= \int_{[0,1]^n} 1_{[0,\infty)}(A(x))\, d\mu \\ &= 1.\mu(\mathbf{H}_A) + 1.\mu(\mathbf{P}_A) + 0.\mu(\mathbf{N}_A) \\ &= 0\,. \end{aligned}$$

Since simple functions are linear combinations of characteristic functions, see Definition 2.4.20, and due to the linearity of the bounded linear functional F, it follows that for any simple function φ we have $F(\varphi) = 0$.

By Theorem 2.6.15, we have that the space of simple functions is dense in the space of essentially bounded measurable functions $\mathbf{L}_\infty$. Therefore for any $\epsilon > 0$ and any $h \in \mathbf{L}_\infty$ there exists a simple function φ, such that

$$\|h - \varphi\|_\infty < \epsilon\,.$$

Hence, $h = \varphi$ μ-almost everywhere and by Corollary 2.4.38, we have

$$\begin{aligned} F(h) &= \int_{[0,1]^n} h(A(x))\, d\mu \\ &= \int_{[0,1]^n} \varphi(A(x))\, d\mu \\ &= 0\,. \end{aligned}$$

Therefore the bounded linear functional F annihilates $\mathbf{L}_\infty$.

In particular, this is the case for any affine function $A : \mathbb{R}^n \to \mathbb{R}$ and the bounded measurable functions $\sin(A(x))$ and $\cos(A(x))$. This gives us the following.

$$\begin{aligned} F(\cos(A(x)) + \imath \sin(A(x))) &= \int_{[0,1]^n} \cos(A(x)) + \imath \sin(A(x))\, d\mu \\ &= \int_{[0,1]^n} \exp(\imath\, A(x))\, d\mu \\ &= 0\,. \end{aligned}$$

This means that the fourier transform of the finite signed measure μ is zero, see Definition 2.7.18. By Theorem 2.7.19, we know that the fourier transform is a continuous, linear, one-to-one mapping and this implies that the measure μ must be the zero measure.

Hence, σ is discriminatory. □

Theorem 3.2.5. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any continuous sigmoidal function. Then $\mathcal{N}_\sigma^n$ is dense in $\mathcal{C}_{[0,1]^n}$.*

Proof. By Theorem 3.2.4 we have that σ is a continuous discriminatory function. Applying Theorem 3.2.3 it follows that $\mathcal{N}_\sigma^n$ is dense in $\mathcal{C}_{[0,1]^n}$. □

We have shown using the ideas of Cybenko [11], that for any continuous function in $\mathcal{C}_{[0,1]^n}$ there is a multilayer feedforward artificial neural network in the set $\mathcal{N}_\sigma^n$, with $\sigma : \mathbb{R} \to \mathbb{R}$ being any continuous sigmoidal function, that can approximate the continuous function to any degree of accuracy on the unit hypercube. Note that the restriction of the functions to $[0,1]^n$, the unit hypercube of $\mathbb{R}^n$, is not a problem as it is always possible to scale the inputs into that range. For example consider the following scaling function for the interval $[a,b]$

$$s : [a,b] \to [0,1]\,,\ x \mapsto \frac{1}{b-a}(x-a)\,. \tag{3.6}$$

3.3 The Method of Stone-Weierstrass for Measurable Functions

In section 3.1 we showed using the ideas of Hornik et al. [20], that for any continuous function there exists a multilayer feedforward artificial neural network in the set $\mathcal{N}_\sigma^n$, with $\sigma : \mathbb{R} \to [0,1]$ being an arbitrary squashing function that can approximate the continuous function to any degree of accuracy on any compact subset $K \subseteq \mathbb{R}^n$ in the ρ_K metric.

Continuing to employ the ideas of Honrik et al. [20], we will now extend the set of functions that can be approximated from the set of all continuous functions $\mathcal{C}^n$ to the set of all measurable functions $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$, on the Borel measurable space $(\mathbb{R}^n, \mathbb{B})$ for a probability measure μ defined on $\mathbb{B}$. Note that the requirement for μ to be a probability measure is merely one of convenience born out of practical considerations. The following concepts and ideas will hold for any finite measure μ defined on $\mathbb{B}$ [20].

Proposition 3.3.1. *Let $(\mathbb{R}^n, \mathbb{B})$ be a measurable space and $\sigma : \mathbb{R} \to \mathbb{R}$ be any Borel measurable function. Then $\mathcal{N}_\sigma^n$ and $\mathcal{A}_\sigma^n$ are subsets of $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$.*

Proof. We firstly show that $\mathcal{N}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B})$. Let $f \in \mathcal{N}_\sigma^n$ then

$$f(x) = \sum_{j=1}^{n} w_j \sigma(A_j(x)).$$

Since σ is measurable and knowing that a linear combination of measurable functions is measurable, see Theorem 2.4.14, it follows that

$$f(x) = \sum_{j=1}^{n} w_j \sigma(A_j(x))$$

is measurable. Hence

$$\mathcal{N}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B}).$$

We now show that $\mathcal{A}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B})$. Let $f \in \mathcal{A}_\sigma^n$ then

$$f(x) = \sum_{j=1}^{d} w_j \cdot \prod_{k=1}^{l_j} \sigma(A_{jk}(x)).$$

Since σ is measurable and knowing that a product of measurable functions is measurable, it follows that for all $j \in \{1, \ldots, n\}$, that

$$\prod_{k=1}^{l_j} \sigma(A_{jk}(x))$$

is measurable. Also since a linear combination of measurable functions is measurable, it follows that

$$f(x) = \sum_{j=1}^{d} w_j \cdot \prod_{k=1}^{l_j} \sigma(A_{jk}(x))$$

is measurable. Hence

$$\mathcal{A}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B}).$$

□

Proposition 3.3.2. *The space of continuous functions $\mathcal{C}^n$ on $\mathbb{R}^n$ is a subset of the space of Borel measurable functions $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$,*

$$\mathcal{C}^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B}).$$

Proof. Since continuous functions are measurable, see Example 2.4.12, it follows that

$$\mathcal{C}^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B}).$$

□

Lemma 3.3.3. *Let $(\mathbb{R}^n, \mathbb{B})$ be a measurable space. Then for any finite measure μ on $\mathbb{B}$, $\mathcal{C}^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$.*

Proof. Pick an arbitrary $f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ and choose any $\epsilon > 0$. We must find a $g \in \mathcal{C}^n$, such that

$$\begin{aligned} \rho_\mu(f, g) &= \inf \{\tilde{\epsilon} > 0 : \mu \{x \in \mathbb{R}^n : |f(x) - g(x)| > \tilde{\epsilon}\} < \tilde{\epsilon}\} \\ &< \epsilon. \end{aligned}$$

For sufficiently large $M \in \mathbb{N}$,

$$\int \min\left\{\left|f \cdot 1_{\{|f|<M\}}(x) - f(x)\right|, 1\right\} \, d\mu < \frac{\epsilon}{2}.$$

Since for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that for all $n \geq n_0$,

$$f \cdot 1_{\{|f|<M_n\}} \to f.$$

By Theorem 2.6.17 and Theorem 2.6.18, there exists a continuous function $g \in \mathcal{C}^n$, such that

$$\int \min\left\{\left|f \cdot 1_{\{|f|<M\}}(x) - g(x)\right|, 1\right\} \, d\mu < \frac{\epsilon}{2}.$$

By the triangle inequality,

$$\begin{aligned} |f(x) - g(x)| &= \left|f(x) - f \cdot 1_{\{|f|<M\}}(x) + f \cdot 1_{\{|f|<M\}}(x) - g(x)\right| \\ &\leq \left|f(x) - f \cdot 1_{\{|f|<M\}}(x)\right| + \left|f \cdot 1_{\{|f|<M\}}(x) - g(x)\right|. \end{aligned}$$

Hence,

$$\begin{aligned} \min\{|f(x) - g(x)|, 1\} \leq \min&\left\{\left|f(x) - f \cdot 1_{\{|f|<M\}}(x)\right|, 1\right\} \\ &+ \min\left\{\left|f \cdot 1_{\{|f|<M\}}(x) - g(x)\right|, 1\right\}, \end{aligned}$$

which implies that

$$\begin{aligned} \int \min\{|f(x) - g(x)|, 1\} \, d\mu &\leq \int \min\left\{\left|f(x) - f \cdot 1_{\{|f|<M\}}(x)\right|, 1\right\} \, d\mu \\ &+ \int \min\left\{\left|f \cdot 1_{\{|f|<M\}}(x) - g(x)\right|, 1\right\} \, d\mu \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Therefore

$$\rho_\mu(f, g) < \epsilon$$

and hence $\mathcal{C}^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. □

Theorem 3.3.4. *Let $(\mathbb{R}^n, \mathbb{B})$ be a measurable space. For every continuous non-constant function $\sigma : \mathbb{R} \to \mathbb{R}$ and every probability measure μ on $(\mathbb{R}^n, \mathbb{B})$, $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$.*

Proof. Given any continuous non-constant function σ, it follows from Theorem 3.1.6 that $\mathcal{A}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$. So for every compact subset $K \subseteq \mathbb{R}^n$, $\mathcal{A}_\sigma^n$ is ρ_K-dense in $\mathcal{C}^n$. This means that for all $\epsilon > 0$ and any $g \in \mathcal{C}^n$, there exists an $f \in \mathcal{A}_\sigma^n$ such that

$$\begin{aligned}\rho_K &= \sup_{x \in K} |f(x) - g(x)| \\ &< \epsilon.\end{aligned}$$

For any compact subset $K \subseteq \mathbb{R}^n$, let (f_n) be a sequence of functions in $\mathcal{A}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ such that for all $\epsilon_n > 0$ and $g \in \mathcal{C}^n$ with $\epsilon_n \geq \epsilon_{n+1}$ we have that

$$\rho_K(f_n, g) < \epsilon_n \,.$$

Then as $n \to \infty$,

$$\rho_K(f_n, g) \to 0.$$

This means that (f_n) is a sequence of functions in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ that converges uniformly on compacta to the function g. By Lemma 2.6.24, we have that

$$\rho_\mu(f_n, g) \to 0.$$

Therefore $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{C}^n$.

Now by Lemma 3.3.3, we have that $\mathcal{C}^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. Therefore for $\frac{\epsilon}{2} > 0$ and any $f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ there exists a $g \in \mathcal{C}^n$, such that

$$\rho_\mu(f, g) < \frac{\epsilon}{2} \,.$$

Due to the fact that $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{C}^n$, for $\frac{\epsilon}{2} > 0$ and any $g \in \mathcal{C}^n$, there exists an $h \in \mathcal{A}_\sigma^n$ such that

$$\rho_\mu(g, h) < \frac{\epsilon}{2} \,.$$

Since ρ_μ is a metric on $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ and by the triangle inequality,

$$\begin{aligned}\rho_\mu(f, h) &\leq \rho_\mu(f, g) + \rho_\mu(g, h) \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \,.\end{aligned}$$

Hence $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. □

This means that for any measurable function there exists a network in $\mathcal{A}_\sigma^n$ which is capable of approximating the measurable function to any degree of accuracy. This result holds for any continuous non-constant activation function σ and any probability measure μ on $(\mathbb{R}^n, \mathbb{B})$.

In attempting to strengthen the approximation capabilities of the set of networks $\mathcal{A}_\sigma^n$ in the set of all measurable functions $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$, we will follow in the same vein as with the case of the set of all continuous functions $\mathcal{C}^n$. This is done by relaxing the restrictions on the activation functions σ and using a particular type of activation function called a squashing function. See Definition 3.1.7.

Theorem 3.3.5. *Let $(\mathbb{R}^n, \mathbb{B})$ be a measurable space. For any squashing function $\sigma : \mathbb{R} \to [0, 1]$ and any probability measure μ on $(\mathbb{R}^n, \mathbb{B})$, $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$.*

Proof. From Theorem 3.1.9, we know that for any squashing function σ, the set of networks $\mathcal{A}_\sigma^n$ are uniformly dense on compacta in $\mathcal{C}^n$. So for every compact subset $K \subseteq \mathbb{R}^n$, $\mathcal{A}_\sigma^n$ is ρ_K-dense in $\mathcal{C}^n$. This means that for all $\epsilon > 0$ and any $g \in \mathcal{C}^n$, there exists an $f \in \mathcal{A}_\sigma^n$ such that

$$\begin{aligned} \rho_K &= \sup_{x \in K} |f(x) - g(x)| \\ &< \epsilon. \end{aligned}$$

For any compact subset $K \subseteq \mathbb{R}^n$, let (f_n) be a sequence of functions in $\mathcal{A}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ such that for all $\epsilon_n > 0$ and $g \in \mathcal{C}^n$ with $\epsilon_n \geq \epsilon_{n+1}$ we have that

$$\rho_K(f_n, g) < \epsilon_n \,.$$

Then as $n \to \infty$,

$$\rho_K(f_n, g) \to 0.$$

This means that (f_n) is a sequence of functions in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ that converges uniformly on compacta to the function g. By Lemma 2.6.24, we have

that

$$\rho_\mu(f_n, g) \to 0.$$

Therefore $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{C}^n$.

Now by Lemma 3.3.3, we have that $\mathcal{C}^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. Therefore for $\frac{\epsilon}{2} > 0$ and any $f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ there exists a $g \in \mathcal{C}^n$, such that

$$\rho_\mu(f, g) < \frac{\epsilon}{2}\,.$$

Due to the fact that $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{C}^n$, for $\frac{\epsilon}{2} > 0$ and any $g \in \mathcal{C}^n$, there exists an $h \in \mathcal{A}_\sigma^n$ such that

$$\rho_\mu(g, h) < \frac{\epsilon}{2}\,.$$

Since ρ_μ is a metric on $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ and by the triangle inequality,

$$\begin{aligned}\rho_\mu(f, h) &\leq \rho_\mu(f, g) + \rho_\mu(g, h) \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,.\end{aligned}$$

Hence $\mathcal{A}_\sigma^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. □

We have shown that for any measurable function there exists a network in $\mathcal{A}_\sigma^n$ which is capable of approximating the measurable function to any degree of accuracy. This result holds for an arbitrary squashing function σ and any probability measure μ on $(\mathbb{R}^n, \mathbb{B})$. As before in section 3.1, we now extend this result to the set of multilayer feedforward artificial neural networks $\mathcal{N}_\sigma^n$.

Theorem 3.3.6. *Let $(\mathbb{R}^n, \mathbb{B})$ be a measurable space. For any squashing function $\sigma : \mathbb{R} \to [0, 1]$ and any probability measure μ on $(\mathbb{R}^n, \mathbb{B})$, $\mathcal{N}_\sigma^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$.*

Proof. From Theorem 3.1.12, we know that for any squashing function σ the set of networks $\mathcal{N}_\sigma^n$ are uniformly dense on compacta in $\mathcal{C}^n$. So for every

compact subset $K \subseteq \mathbb{R}^n$, $\mathcal{N}_\sigma^n$ is ρ_K-dense in $\mathcal{C}^n$. This means that for all $\epsilon > 0$ and any $g \in \mathcal{C}^n$, there exists an $f \in \mathcal{N}_\sigma^n$ such that

$$\begin{aligned} \rho_K &= \sup_{x \in K} |f(x) - g(x)| \\ &< \epsilon. \end{aligned}$$

For any compact subset $K \subseteq \mathbb{R}^n$, let (f_n) be a sequence of functions in $\mathcal{N}_\sigma^n \subseteq \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ such that for all $\epsilon_n > 0$ and $g \in \mathcal{C}^n$ with $\epsilon_n \geq \epsilon_{n+1}$ we have that

$$\rho_K(f_n, g) < \epsilon_n \,.$$

Then as $n \to \infty$,

$$\rho_K(f_n, g) \to 0.$$

This means that (f_n) is a sequence of functions in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ that converges uniformly on compacta to the function g. By Lemma 2.6.24, we have that

$$\rho_\mu(f_n, g) \to 0.$$

Therefore $\mathcal{N}_\sigma^n$ is ρ_μ-dense in $\mathcal{C}^n$.

Now by Lemma 3.3.3, we have that $\mathcal{C}^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. Therefore for $\frac{\epsilon}{2} > 0$ and any $f \in \mathcal{M}(\mathbb{R}^n, \mathbb{B})$ there exists a $g \in \mathcal{C}^n$, such that

$$\rho_\mu(f, g) < \frac{\epsilon}{2} \,.$$

Due to the fact that $\mathcal{N}_\sigma^n$ is ρ_μ-dense in $\mathcal{C}^n$, for $\frac{\epsilon}{2} > 0$ and any $g \in \mathcal{C}^n$, there exists an $h \in \mathcal{N}_\sigma^n$ such that

$$\rho_\mu(g, h) < \frac{\epsilon}{2} \,.$$

Since ρ_μ is a metric on $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$ and by the triangle inequality,

$$\begin{aligned} \rho_\mu(f, h) &\leq \rho_\mu(f, g) + \rho_\mu(g, h) \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \,. \end{aligned}$$

Hence $\mathcal{N}_\sigma^n$ is ρ_μ-dense in $\mathcal{M}(\mathbb{R}^n, \mathbb{B})$. □

We have shown using the ideas of Hornik et al. [20], that for an arbitrary squashing function $\sigma : \mathbb{R} \to [0,1]$, any probability measure μ on the measurable space $(\mathbb{R}^n, \mathbb{B})$, that for any measurable function, and any given degree of accuracy there is a multilayer feedforward artificial neural network in $\mathcal{N}_\sigma^n$ which can approximate the measurable function to the specified degree of accuracy in the ρ_μ metric.

The next results will look at the approximation capabilities of the multilayer feedforward artificial neural networks $\mathcal{N}_\sigma^n$, for an arbitrary squashing function $\sigma : \mathbb{R} \to [0,1]$, in the Lebesgue space $\mathbf{L}_p = \mathbf{L}_p(\mathbb{R}^n, \mathbb{B}, \mu)$ for a probability measure μ on $(\mathbb{R}^n, \mathbb{B})$. See Definition 2.5.5.

Corollary 3.3.7. *Let $(\mathbb{R}^n, \mathbb{B})$ be a measurable space. For any squashing function $\sigma : \mathbb{R} \to [0,1]$ and any probability measure μ on $(\mathbb{R}^n, \mathbb{B})$, if there is a compact subset $K \subseteq \mathbb{R}^n$ such that $\mu(K) = 1$, then $\mathcal{N}_\sigma^n$ is ρ_p-dense in $\mathbf{L}_p = \mathbf{L}_p(\mathbb{R}^n, \mathbb{B}, \mu)$ for every $p \in [1, \infty)$.*

Proof. For any $g \in \mathbf{L}_p$ and any $\epsilon > 0$, we must show that there exists a function $f \in \mathcal{N}_\sigma^n$, such that

$$\rho_p(f, g) < \epsilon \,.$$

By Theorem 2.6.17 and Theorem 2.6.18, it follows that for every bounded function $h \in \mathbf{L}_p$ there is a continuous function f', such that

$$\rho_p(h, f') < \frac{\epsilon}{3} \,.$$

For sufficiently large $M \in \mathbb{R}$, letting $h = g \cdot 1_{\{|g| \leq M\}}$, we have that

$$\rho_p(g, h) < \frac{\epsilon}{3} \,.$$

Since for all $\epsilon > 0$ there exists $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that for all $n \geq n_0$,

$$g \cdot 1_{\{|g| < M_n\}} \to g \,.$$

Because $\mathcal{N}_\sigma^n$ is uniformly dense on compacta in $\mathcal{C}^n$, there exists a function $f \in \mathcal{N}_\sigma^n$, such that

$$\sup_{x \in K} |f(x) - f'(x)|^p < \left(\frac{\epsilon}{3}\right)^p .$$

From our assumption that $\mu(K) = 1$ and Corollary 2.4.42, we have

$$\begin{aligned}
\rho_p(f', f) &= \left(\int_K |f'(x) - f(x)|^p \, d\mu \right)^{\frac{1}{p}} \\
&< \left(\int_K \left(\frac{\epsilon}{3}\right)^p \, d\mu \right)^{\frac{1}{p}} \\
&= \left[\left(\frac{\epsilon}{3}\right)^p \cdot \mu(K) \right]^{\frac{1}{p}} \\
&= \frac{\epsilon}{3} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\rho_p(g, f) &\leq \rho_p(g, h) + \rho_p(h, f') + \rho_p(f', f) \\
&< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon .
\end{aligned}$$

Thus $\mathcal{N}_\sigma^n$ is ρ_p-dense in $\mathbf{L}_p = \mathbf{L}_p(\mathbb{R}^n, \mathbb{B}, \mu)$. □

Corollary 3.3.8. *Let $([0,1]^n, \mathbb{B})$ be a measurable space. For any squashing function $\sigma : \mathbb{R} \to [0,1]$ and any probability measure μ on $([0,1]^n, \mathbb{B})$, the set $\mathcal{N}_\sigma^n$ is ρ_p-dense in $\mathbf{L}_p = \mathbf{L}_p([0,1]^n, \mathbb{B}, \mu)$ for every $p \in [1, \infty)$.*

Proof. We know that $[0,1]^n$ is compact. Letting $K = [0,1]^n \subseteq \mathbb{R}^n$ and applying Corollary 3.3.7 with $\mu([0,1]^n) = 1$, it follows that $\mathcal{N}_\sigma^n$ is ρ_p-dense in $\mathbf{L}_p$. □

We have shown using the ideas of Hornik et al. [20], that for an arbitrary squashing function $\sigma : \mathbb{R} \to [0,1]$, any probability measure μ on the measurable space $(\mathbb{R}^n, \mathbb{B})$ and compact subset $K \subseteq \mathbb{R}^n$ such that $\mu(K) = 1$, that for any integrable function in the Lebesgue space $\mathbf{L}_p = \mathbf{L}_p(\mathbb{R}^n, \mathbb{B}, \mu)$, and any given degree of accuracy there is a multilayer feedforward artificial neural network in $\mathcal{N}_\sigma^n$ which can approximate the integrable function to the specified degree of accuracy in the ρ_p metric.

Chapter 4

Conclusions

We have shown that for a broad range of activation functions (continuous sigmoidal and arbitrary squashing) that feedforward artificial neural networks are capable of approximating any continuous function on a compact subset of $\mathbb{R}^n$ to any degree of accuracy [11, 13, 20]. Further for an arbitrary squashing function and any probability measure on the Borel subsets of $\mathbb{R}^n$ that feedforward artificial neural networks are capable of approximating measurable functions and Lebesgue integrable functions [7, 20].

The approaches taken by Cybenko [11] and Hornik et al. [20] differ significantly. Hornik et al. approach the problem by showing that the set of functions generated by multilayer feedword artificial neural networks satisfies the requirements for the Stone-Weierstrass Theorem and is hence dense in the space of all continuous functions and the space of measurable functions. While Cybenko takes the novel approach that for a particular type of activation function, which Cybenko defines as discriminatory in the literature, and shows through an elegant application of the Hahn-Banach and Riesz Representation Theorems that the set of functions generated by multilayer feedword artificial neural networks are dense in the space of all continuous functions on the unit hypercube of $\mathbb{R}^n$. For a further overview of the different methods employed see Melody [25].

The results due to Hornik et al. [20] relax the restrictions on the activation functions from being continuous and sigmoidal [11] to that of merely being a squashing function. This extension covers most activation functions under consideration for practical applications. It follows that the failure to create such a network in practical applications is due to a lack of training, inadequate numbers of neurons in the hidden layer, or that the mapping to be approximated is not of a deterministic nature [14, 18, 20].

Even though the results presented here are for single output feedforward artificial neural networks only, these can be extended to the multi-dimensional cases by using the appropriate multi-dimensional metric [3, 20]. This leads to the intuitive idea that a multi-dimensional output mapping is merely a collection of component single output mappings. As such, one can create many single output feedforward artificial neural networks, linking them together in the appropriate neural network architecture [6, 16].

All of the results presented here have been of an existential nature which do not address key concerns for practical applications. These range from questions on how to determine the network architecture, how many hidden neurons are required to achieve a desired degree of accuracy, or how to train such networks? The question of network architecture has been addressed by Kůrková [22] through the application of Kolmogorov's Theorem and by Cotter [9] designing a network architecture that will ensure the set of functions represented by the feedforward artificial neural networks satisfies the requirements of The Stone-Weierstrass Theorem. There is also an interesting application of Wavelet Theory to the design of the network architecture for multilayer feedforward artificial neural networks in the work by Csáji [10]. The method of training for feedforward artificial neural networks has been explored by Hecht-Nielsen [17] through the application of Backpropagation. Even though the requirement for a continuum of hidden neurons [21] is no

longer required [11, 13, 20], determining the number of neurons in the hidden layer required for a desired level of accuracy is still of fundamental importance. To address this issue Blum and Li [5] have given an upper bound for the number of neurons required in the hidden layer. Unfortunately certain a priori knowledge must be known about the mapping to be approximated, otherwise no such bounds may be accurately estimated[5, 36].

In most practical applications in which feedforward artificial neural networks are used to approximate some unknown mapping, only a partial function table of input-output values is known. This problem is similar to extrapolation of functions as only local information is available. Therefore further information about the mapping to approximated must be determined. Hornik et al. [20] and Attali and Pagès [3] address this by approximating the derivatives of the unknown mapping, while Park and Sandberg [28, 29] and Liao et al [23] have proven that Radial-Basis function networks are capable of approximating continuous and integrable functions.

Therefore a solid theoretical foundation for feedforward artificial neural networks being universal approximators has been developed. In terms of applications, further research is warranted in the area of constructive proofs related to the relationship of the number of neurons in the hidden layer and the desired degree of accuracy for approximation [36, 35].

Bibliography

[1] Biological neuron figure. http://www.abovetopsecret.com/forum/thread532721/pg1, June 2010.

[2] R. Ash. *Real Analysis and Probability*. Academic Press, 1972.

[3] J. G. Attali and G. Pagès. Approximations of functions by a multilayer perceptron: a new approach. *Neural Networks*, 10(6):1069–1081, 1997.

[4] R. G. Bartle. *The Elements of Integration and Lebesgue Measure*. John Wiley and Sons, 1995.

[5] E. K. Blum and L. K. Li. Approximation theory and feedforward networks. *Neural Networks*, 4:511–515, 1991.

[6] M. Burton. *Neural Networks*. Rhodes University, 2008. Artificial Neural Networks Module.

[7] J. L. Castro, Mantas C. J., and Benítez. Neural networks with a continuous squashing function in the ouput are universal approximators. *Neural Networks*, 13:561–563, 2000.

[8] D. L. Cohn. *Measure Theory*. Birkhäuser, 1980.

[9] N. E. Cotter. The Stone-Weirstrass theorem and its application to neural networks. *IEEE Transactions on Neural Networks*, 1(4):290–295, December 1990.

[10] B. C. Csáji. Approximation with artifical neural networks. Master's thesis, Eötvös Loránd University, Hungary, 2001.

[11] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals Systems*, 2:303–314, 1989.

[12] R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2004.

[13] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

[14] A. R. Gallant and H. White. There exists a neural network that does not make avoidable mistakes. *IEEE Second International Conference on Neural Networks*, pages 657–664, 1988.

[15] P. R. Halmos. *Measure Theory.* Springer-Verlag, 1974.

[16] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 2nd edition, 1998.

[17] R. Hecht-Nielsen. Theory of the back propagation neural network. *Proceedings of the International Conference on Neural Networks*, pages 593–608, 1989.

[18] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[19] K. Hornik. Some new results on neural network approximation. *Neural Networks*, 6:1069–1072, 1993.

[20] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[21] B. Irie and S. Miyake. Capabilities of three layer perceptrons. In *IEEE Second International Conference on Neural Networks*, volume 1, pages 641–648, 1988.

[22] V. Kůrková. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5:501–506, 1992.

[23] Y. Liao, S. Fang, and H. L. W. Nuttle. Relaxed conditions for radial-basis function networks to be universal approximators. *Neural Networks*, 16:1019–1028, 2003.

[24] W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysic*, 7:115–133, 1943.

[25] J. M. Melody. On universal approximation using neural networks. Project for 4th year mathematics course, June 1999.

[26] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[27] D. Nauck, F. Klawonn, and R. Kruse. *Foundations of Neuro-Fuzzy Systems*. John Wiley and Sons, 1997.

[28] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3:246–257, 1991.

[29] J Park and I. W. Sandberg. Approximation and radial-basis-function networks. *Neural Computation*, 5:305–316, 1993.

[30] J. D. Pryce. *Basic Methods in Linear Functional Analysis*. Hutchinson and Co., 1973.

[31] R. D. Reed and R. J. Marks, II. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. The MIT Press, 1998.

[32] R. Rojas. *Neural Networks: A Systematic Introduction*. Springer-Verlag, 1996.

[33] W. Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 2nd edition, 1964.

[34] W. Rudin. *Functional Analysis*. McGraw-Hill Series in Higher Mathematics. McGraw-Hill, 1973.

[35] F. Scarselli and A. C. Tsoi. Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, 11(1):15–37, 1998.

[36] D. Tikk, L. T. Kóczy, and T. Gedeon. A survey on universal approximation and its limits in soft computing techniques. *International Journal of Approximate reasoning*, 33:185–202, 2003.