Approximation of Continuous Functions by Artificial Neural Networks

By

Zongliang Ji

* * * * * * * * *

Submitted in partial fulfillment

of the requirements for

Honors in the Department of Mathematics Union College

June, 2018

Abstract

Ji, Zongliang

ADVISOR: George Todd

An artificial neural network is a biologically-inspired system that can be trained to perform computations. Recently, techniques from machine learning have trained neural networks to perform a variety of tasks. It can be shown that any continuous function can be approximated by an artificial neural network with arbitrary precision. This is known as the universal approximation theorem. In this thesis, we will introduce neural networks and one of the first versions of this theorem, due to Cybenko. He modeled artificial neural networks using sigmoidal functions and used tools from measure theory and functional analysis.

Approximation of Continuous Functions by Artificial Neural Networks

Zongliang Ji

September 17, 2018

1 Introduction

Artificial neural networks are computing systems vaguely inspired by the biological neural networks that constitute animal brains. As mostly used in machine learning, an area recently further developed by the increasing data collection ability and computation ability, artificial neural networks play a crucial role in the development of machine learning.

Fascinated by the variety of applications of artificial neural networks in the real world, we sometimes have a shallow understanding of why and how artificial neural network could be a such useful tool in the real world applications. This thesis provides a mathematical background of why artificial neural networks are such powerful computing systems to perform approximation.

The main theorem that shows this ability of artificial neural network is called universal approximation theorem. The theorem states that given a continuous function defined on a certain domain, we are able to use a single hidden layer neural network to approximate this continuous function to arbitrary precision.

The proof of this universal approximation theorem uses two main theorem from real analysis: the Riesz Representation Theorem and the Hahn-Banach Theorem. Thus, this thesis will begin by building up the knowledge and mathematical background from measure theory to integration with respect to a given measure. Then we will show how to prove the Riesz Representation theorem and the Hahn-Banach Theorem using the background knowledge we state and prove at first.

In order to prove this excellent property of artificial neural network, we need to first have a basic understanding of artificial neural network and the component that is crucial to it. Then, we need to have a mathematical definition of neural networks and the definition of the type of continuous function we are

trying to approximate.

Finally, by stating the fundamental construction of the artificial neural network in our context and knowing the statement and proof of the basic two theorem, we are able to illustrate what we mean by approximating any continuous function by an artificial neural network with arbitrary precision. This will be the formal statement of the universal approximation theorem and we will utilize the tools we built up to eventually prove this theorem.

In this thesis, we will closely follow [2].

2 Measure and its construction

2.1 Why we need measure

Suppose we are given a subset of $E \subseteq \mathbb{R}^n$ with $n \in \mathbb{N}$, and that we want to have a function μ with $\mu(E) \in [0, \infty]$, such that μ holds the following properties:

(i) If $E_1, E_2, E_3, ...$ is a finite or infinite sequence of disjoint sets, then $\mu(E_1 \cup E_2 \cup E_3 \cup ...) = \mu(E_1) + \mu(E_2) + ...$

(ii) If E is congruent to F (that is, if E can be transformed into F by translations, rotations, and reflections), then $\mu(E) = \mu(F)$

(iii) $\mu(Q) = 1$, where Q is the unit cube $Q = \{x \in \mathbb{R}^n : 0 \leq x_j < 1 \text{ for } j = 1, ..., n\}$.

These properties follow our intuition of measuring daily objects, when you want to measure the total volume of two individual balls you should add up the volume of each to get the result, and if you kick the ball or move the ball into somewhere else, the volume should be the same. Also, to measure the volume of the ball you also need to have metric like cm^3 or m^3.

However, these nice properties of μ are not able to be fulfilled when we choose the domain of the μ to be the power set of $\mathbb{R}^n$, this is because $\mathbb{R}^n$ has a lot of messy subsets that are not easy for us to measure. (See the section 1.1 in Folland's real analysis text for a concrete counter-example.) So, instead of trying to use a measure to take measurement of every subset of $\mathbb{R}^n$, we try to construct μ on class of subsets of $\mathbb{R}^n$ that includes all the sets that we normally deal with. These new classes or families of subsets of $\mathbb{R}^n$ are called the σ-algebra.

2.2 σ-algebra

Definition 2.1. An *algebra* of sets on X is a nonempty collection $\mathcal{A}$ of subsets of X that is closed under finite unions and complements. If $E_1, .., E_n \in \mathcal{A}$, then $\cup_1^n E_j \in \mathcal{A}$; and if $E \in \mathcal{A}$, then $E^c \in \mathcal{A}$.

Definition 2.2. σ-algebra is an algebra that is closed under countable unions.

Since $\cap_j E_j = (\cup_j E_j^c)^c$, algebras are also closed under finite intersections. Also, if $\mathcal{A}$ is an algebra, then $\emptyset \in \mathcal{A}$ and $X \in \mathcal{A}$ since $\emptyset = E \cap E^c$ and $X = E \cup E^c$ for $E \in \mathcal{A}$.

Example 2.3. σ-algebra

1. If $X = \{a, b, c, d\}$, one possible σ-algebra on X is $\mathcal{A} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$. In general, a finite algebra is always a σ-algebra.

2. If $\{A_1, A_2, A_3, ...\}$ is a countable partition of X then the collection of all unions of sets in the partition (including the empty set) is a σ-algebra.

Definition 2.4. If X has a topology, then we define a Borel σ-algebra on X , as the σ-algebra generated by the family of open sets in X, which is denoted by $\mathcal{B}_X$.

2.3 Measure

We follow Folland [2] that we want the range of our measure to be $[0, \infty]$, and we just defined a family of sets algebra for the domain of the measure. Let's define measure.

Definition 2.5. Let X be a set that is able to generate a σ-algebra $\mathcal{M}$ from it. A *measure* on $\mathcal{M}$ (or on $(X, \mathcal{M})$) is a function $\mu : \mathcal{M} \to [0, \infty]$ such that

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

ii If $\{E_j\}_1^\infty$ is a sequence of disjoint sets in $\mathcal{M}$, then $\mu(\cup_1^\infty E_j) = \sum_1^\infty \mu(E_j)$.

Property (ii) is called countable additivity and implies finite additivity: If $\{E_j\}_1^n$ is a sequence of disjoint sets in $\mathcal{M}$, then $\mu(\cup_1^n E_j) = \sum_1^n \mu(E_j)$.

If X is a set and $\mathcal{M} \subseteq \mathcal{P}(X)$ is a σ-algebra, $(X, \mathcal{M})$ is called a measurable space and the sets in $\mathcal{M}$ are called measurable sets. If μ is a measure on $(X, \mathcal{M})$, then $(X, \mathcal{M}, \mu)$ is called a measure space.

If $\mu(X) < \infty$, then μ is called finite.

If $X = \cup_1^\infty E_j$ where $E_j \in \mathcal{M}$ and $\mu(E_j) < \infty$ for all j, μ is called σ-finite.

Theorem 2.6. *Let $(X, \mathcal{M}, \mu)$ be a measure space.*

1. *(Monotonicity) If $E, F \in \mathcal{M}$, and $E \subseteq F$, then $\mu(E) \leq \mu(F)$*
2. *(Subadditivity) If $\{E_j\}_1^\infty \subseteq \mathcal{M}$, then $\mu(\cup_1^\infty E_j) \leq \sum_1^\infty \mu(E_j)$*
3. *(Continuity from below) If $\{E_j\}_1^\infty \subseteq \mathcal{M}$ and $E_1 \subseteq E_2 \subseteq ...$, then $\mu(\cup_1^\infty) = \lim_{j\to\infty} \mu(E_j)$.*
4. *(Continuity from above) If $\{E_j\}_1^\infty \subseteq \mathcal{M}$, $E_1 \supseteq E_2 \supseteq ...$, and $\mu(E_1) < \infty$, then $\mu(\cap_1^\infty E_j) = \lim_{j\to\infty} \mu(E_j)$.*

Proof. See Folland's proof of Theorem 1.8. □

Definition 2.7. If $(X, \mathcal{M}, \mu)$ is a measure space, a set $E \in \mathcal{M}$ such that $\mu(E) = 0$ is called a *null* set. A measure whose domain includes all subsets of null sets is called *complete*.

In this section, we define what a measure is. It is a map from a σ-algebra to $[0, \infty]$. It evaluates the empty set as 0 and has the additivity property. Let's summarize the material we introduced up to now.

We want to have a function that gives a value to every subset of an abstract set (like $\mathbb{R}^n$), then we found out that there are some subsets that are not easy to be measured by the properties we want for our function. We call this function a "measure". To make this measure have a suitable domain, we defined a family of subsets from a set that we are able to measure and this family of sets is called σ-algebra. Then, we define our measure μ to map from $\mathcal{M}$ which is the σ-algebra of the set X to $[0, \infty]$. X is the set that we want to measure its subsets, $\mathcal{M}$ is a family (collection) of sets that we are able to assign value on and all the elements in $\mathcal{M}$ is a subset of X.

Now that we have the definition of measure, however, it's hard for us to construct a concrete example of a measure using this definition. Since we first need to come up with the σ-algebra of the set and come up with the function that fulfills the two properties in the definition.

In the next section, we will describe a way (Caratheodory's Theorem) to use some functions with a "loose" definition from which it will be easier to get the measure we want.

2.4 Outer Measure

Before introducing the idea of outer measure, it's good to make an analogy to how we can try to measure a area of a region, we first wrap a bigger shape out side this region that is easy to measure and then we shrink this region little by little then finally we make the bigger region as same as the region we first want to measure. We also do this from inside the region to match up with the region value from outside.

The outer measure has the same intuition of this area measurement approach from the outside.

Definition 2.8. An *outer measure* on a nonempty set X is a function $\mu^* : \mathcal{P}(X) \to [0, \infty]$ that satisfies:

i $\mu^*(\emptyset) = 0$,

ii $\mu^*(A) \leq \mu^*(B)$ if $A \subseteq B$,

iii $\mu^*(\cup_1^\infty A_j) \leq \sum_1^\infty \mu^*(A_j)$.

The domain of the outer measure is easier to find since is just the power set of X. We now show that why it's also easy to fulfill the properties of a outer measure. The intuition is that we take arbitrary subset family ε of X like those "bigger regions" and then we use countable unions of sets in ε to form a closer and closer region to the desired one.

Proposition 2.9. *Let $\varepsilon \subseteq \mathcal{P}(X)$ be arbitrary and $\rho : \varepsilon \to [0, \infty]$ be such that $\emptyset \in \varepsilon$, $X \in \varepsilon$, and $\rho(\emptyset) = 0$. For any $A \subseteq X$, define*

$\mu^*(A) = \inf\{\sum_1^\infty \rho(E_j) : E_j \in \varepsilon \text{ and } A \subseteq \cup_1^\infty E_j\}$

Then μ^* is an outer measure.

Proof. We first want to show that μ^* here is well-defined. Since $X \in \varepsilon$, we can have a $\{E_j\}_1^\infty \subseteq \varepsilon$ when all the $E_j = X$. Thus, we have that $A \subseteq \cup_1^\infty E_j$. Then μ^* here is well-defined.

Since $A \subseteq X$ is arbitrary and the result of $\mu^*(A)$ is the sum of the function ρ whose range is $[0, \infty]$, then the range of the μ^* is also $[0, \infty]$. We know that $\rho(\emptyset) = 0$ so if we take $E_j = \emptyset, \forall j$, then $\sum_1^\infty \rho(E_j) = 0$ which means $\mu^*(A) = 0$ since 0 is the smallest element in the set.

Then we want to show property (ii), if $A \subseteq B$, then $\{\sum_1^\infty \rho(E_j) : E_j \in \varepsilon \text{ and } B \subseteq \cup_1^\infty E_j\} \subseteq \{\sum_1^\infty \rho(E_j) : E_j \in \varepsilon \text{ and } A \subseteq \cup_1^\infty E_j\}$ since there exists E_j such that $A \subset E_j \subset B$.

To prove property (iii), suppose $\{A_j\}_1^\infty \subseteq \mathcal{P}(X)$ and let $\epsilon > 0$. For each j, $\exists\{E_j^k\}_{k=1}^\infty \subseteq \varepsilon$ such that $A_j \subseteq \cup_{k=1}^\infty E_j^k$ and $\sum_{k=1}^\infty \rho(E_j^k) \leq \mu^*(A_j) + \epsilon * 2^{-j}$ since $\mu^*(A)$ is the infimum of the set. If $A = \cup_1^\infty A_j$, then $A \subseteq \cup_{k=1}^\infty E_j^k$ and $\sum_{j,k}^\infty \leq \sum_j \mu^*(A_j) + \epsilon$. Since $\mu^*(A)$ is the infimum of the set, then $\mu^*(A) \leq \sum_j \mu^*(A_j) + \epsilon$. Since we say that $\epsilon > 0$ is arbitrary, we are done. □

Now, we show that we can basically use any function ρ and any family of subsets of X to generate a outer measure. We will then use μ^* to generate our measure μ. Before that we need to define a crucial term for our next proof.

Definition 2.10. If μ^* is an outer measure on X, a set $A \subseteq X$ is called μ^*- *measurable* if

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c), \forall E \subseteq X.$$

Since $E \subseteq (E \cap A) \cup (E \cap A^C)$, by the definition of outer measure, $\forall E \subseteq X$,

$$\mu^*(E) \leq \mu^*((E \cap A) \cup (E \cap A^C)) \leq \mu^*(E \cap A) + \mu^*(E \cap A^c).$$

Thus, to show that A is μ^*-measurable, it suffices to show that the reverse inequality and if $\mu^*(E) = \infty$, then the reverse inequality works, so A is μ^*-measurable iff

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c) \forall E \subseteq X \text{ s.t } \mu^*(E) < \infty.$$

This plays in our intuition of constructing the measure is that $\mu^*(E) - \mu^*(E \cap A^c) = \mu^*(E \cap A)$ which is the "inner measure" of A is as same as "outer measure" of A. Now, with everything in our hand, let's show how to construct measure using an outer measure.

Theorem 2.11 (Caratheordory's Theorem)**.** *If μ^* is an outer measure on X, the collection $\mathcal{M}$ of μ^*-measurable sets is a σ-algebra, and the restriction of μ^* to $\mathcal{M}$ is a complete measure.*

Proof. (Follows the proof of Caratheordory's Theorem in Folland [2].)

We want to show that $\mathcal{M}$ is an algebra by first showing that $\mathcal{M}$ is closed under complement since definition of $A \in \mathcal{M}$ as a μ^*-measurable set is symmetric. Then we want to show that it's closed under finite unions.

Let $A, B \in \mathcal{M}$ and $E \subseteq X$, apply definition of μ^*-measurable two times,

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap A) + \mu^*(E \cap A^c) \\ &= \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B) + \mu^*(E \cap A^c \cap B^c)\end{aligned}$$

Since $(A \cup B) = (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$, so by property (iii) of μ^*, we have that,

$$\mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B) \geq \mu^*(E \cap (A \cup B)),$$

so that

$$\mu^*(E) \geq \mu^*(E \cap (A \cup B)) + \mu^*(E \cap (A \cup B)^c) \text{ since } (A \cup B)^C = (A^c \cap B^c).$$

Then by definition of μ^*-measurable, we have that $A \cup B \in \mathcal{M}$, by induction, $\mathcal{M}$ is closed under finite union, so $\mathcal{M}$ is an algebra.

If $A \cap B = \emptyset$, we have that $\mu^*(A \cup B) = \mu^*((A \cup B) \cap A) + \mu^*((A \cup B) \cap A^c) = \mu^*(A) + \mu^*(B)$, so μ^* is finitely additive on $\mathcal{M}$.

To show that $\mathcal{M}$ is a σ-algebra, it will suffice to show that $\mathcal{M}$ is closed under countable disjoint unions, since we already have an algebra. If $\{A_j\}_1^\infty$ is a sequence of disjoint sets in $\mathcal{M}$, let $B_n = \cup_1^n A_j$ and $B = \cup_1^\infty A_j$. Then for any $E \subseteq X$,

$$\begin{aligned}\mu^*(E \cap B_n) &= \mu^*(E \cap B_n \cap A_n) + \mu^*(E \cap B_n \cap A_n^c) \\ &= \mu^*(E \cap A_n) + \mu^*(E \cap B_{n-1}) \\ &= \sum_1^n \mu^*(E \cap A_j) \qquad \text{by induction}\end{aligned}$$

Therefore,

$$\mu^*(E) = \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) \geq \sum_1^n \mu^*(E \cap A_j) + \mu^*(E \cap B^c),$$

since $B_n \subseteq B$ by property (ii).

If we take the let $n \to \infty$ we obtain

$$\begin{aligned}\mu^*(E) &\geq \sum_1^\infty \mu^*(E \cap A_j) + \mu^*(E \cap B^c) \\ &\geq \mu^*(\cup_1^\infty (E \cap A_j)) + \mu^*(E \cap B^c) \\ &= \mu^*(E \cap B) + \mu^*(E \cap B^c) \\ &\geq \mu^*(E),\end{aligned}$$

by property (ii), (iii) and the definition of B.

Thus, $\mu^*(E \cap B) + \mu^*(E \cap B^c) = \mu^*(E)$. It follows that $B \in \mathcal{M}$ and then $\mathcal{M}$ is a σ-algebra.

We already showed that μ^* is finitely additive. If we let $E = B$, then

$$\mu^*(B) = \sum_1^\infty \mu^*(E \cap A_j) + \mu^*(\emptyset) = \sum_1^\infty \mu^*(A_j),$$

then μ^* is countable additive in on $\mathcal{M}$. By definition of the outer measure, $\mu^*(\emptyset) = 0$. Thus, μ^* is a measure restricted on $\mathcal{M}$ (μ^* is a measure when the domain is $\mathcal{M}$).

We then want to show that μ^* is complete measure on $\mathcal{M}$. Let $\mu^*(A) = 0$ for arbitrary A, then for any $E \subseteq X$, we have

$$\mu^*(E) \leq \mu^*(E \cap A) + \mu^*(E \cap A^c) = \mu^*(E \cap A^c) \leq \mu^*(E),$$

by property (ii) of outer measure, so that $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$. Thus, $A \in \mathcal{M}$. Therefore, $\mu^*|\mathcal{M}$ is a complete measure.

□

In this section, we provide a easier and concrete way to construct measure from an outer measure. By Proposition 2.9, we are able to generate an outer measure for an arbitrary subset. This provides us an easy way to find an outer measure.

Then by applying Caratheodory's Theorem, we are able to get a measure that we want by restricting the easily generated outer measure on a collection of μ^*-measurable sets.

By combining the proposition and the theorem, given an arbitrary set X, we are able to find a measure on X in an relatively easier way than just using the definition of measure itself.

3 Integration with measure

In the previous section, we introduced measure and its construction. In this section, we are trying to define integration with respect to a measure over a measure space. Recall integration in a calculus class, $\int_a^b f(x)dx$ is defined as a limit of Riemann sums, which add up the area under the curve of f from a to b. It turns out that we are also able to integrate with respect to different measure spaces. The reason of introducing this new definition of integration is that there are some functions that Riemann integral cannot calculate.

We also need to be able to integrate with respect to the measure μ for the Riesz Representation Theorem. The measure μ here is a special measure called Radon measure that is a restricted version of the Borel measure which we will introduce in the next section. In this section, we will introduce the integration on abstract measure spaces for the future use of the integration of a radon measure in the Riesz representation theorem.

3.1 Measurable Functions

Since we are trying to find the integration of a function with respect to a measure, the functions we choose should have some desired property for the convenience of the definition. These functions are called measurable functions.

We recall that any mapping $f : X \to Y$ between two sets have a inverse mapping $f^{-1} : P(Y) \to P(X)$ which is defined by $f^{-1}(E) = \{x \in X : f(x) \in E\}$. This inverse mapping preserves unions, intersections and

complements. Recall that as introduced in the last section, algebra and σ-algebra are closed under unions and complements. Thus, if $\mathcal{N}$ is a σ-algebra on Y, then $\{f^{-1}(E) : E \in \mathcal{N}\}$ is also a σ-algebra on X.

Definition 3.1. If $(X, \mathcal{M})$ and $(Y, \mathcal{N})$ are measurable spaces, a mapping $f : X \to Y$ is called $(\mathcal{M}, \mathcal{N})$-measurable, or just measurable, if $f^{-1}(E) \in \mathcal{M}$, $\forall E \in \mathcal{N}$

We should know that the composition of measurable mapping is measurable.

Corollary 3.2. *If X and Y are metric spaces, every continuous $f : X \to Y$ is $(\mathcal{B}_X, \mathcal{B}_Y)$-measurable.*

Proof. To show that f is $(\mathcal{B}_X, \mathcal{B}_Y)$-measurable, we need to show that $\forall E \in \mathcal{B}_Y$, $f^{-1}(E) \in \mathcal{B}_X$. Let $E \in \mathcal{B}_Y$ be arbitrary. Since $\mathcal{B}_Y$ is a σ-algebra of open sets. Thus, E is a open set. Since f is continuous, by definition of definition of continuous in real analysis class (f is continuous iff $f^{-1}(U)$ is open in X for every open $U \subseteq Y$), we have that $f^{-1}(E)$ is open. Then $f^{-1}(E) \in \mathcal{B}_X$, as desired. □

If $(X, \mathcal{M})$ is a measurable space, a real- or complex-valued function f on X will be called $\mathcal{M}$-measurable, or just measurable, if it is $(\mathcal{M}, \mathcal{B}_{\mathbb{R}})$ or $(\mathcal{M}, \mathcal{B}_{\mathbb{R}})$ measurable.

With the definition of measurable function, we are able to yield a lot of nice propositions of the measurable functions.

Proposition 3.3. *1. A function $f : X \to \mathbb{C}$ is $\mathcal{M}$-measurable iff Ref and Imf are $\mathcal{M}$-measurable.*

2. If $f, g : X \to \mathbb{C}$ are $\mathcal{M}$-measurable, then so are $f + g$ and fg.

3. If $f, g : X \to \bar{\mathbb{R}}$ are measurable, then so are $max(f, g)$ and $min(f, g)$.

Proof. See Folland Chapter 2. □

For future reference, we present two useful decompositions of functions.

Definition 3.4. If $f, g : X \to \bar{\mathbb{R}}$ where $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, we define the positive and negative parts of f to be

$$f^+(x) = \max(f(x), 0),$$

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

Then by this definition $f = f^+ + f^-$. If f is measurable, so are f^+ and f^-, by Proposition 2.3.

We just defined the measurable function and now we try to discuss functions that are building blocks for the theory of integration. These functions are called the characteristic function and simple function. Simple function will be used later to approximate any arbitrary measurable functions.

Definition 3.5. (Folland Section 2.1) Suppose that $(X, \mathcal{M})$ is a measurable space. If $E \subseteq X$, the characteristic function $\mathcal{X}_E$ of E (sometimes called indicator function of E denoted by 1_E) is defined by

$$\mathcal{X}_E(x) \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}$$

Since the image of the characteristic function is $\{0, 1\}$, then $\mathcal{X}_E$ is measurable iff $E \in \mathcal{M}$.

Definition 3.6. (Folland Section 2.1) A simple function on X is a finite linear combination, with complex coefficients, of characteristic functions of sets in $\mathcal{M}$. Equivalently, $f : X \to \mathbb{C}$ is simple iff f is measurable and the range of f is a finite subset of $\mathbb{C}$. The standard representation of f is:

$$f = \sum_{1}^{n} z_j \mathcal{X}_{E_j}, \text{ where } E_j = f^{-1}(\{z_j\}) \text{ and range}(f) = \{z_1, ..., z_n\}.$$

This definition show that f is a linear combination of characteristic functions and the union of these characteristic functions is X. By definition of simple function, $f + g$ and fg are simple function if f and g are simple function. We defined the simple function using characteristic function and now try to show that simple functions can approximate measurable functions.

Theorem 3.7. *Let $(X, \mathcal{M})$ be a measurable space.*

a If $f : X \to [0, \infty]$ is measurable, there is a sequence $\{\phi_n\}$ of simple functions such that $0 \leq \phi_1 \leq \phi_2 \leq ... \leq f$, $\phi_n \to f$ point-wise, and $\phi_n \to f$ uniformly on any set on which f is bounded.

b If $f : X \to \mathbb{C}$ is measurable, there is a sequence $\{\phi_n\}$ of simple functions such that $0 \leq |\phi_1| \leq |\phi_2| \leq ... \leq |f|$, $\phi_n \to f$ point-wise, and $\phi_n \to f$ uniformly on any set on which f is bounded.

Proof. (a) For $n = 0, 1, 2, ...$ and $0 \leq k \leq 2^{2n} - 1$ and $k \in \mathbb{Z}$, let

$$E_n^k = f^{-1}((k2^{-n}), (k+1)2^{-n}]) \text{ and } F_n = f^{-1}((2^n, \infty]),$$

and define

$$\phi_n = \sum_{k=0}^{2^{2n}-1} k2^{-n}\mathcal{X}_{E_n^k} + 2^n\mathcal{X}_{F_n}$$

The definition of the ϕ_n is easier to understand graphically. By every increment of n, we increase the the value where range is $(2^n, \infty]$ and increase the number of separation on the domain by dividing the domain sets into smaller and smaller intervals corresponding to the range. We can see that $\phi_n \leq \phi_{n+1} \forall n$ since both of the terms of the ϕ_n are increasing with increasing of n. From the first term of the equation, we can see that $0 \leq f - \phi_n \leq 2^{-n}$ where $f \leq 2^n$. Thus, with the increasing n, the difference between ϕ_n and f will decrease. Thus, $\phi_n \to f$ point-wise.

(b) If $f = g + ih$, then we can apply part (a) on positive and negative part of g and h. We will have sequences ψ^+, ψ^-, ζ^+, ζ^- of nonnegative simple functions that increase (approach) to g^+, g^-, h^+, h^-. Let $\phi_n = (\psi^+ - \psi^-) + i(\zeta^+ - \zeta^-)$, since each of these sequence converge to f then $\phi_n \to f$ point-wise. □

In this section, we showed the measurable function that we need to integrate on. We also showed that we can use simple functions which are easy to construct to approximate any arbitrary nonnegative functions by Theorem 3.7. In the next section, we try to define the integration of nonnegative measurable functions with respect to a measure μ.

3.2 Integration of nonnegative functions

In this section we fix a measure space $(X, \mathcal{M}, \mu)$, and we define

$$L^+ = \{f | f : X \to [0, \infty], f \text{ measurable.}\}$$

Definition 3.8. If ϕ is a simple function in L^+ with standard representation $\phi = \sum_1^n a_j \mathcal{X}_{E_j}$, we define the integral of ϕ with respect to μ by

$$\int \phi d\mu = \sum_1^n a_j \mu(E_j)$$

We note that $\int \phi d\mu$ may equal to ∞ since $\mu(E_j)$ may be infinite. If $A \in \mathcal{M}$, then $\phi_{\mathcal{X}_A}$ is also simple by definition ($\phi_{\mathcal{X}_A} = \sum a_j \mathcal{X}_{A \cap E_j}$), and we define $\int_A \phi d\mu = \int \phi_{\mathcal{X}_A} d\mu = \int_A \phi = \int_A \phi(x) d\mu(x)$ and $\int = \int_X$.

Proposition 3.9. *Let ϕ and ψ be simple functions in L^+*

a If $c \geq 0$, $\int c\phi = c \int \phi$

b $\int(\phi + \psi) = \int \phi + \int \psi$

c If $\phi \leq \psi$, then $\int \phi \leq \int \psi$

d The map $A \to \int_A \phi d\mu$ is measure on $\mathcal{M}$.

Proof. See Folland Chapter 2. □

After defining the integration of simple functions, we now extend the integral to all functions $f \in L^+$ be defining

Definition 3.10.

$$\int f d\mu = \sup \left\{ \int \phi d\mu : 0 \leq \phi \leq f, \phi \text{ simple} \right\}$$

The above definition makes sense because the family of simple functions over which the supremum is taken includes f itself. By the definition of $\int f$ and Proposition 3.9, we have that

$$\int f \leq \int g \text{ whenever } f \leq g, \text{ and } \int cf = c \int f \text{ for all } c \in [0, \infty)$$

However, like what we did for measure, it is nice to have a definition of the integral of all the $f \in L^+$, but the definition of $\int f$ takes a supremum over a huge family of simple functions, so it is difficult to calculate or evaluate $\int f$ directly from the definition. We will now show and prove a convergence theorem that allows us to compute $\int f$ in a relatively easier way.

Theorem 3.11. *The Monotone Convergence Theorem. If $\{f_n\}$ is a sequence in L^+ such that $f_j \leq f_{j+1}$ for all j, and $f = \lim_{n\to\infty} f_n$, then $\int f = \lim_{n\to\infty} \int f_n$*

Proof. (Follows from Folland's proof of Theorem 2.14.)

Assume $\{f_n\}$ is a sequence in L^+ such that $f_j \leq f_{j+1}$ for all j, and $f = \lim_{n\to\infty} f_n$, we want to show that $\int f = \lim_{n\to\infty} \int f_n$. Since $\int f \leq \int g$ whenever $f \leq g$, we have that $\{\int f_n\}$ is an increasing sequence of numbers. Thus, $\{\int f_n\}$ has a limit which may be ∞. Moreover, since $f = \lim_{n\to\infty} f_n (sup_n f_n)$, we have that $f_n \leq f \forall n$, so $\lim_{n\to\infty} \int f_n \leq \int f$. We remain to show that $\lim_{n\to\infty} \int f_n \geq \int f$.

To achieve the reverse inequality. We let $\alpha \in (0, 1)$, be arbitrary, let ϕ be a simple function with $0 \leq \phi \leq f$, and $E_n = \{x : f_n(x) \leq \alpha\phi(x)\}$. Since f_n is increasing, we have that $\{E_n\}$ is an increasing sequence of measurable sets and $\cup_1^\infty E_n = X$. Since $E_n \subseteq X$, we have that $\int f \geq \int_{E_n} f_n \geq \alpha \int_{E_n} \phi$. By Proposition 2.11d, we have that map $E_n \to \int_{E_n} \phi d\mu$ is a measure on $\mathcal{M}$. By theorem 1.5c, we have that $E_1 \subseteq E_2 \subseteq ...$ and $\int_{E_n} \phi d\mu$ is the image of the measure, we have that $\lim_{n\to\infty} \int_{E_n} \phi = \int_{\cup_1^\infty E_n} \phi = \int_X \phi = \int \phi$.

Thus,if we take the limit of both side for $\int f \geq \alpha \int_{E_n} \phi$, we will have that

$$\lim_{n\to\infty} \int f_n \geq \alpha \int \phi.$$

Since α is arbitrary, then the above inequality works for all $\alpha < 1$. If we take the supremum of over all simple function $\phi \leq f$, by the definition and the inequality above, $\lim_{n\to\infty} \int f_n \geq \int f$, as desired. □

Now, with the Monotone Convergence Theorem, we can compute $\int f$ by computing $\lim \int \phi_n$ where $\{\phi_n\}$ is a sequence of simple functions that increase to f, in Theorem 2.7 shows that this kind of sequence must exists.

After developing a way to compute integration of any non-negative functions, we now establish the additivity of the integral.

Theorem 3.12. *If $\{f_n\}$ is a finite or infinite sequence in L^+ and $f = \sum_n f_n$, then $\int f = \sum_n \int f_n$.*

Proof. (Follows Folland's proof of Theorem 2.15.)

Let $f_1, f_2 \in L^+$, by Theorem 2.7, we can find $\{\phi_n\}$ and $\{\psi_n\}$ of nonnegative simple functions that increase to f_1 and f_2. Then $\{\phi_n + \psi_n\}$ increases to $f_1 + f_2$, so by the monotone convergence theorem and Theorem 2.9b, we have that

$$\int (f_1 + f_2) = \lim \int (\phi_j + \psi_j) = \lim \int \phi_j + \lim \int \psi_j = \int f_1 + \int f_2$$

By induction, we have that $\int \sum_1^N f_n = \sum_1^N \int f_n$ for any finite N. Letting $N \to \infty$ and applying the monotone convergence theorem, we have that $\int \sum_1^\infty f_n = \sum_1^\infty \int f_n$. □

We now defined the integration of all nonnegative functions. In the next section we are going to define the integral on any real-valued measurable functions f.

3.3 Integration of Complex Functions

We continue to work in a measure space $(X, \mathcal{M}, \mu)$. We now want to define the integral of complex functions.

Definition 3.13. If f^+ and f^- are the positive and negative parts of f and at least one of $\int f^+$ and $\int f^-$ is finite, we define,

$$\int f = \int f^+ - \int f^-$$

If both f^+ and f^- are finite, we say that f is integrable. Since $|f| = f^+ + f^-$, we know that f is integrable iff $\int |f| < \infty$. Now we defined the integral of all the real-valued functions. Let us find out what do these integrable real-valued functions on X look like.

Proposition 3.14. *The set of integrable real-valued functions on X is a real vector space, and the integral is a linear functional on it.*

Before proving this proposition, let's look at the term in this proposition. We want to show that that the set of integrable real-valued functions f that we defined above to be a vector space. To show that a set is a vector space, it must have to operations: scalar multiplication and addition and also follows axioms like closure, commutativity and associativity, etc. We want to show that the integral as a linear functional which is a relative new term that also shows up in the Riesz Representation theorem that we are trying to prove. Thus, let's first have a brief introduction on linear functional.

Definition 3.15. Let $\mathcal{X}$ be a vector space over K, where $K = \mathbb{R}$ or $\mathbb{C}$. A linear map from $\mathcal{X}$ to K is called a linear functional on $\mathcal{X}$.

We will show more of linear functionals in the future sections about proving Hahn-Banach Theorem. For now, from the definition of the linear functional, we can tell that a linear functional maps a vector in the vector space to real or complex number. This follows naturally that if the set of all integrable real-value linear functions is a vector space, then the integral we defined is always a value in $\mathbb{R}$ or $\mathbb{C}$. Thus, the integral would be a linear functional. Now, we have want we need for the proof of Prop 2.14.

Proof. It is quite straight-forward to show that the set of integrable real-valued functions on X is a real vector space since we have $|af + bg| \leq |a||f| + |b||g|$. We know that $f + (-f) = 0$, $f + (g + h) = (f + g) + h$, $f + g = g + h$, $1f = f$, $(ab)f = a(b)f$, $a(f + g) = af + ag$, $0 + f = f + 0 = f$ and $(a + b)f = af + bf$. To show that it is closed under addition, suppose f, g are integrable and let $h = f + g$. Then

$$h^+ - h^- = f^+ - f^- + g^+ - g^-$$
$$h^+ + f^- + g^- = h^- + f^+ + g^+$$
$$\int h^+ + \int f^- + \int g^- = \int h^- + \int f^+ + \int g^+ \text{ by Theorem 2.14}$$

Then, we have that

$$\int h = \int h^+ - \int h^- = \int f^+ - \int f^- + \int g^+ - \int g^- = \int f + \int g$$

To show that the set is closed under scalar multiplication, we just need to show that $\int af = a\int f$, $\forall a \in \mathbb{R}$ which follows from the definition and proposition 2.9a.

The integral is a linear functional since the image of the integration is a real number. □

After defining the integral of real-valued measurable function, we move on to complex-valued measurable function f.

Definition 3.16. If f is a complex-valued measurable function, we say that f is integrable if $\int |f| < \infty$. More generally, if $E \in \mathcal{M}$, f is integrable on E if $\int_E |f| < \infty$. Since $|f| \leq |\mathrm{Re}f| + |\mathrm{Im}f| \leq 2|f|$, f is integrable, and in this case we define

$$\int f = \int \mathrm{Re}f + \int \mathrm{Im}f$$

The space of complex-valued integrable functions is a complex vector space and the integral is a complex-linear functional on it. We denote this space L^1.

In this section, we defined the integral on measurable functions that will be useful for understanding the Riesz Representation Theorem. We also give a definition of linear functional. In the next section, we are going to state and prove the Riesz Representation Theorem.

4 Riesz Representation Theorem

In the beginning of this thesis, we said that we need two major theorems to prove the theorem in Cybenko's paper [1]. We then used two sections to build up what we need to prove the first theorem, Riesz Representation Theorem which connects linear functionals and measures. In this section, we will prove the Riesz Representation Theorem. However, before that we need to introduce some background knowledge for that we will use in the proof of Riesz Representation Theorem.

4.1 Radon Measure and Point Set Topology

In the previous section, we found out that measurable functions can be approximated by continuous functions when we explored how to compute the integration of a measurable function. We want to show the measures that have something similar holds for more general spaces. We also want to show that certain linear functionals on spaces of continuous functions are gives by integration against such measures which is essentially the Riesz Representation Theorem.

In this section, we will define our X as a special space called Locally Compact Hausdorff (LCH) space. The measure that we will focus on are the Borel measures, the σ-algebra generated by open sets that we mentioned in section 1 and its behavior on X.

Definition 4.1. X is Hausdorff Space if $\forall x, y \in X$, if $x \neq y$, then there exist disjoint open sets U, V such that $x \in U$ and $y \in V$.

Definition 4.2. A Hausdorff Space is said to be locally compact if every point has a compact neighborhood (an compact set that contains this point).

There is a important lemma on the LCH that will help us to prove Riesz Representation.

Lemma 4.3. *Urysohn's Lemma for LCH*

Let $K \subseteq U \subseteq X$ and X be a LCH space. Let U be a open set and K be a compact set. Then $\exists f : X \to [0, 1]$ continuous such that f is 1 on K and f is 0 outside of a compact subset of U.

Proof. See Folland Chapter 3. □

Riesz Representation also deals with a specific set of functions that are in the set $C_C(X)$.

Definition 4.4. If f is a function on a topological space X, then the support of f, written $\text{supp}(f)$, is the smallest closed subset of X outside of which f is zero.

$$C_c(X) = \{f : X \to \mathbb{C} | \text{ supp}(f) \text{ is compact, } f \text{ is continuous}\}$$

$C_c(X)$ is called the space of continuous functions on X with compact support.

The following are some definitions and propositions from point-set topology and will be used to prove the Riesz Representation Theorem.

Definition 4.5. If U is open in X and $f \in C_c(X)$, then

$$f \prec U$$

means that $0 \leq f \leq 1$ and $\text{supp}(f) \subset U$. Read as "f is subordinate to U".

The following definition and proposition will help us define a collection of interesting functions that will be useful for the proof and understanding of the Riesz Representation Theorem.

Definition 4.6. If X is a topological space and $E \subset X$, a partition of unity on E is a collection $\{h_\alpha\}_{\alpha \in A}$ of functions in $C(X, [0,1])$ such that

1. $\forall\ x \in X$, x has a neighborhood on which only finitely many h_α's are nonzero;
2. $\sum_{\alpha \in A} h_\alpha(x) = 1$ for $x \in E$.

A partition of unity $\{h_\alpha\}$ is subordinate to an open cover $\mathcal{U}$ of E if for each α there exists $U \in \mathcal{U}$ with $\text{supp}(h_\alpha) \subset U$

Proposition 4.7. *(Folland 4.41)*

Let X be an LCH space, K a compact subset of X, and $\{U_j\}_1^n$ an open cover of K. There is a partition of unity on K subordinate to $\{U_j\}_1^n$ consisting of compactly supported functions.

Recall that a linear functional is a map from a vector space to a scalar.

Definition 4.8. A linear functional I on $C_c(X)$ is positive if $I(f) \geq 0$ whenever $f \geq 0$.

Now, we start to develop the measure that we Riesz Representation use which is called the Radon Measure.

Definition 4.9. Let μ be a Borel measure on X and E a Borel subset (subset of a Borel σ-algebra on X) of X. The measure μ is called outer regular on E if

$$\mu(E) = \inf\{\mu(U) : E \subseteq U,\ U \text{ open}\}$$

and inner regular on E if

$$\mu(E) = \sup\{\mu(K) : K \subseteq E,\ E \text{ compact}\}$$

If μ is outer and inner regular on all Borel sets, μ is called regular.

Here, we define the Radon measure based on inner and outer regularity for the future use in the Riesz Representation Theorem.

Definition 4.10. A Radon measure on X is a Borel measure that is finite on all compact sets, outer regular on all Borel sets, and inner regular on all open sets.

Now we have everything we need to understand (section 1 and 2) and prove (section 3.1) the Riesz Representation. Let's prove it.

4.2 Proof of Riesz Representation

Riesz Representation is one of the two crucial theorems that we need to prove the Universal Approximation Theorem in Cybenko's paper. This theorem describes that certain linear functionals of continuous functions are given by integration against Radon measures. This gives us an important link between measure theory and functional analysis. We will introduce more about functional analysis when we prove the other important theorem called Hahn-Banach Theorem in the future.

Theorem 4.11. *The Riesz Representation Theorem*

If I is a positive linear functional on $C_c(X)$, there is a unique Radon measure μ on X such that $I(f) = \int f d\mu$ for all $f \in C_c(X)$. Moreover, μ satisfies

(a) $\mu(U) = \sup\{I(f) : f \in C_c(X), f \prec U\}$ *for all open* $U \subset X$

(b) $\mu(K) = \inf\{I(f) : f \in C_c(X), f \geq \mathcal{X}_K\}$ *for all compact* $K \subset X$.

Proof. Assume I is a positive linear functional on $C_c(X)$.

We start by showing the uniqueness of the Radon measure μ since the proof of the uniqueness suggests how we going to prove the existence.

Assume the Radon measure μ on X such that $I(f) = \int f d\mu$ for all $f \in C_c(X)$, and let $U \subset X$ be open. We want to show that μ is determined by I on all Borel sets. Then, given a linear functional I, there is only one μ.

We will show part (a) to show that μ is determined by I. To show part (a), we need to show that

$$\mu(U) = \sup\{I(f) : f \in C_c(X), f \prec U\} \text{ for all open } U \subset X.$$

We first want to show that $I(f) \leq \mu(U)$ whenever $f \prec U$. Since U is open and μ is a Radon measure, μ is inner regular on all open sets. Then, we have that

$$\mu(U) = \sup\{\mu(K) : K \subset U, K \text{ compact}\}.$$

Since $f \in C_c(X)$, we have that the $\text{supp}(f)$ is compact. Since $f \prec U$, we have that $\text{supp}(f) \subset U$. Thus, by inner regularity of U, we have that $\mu(\text{supp}(f)) \leq \mu(U)$. We remain to show that $I(f) \leq \mu(\text{supp}(f))$. Since

$f \prec U$, we have that $0 \leq f \leq 1$. By the assumption and the definition of the integration with respect to measure, we have that

$$I(f) = \int f d\mu = \sum_1^n a_j \mu(E_j) \leq \sum_1^n \mu(E_j)$$

where $\cup_1^n E_j \subset X$ and $f(x) \neq 0$ when $x \in E_j$. Since $\text{supp}(f)$ is by definition the closure of $\{x : f(x) \neq 0\}$. Thus, $\cup_1^n E_j \subset \text{supp}(f)$. By finite additivity and monotonicity of measure, we have that

$$I(f) \leq \sum_1^n \mu(E_j) \leq \mu(\cup_1^n E_j) \leq \mu(\text{supp}(f)),$$

as desired.

We now need to show that $\mu(U)$ is the supremum of the set of $I(f)$. Let $K \subset U$ be compact, by Urysohn's Lemma for LCH (3.3), we have a $f \in C_c(X)$ such that $f \prec U$ and $f = 1$ on K. Thus, we have that $\mu(K) \leq \int f d\mu = I(f)$ since $K \subset A$ where $A = \{x : f(x) \neq 0\}$. Then we have that $I(f)$ is an upper bound of $\{\mu(K) : K \subset U, K \text{ compact}\}$. By the inner regularity of U, we have that $I(f) \geq \mu(U)$ ($\mu(U)$ is the least upper bound of the set $\{\mu(K)\}$). Since we show in the previous part that $I(f) \leq \mu(U)$ when $f \prec U$ and $f \in C_c(X)$. Thus,

$$\mu(U) = \sup\{I(f) : f \in C_c(X), f \prec U\} \text{ for all open } U \subset X.$$

By showing part (a), we show that μ is determined by I on open sets. μ is also determined by all Borel sets because Radon measure is outer regular on all Borel sets and outer regularity says that measure of the Borel set is the infimum of measure of open sets.

After showing the uniqueness, we know that we show part (a) along the way. Thus, we will show the existence of such a Radon Measure μ by defining it using part (a).

We define

$$\mu(U) = \sup\{I(f) : f \in C_c(X), f \prec U\} \text{ for all open } U \subset X.$$

We then try to show that this μ is a Radon Measure which is the proof of the existence.

The outline of proving such μ is a Radon Measure is by proving μ is a Borel measure on X and then show that μ is outer regular on all Borel sets and inner regular on all open sets. Then show that such μ has the property of $I(f) = \int f d\mu$.

After defining μ, for arbitrary $E \subseteq X$, we define

$$\mu^*(E) = \inf\{\mu(U) : E \subset U, U \text{ open }\}.$$

By definition of μ, we have that $\mu(U) \leq \mu(V)$ if $U \subseteq V$, hence $\mu^*(A) = \mu(A)$ if A is open by the definition of $\mu^*(A)$.

We then show that

(i) μ^* is an outer measure.

(ii) Every open set is μ^*-measurable.

At this point, we can apply Caratheodory's Theorem to construct our measure. Since $\mathcal{B}_X$ is a collection of open sets and open sets are μ^*-measurable, $\mathcal{B}_X$ is a Borel σ-algebra and $\mu^*|\mathcal{B}_X$ is a Borel measure. Since $\mu^*(A) = \mu(A)$ when A open, we have that $\mu^*|\mathcal{B}_X = \mu$ a Borel measure. By the definition of μ^*, with the restriction here that $E \in \mathcal{B}_X$, we have that μ is outer regular. We then show that

(iii) μ satisfies part (b).

Part (b) implies that μ is finite on the compact set since $I(f)$ is finite and implies μ is inner regular on all open sets since let U be open, let $\alpha < \mu(U)$. We choose $f \in C_c(X)$ such that $f \prec U$ and $I(f) > \alpha$ and let $K = \text{supp}(f)$. Thus, $\text{supp}(f) \subset U$. If $g \in C_c(X)$ and $g \geq \mathcal{X}_K$, then $g - f \geq 0$. By property of linear functionals, we have that $I(g) \geq I(f) > \alpha$. By part b, since $\mu(K)$ is a infimum, then $\mu(K) > \alpha$. Since $\mu(\text{supp}(f)) \geq I(f)$, we have that $\mu(K) \geq I(f)$. Thus, by definition of $\mu(U)$, $\mu(U) \leq \mu(K)$. However, $K \subset U$ by $f \prec U$. Then we have that $\mu(U)$ is the supremum of $\{\mu(K) : K \subseteq E, K \text{ compact }\}$. Thus, μ is inner regular on U. Finally, we need to prove that

(iv) $I(f) = \int f d\mu$

With all above, this completes the proof of the Riesz representation Theorem.

Proof of (i):

We want to show that μ^* is an outer measure by using Proposition 1.8. By the definition of μ, we have that $\mu : \mathcal{C} \to [0, \infty]$ where $\mathcal{C}$ is a collection of open sets. (Since $f \prec U$, $0 \leq f \leq 1$, then $I(f) \geq 0$. The range of μ is $[0, \infty]$.) Also, since X and $\emptyset$ are open, we have that $X, \emptyset \in \mathcal{C}$. We also have that $\mu(\emptyset) = 0$ by the definition of μ.

It suffices to show that if $\{U_j\}$ is a collection of open sets and $U = \cup_1^\infty U_j$ where $U \subset X$ is an arbitrary open set, then $\mu(U) \leq \sum_1^\infty \mu(U_j)$. Since by the definition of μ^*, for arbitrary $E \subset X$, we have that

$$\mu^*(E) = \inf\{\mu(U) : E \subset U, U \text{ open}\}$$

Since $U \subset \mathcal{C}$ and $\mu(U) \leq \sum_1^\infty \mu(U_j)$ and $U = \cup_1^\infty U_j$, we have that

$$\mu^*(E) = \inf\left\{\sum_1^\infty \mu(U_j) : U_j \text{ open}, E \subset \cup_1^\infty U_j = U\right\}$$

Thus, by Proposition 1.8, μ^* is an outer measure.

We remain to show that if $\{U_j\}$ is a collection of open sets and $U = \cup_1^\infty U_j$ where $U \subset X$ is an arbitrary open set, then $\mu(U) \leq \sum_1^\infty \mu(U_j)$.

Assume $\{U_j\}$ is a collection of open sets and $U = \cup_1^\infty U_j$. By the definition of μ, we have that $f \in C_C(X)$, and $f \prec U$. Since $f \prec U$, we have that $\text{supp}(f) \subset U$. Then $\text{supp}(f) \subset \cup_1^\infty U_j$. Let $K = \text{supp}(f)$, since $f \prec U$, we have that K is compact. Thus, relabeling if necessary K has an finite open cover $\cup_1^n U_j$. By Proposition 4.7, we have that there exits a partition of unity on K subordinate to $\cup_1^n U_j$ consisting of compactly supported functions.

By the definition of partition of unity, we have $g_1, ..., g_n \in C_c(X)$ with $g_j \prec U_j$ and $\sum_1^n g_j = 1$ on K.

Then we have that $f = f\sum_1^n g_j = \sum_1^n fg_j$. Since $f \prec U$, then $0 \leq f \leq 1$. Then $fg_j \leq g_j$, we have that $fg_j \prec U_j$. Thus, by the definition of μ, we have that

$$I(f) = \sum_1^n I(fg_j) \leq \sum_1^n \mu(U_j) \leq \sum_1^\infty \mu(U_j).$$

Since by definition of μ, $\mu(U)$ is the least upper bound of $I(f)$ and $f \in C_C(X)$ is arbitrary. Thus, we have that $\mu(U) \leq \sum_1^\infty \mu(U_j)$, as desired.

Proof of (ii):

Let $U \subset X$ be open, we want to show that U is μ^*-measurable. We must show that, with any $E \subset X$ such that $\mu^*(E) < \infty$,

$$\mu^*(E) \geq \mu^*(E \cap U) + \mu^*(E \cap U^C).$$

First, we suppose E is open. Then $E \cap U$ is open. By the definition of $\mu(E \cap U)$, given a $\varepsilon > 0$, we can choose a $f \in C_C(X)$ such that $f \prec E \cap U$ and $I(f) > \mu(E \cap U) - \varepsilon$. Since supp is closed, then

$\text{supp}(f)^C$ is open. Thus, $E\backslash(\text{supp}(f))$ is open, then we have a function $g \in C_C(X)$ such that $f \prec E\backslash\text{supp}$ and $I(g) > \mu(E\backslash\text{supp}(f)) - \varepsilon$. Since g is defined on $E\backslash\text{supp}(f)$, then when $g > 0$, $f = 0$. Thus, $0 \leq f+g \leq 1$. Since $\text{supp}(f) \subset E \cap U$ and $\text{supp}(g) \subset E\backslash\text{supp}(f)$, we have that $\text{supp}(f) \cap \text{supp}(g) = \emptyset$. Thus, $\text{supp}(f+g) \subset E$. Since $\text{supp}(f+g) \subset E$, we have that $f + g \prec E$ and

$$\begin{aligned}
\mu(E) &\geq I(f+g) \\
&= I(f) + I(g) \\
&> \mu(E \cap U) + \mu(E\backslash\text{supp}(f)) - 2\varepsilon \\
&\geq \mu^*(E \cap U) + \mu^*(E\backslash\text{supp}(f)) - 2\varepsilon \\
&\geq \mu(E \cap U) + \mu(E\backslash U) - 2\varepsilon
\end{aligned}$$

since $\text{supp}(f) \subset U$, then $E\backslash\text{supp}(f) \subseteq E\backslash U$. Let $\varepsilon \to 0$, we have desired inequality.

For general $E \subset X$, since $\mu^*(E) < \infty$ and by the definition of $\mu^*(E)$ (μ^* is the infimum of μ), given a $\varepsilon > 0$, we can choose an open set $V \supset E$ such that $\mu(V) < \mu^*(E) + \varepsilon$. Thus, we have that

$$\begin{aligned}
\mu^*(E) + \varepsilon &> \mu(V) \\
&\geq \mu^*(V \cap U) + \mu^*(V\backslash U) \text{ by the previous paragraph} \\
&\geq \mu^*(E \cap U) + \mu^*(E\backslash U) \text{ since } E \subset V.
\end{aligned}$$

Let $\varepsilon \to 0$, we have the desired inequality.

Proof of (iii):

We want to show that μ satisfies part (b). Let $K \subset X$ be arbitrary compact set. Let $f \in C_C(X)$ and $f \geq \mathcal{X}_K$. We want to show that $\mu(K) = \inf\{I(f)\}$. Given $\varepsilon > 0$, let $U_\varepsilon = \{x : f(x) > 1 - \varepsilon\}$. Since $f \geq \mathcal{X}_K$, we have that $K \subset U_\varepsilon$. Since $f \in C_C(X)$, f is continuous, since $\{f(x) > 1 - \varepsilon\}$ is open, we have that $f^{-1}(\{f(x) > 1 - \varepsilon\}) = U_\varepsilon$ is open. Thus, U_ε is open. If $g \prec U_\varepsilon$, then $0 \leq g \leq 1$. Thus, $\frac{f}{1-\varepsilon} - g \geq 0$ on U_ε. By the linear property of linear functional, we have that

$$I(g) \leq \frac{I(f)}{1-\varepsilon}$$

Thus, $\frac{I(f)}{1-\varepsilon}$ is an upper bound of $I(g)$, then by definition of μ from part (a), we have that $\mu(U_\varepsilon) \leq \frac{I(f)}{1-\varepsilon}$. Since $K \subset U_\varepsilon$, we have

$$\mu(K) \leq \mu(U_\varepsilon) \leq \frac{I(f)}{1-\varepsilon}$$

Let $\varepsilon \to 0$, then $\mu(K) \leq I(f)$. Thus, we've shown that $\mu(K)$ is an lower bound of $I(f)$.

On the other hand, for any open set $U \supset K$, by Urysohon's Lemma on LCH, we can choose an $f \in C_C(X)$ such that $f \geq \mathcal{X}_K$ and $f \prec U$. Thus, by part a, we have that $I(f) \geq \mu(U)$. Since μ is outer regular on K, we have that $\mu(K) = \inf\{\mu(U)\}$. Since $\mu(K)$ is the greatest lower bound of $\mu(U)$ and $\mu(K) \leq I(f) \leq \mu(K)$, we have that $\mu(K)$ is the greatest lower bound of $I(f)$, as desired.

Proof of (iv): (Follows Folland [2].)

We want to show that for a Radon measure μ on X and a positive linear functional I on $C_C(X)$, $I(f) = \int f d\mu$ for all $f \in C_C(X)$. Since $C_C(X)$ is the linear span of $C_C(X, [0,1])$. It suffices for us to show that $I(f) = \int f d\mu$, for all $f \in C_C(X, [0,1])$.

Let $f \in C_C(X, [0,1])$ be arbitrary, we want to show that $I(f) = \int f d\mu$. We have that $0 \leq f \leq 1$. Given $N \in \mathbb{N}$, for $1 \leq j \leq N$, let $K_j = \{x : f(x) \leq \frac{j}{N}\}$ and let $K_0 = \text{supp}(f)$. Define $f_1, ..., f_N \in C_C(X)$ such that

$$f_j = \begin{cases} 0 & \text{if } x \notin K_{j-1} \\ f(x) - \frac{j-1}{N} & \text{if } x \in K_{j-1} \backslash K_j \\ f_j(x) = \frac{1}{N} & \text{if } x \in K_j \end{cases}$$

In other words,

$$f_j = \min\left\{\max\left\{f - \frac{j-1}{N}, 0\right\}, \frac{1}{N}\right\}.$$

Given this definition, we have that $K_0 \supseteq K_1 \supseteq K_2 \supseteq ... \supseteq K_N$ and K_j is compact. We also see that $\sum_1^N f_j = f$ and

$$\frac{\mathcal{X}_{K_j}}{N} \leq f_j \leq \frac{\mathcal{X}_{K_{j-1}}}{N}$$

Thus, by definition of integration with respect to measure (Definition 3.8), we have that

$$\frac{\mu(K_j)}{N} \leq \int f_j d\mu \leq \frac{\mu(K_{j-1})}{N}$$

Since

$$\frac{\mathcal{X}_{K_j}}{N} \leq f_j \leq \frac{\mathcal{X}_{K_{j-1}}}{N},$$

we have that $Nf_j \leq \mathcal{X}_{K_{j-1}}$. Thus $0 \leq Nf_j \leq 1$ and $K_j \subseteq \text{supp}(Nf_j) \subseteq K_{j-1}$. For any open set U that

contains K_{j-1}, we have that $\text{supp}(f_j) \subset U$. Thus, $Nf_j \prec U$. By part (a),

$$I(Nf_j) \leq \mu(U)$$
$$NI(f_j) \leq \mu(U)$$
$$I(f_j) \leq \frac{\mu(U)}{N}$$

Since μ is outer regular on K_{j-1}, we have that

$$I(f_j) \leq \frac{\mu(K_{j-1})}{N}$$

Since

$$\frac{\mathcal{X}_{K_j}}{N} \leq f_j \leq \frac{\mathcal{X}_{K_{j-1}}}{N},$$

we have that $\mathcal{X}_{K_j} \leq Nf_j$. Thus, by part (b),

$$\mu(K_j) \leq I(Nf_j)$$
$$\mu(K_j) \leq NI(f_j)$$
$$\frac{\mu(K_j)}{N} \leq I(f_j)$$

Combine the above two inequality, we have

$$\frac{\mu(K_j)}{N} \leq I(f_j) \leq \frac{\mu(K_{j-1})}{N}.$$

Since $f = \sum_1^N f_j$, we have that

$$\sum_1^N \frac{\mu(K_j)}{N} \leq \int f d\mu \leq \sum_0^{N-1} \frac{\mu(K_{j-1})}{N} \tag{4.12}$$
$$\sum_1^N \frac{\mu(K_j)}{N} \leq I(f) \leq \sum_0^{N-1} \frac{\mu(K_{j-1})}{N}. \tag{4.13}$$

Subtract 4.13 by 4.12, we know that

$$0 \leq I(f) - \int f d\mu \leq \frac{\mu(K_0) - \mu(K_N)}{N} \leq \frac{\mu(\text{supp}(f))}{N}$$

Since $f \in C_C(X, [0,1])$, we have that $\mu(\text{supp}(f)) < \infty$, then let $N \to \infty$, we have that

$$0 \leq I(f) - \int f d\mu \leq 0.$$

Thus,

$$I(f) = \int f d\mu.$$

□

Now, we prove the Riesz Representation Theorem which connects the measure theory and the linear functionals. This theorem is crucial for the proof of theorem in Cybenko's paper[1]. We are now going to explore the second theorem we need for the Cybenko paper [1] which is more related to functional analysis. Functional analysis includes the linear functionals we briefly defined before. In the next section, we will introduce elements in functional analysis and state and prove the second important theorem called the Hahn-Banach Theorem. Then we will use these two theorem to prove the theorem in Cybenko's paper [1] by way of contradiction.

5 Elements of Functional Analysis with Hahn-Banach Theorem

Functional Analysis is the study of infinite-dimensional vector spaces over $\mathbb{R}$ or $\mathbb{C}$ and the linear maps between them. This area is related to linear algebra but different in that the study of functional analysis consider topology related to the vector spaces. In order to show the Universal Approximation Theorem, we also need another crucial theorem called Hahn-Banach Theorem. In this section, we will introduce the material that will help us prove the Hahn-Banach Theorem.

5.1 Normed Vector Spaces

In this section, we will associate vector spaces with a special function called a norm. With such a function, we are able to define a topology on vector space with norm by giving it a metric. This essentially connects linear algebra together with topology which gives us a more powerful tool to consider.

Definition 5.1. Let K denote either $\mathbb{R}$ or $\mathbb{C}$, and let $\mathcal{X}$ be a vector space over K. We denote the zero element of $\mathcal{X}$ by 0. If $x \in \mathcal{X}$, we denote by Kx the one-dimensional subspace spanned by x. Also, if $\mathcal{M}$ and $\mathcal{N}$ are subspaces of $\mathcal{X}$, $\mathcal{M} + \mathcal{N}$ denotes the subspace $\{x + y : x \in \mathcal{M}, y \in \mathcal{N}\}$.

A semi-norm on $\mathcal{X}$ is a function $x \to ||x||$ from $\mathcal{X}$ to $[0, \infty)$ such that

1. $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in \mathcal{X}$ (triangle inequality)
2. $||\lambda x|| = |\lambda|||x||$ for all $x \in \mathcal{X}$ and $\lambda \in K$.

The second property of the semi-norm implies that $||0|| = 0$.

Definition 5.2. A semi-norm such that $||x|| = 0$ only when $x = 0$ is called a norm, and a vector space equipped with a norm is called a normed vector space (or normed linear space.)

If $\mathcal{X}$ is a normed vector space, then $d(x, y) = ||x - y||$ is a metric on $\mathcal{X}$.

5.2 Linear Functionals and the Proof of Hahn-Banach Theorem

In this section, we will state and prove the Hahn-Banach Theorem.

Definition 5.3. Let $\mathcal{X}$ be a vector space over K, where $K = \mathbb{R}$ or $\mathbb{C}$. A linear map from $\mathcal{X}$ to K is called a linear functional on $\mathcal{X}$.

Definition 5.4. If $\mathcal{X}$ is a real vector space, a sub-linear functional on $\mathcal{X}$ is a map $p : \mathcal{X} \to \mathbb{R}$ such that

$p(x + y) \leq p(x) + p(y)$ and $p(\lambda x) = \lambda p(x)$ for all $x, y \in \mathcal{X}$ and $\lambda \geq 0$.

For example, every semi-norm is a sub-linear functional.

We now introduce our second important theorem, the Hahn-Banach Theorem. The theorem guarantees us that there will be interesting linear functional exists outside the subspace of the vector space.

Theorem 5.5. *The Hahn-Banach Theorem*

Let $\mathcal{X}$ be a real vector space, p a sub-linear functional on $\mathcal{X}$, $\mathcal{M}$ a subspace of $\mathcal{X}$, and f a linear functional on $\mathcal{M}$ such that $f(x) \leq p(x)$ for all $x \in \mathcal{M}$.

Then there exists a linear functional F on $\mathcal{X}$ such that $F(x) \leq p(x)$ for all $x \in \mathcal{X}$ and $F|\mathcal{M} = f$.

Proof. (Follows from Folland Chapter 5 [2].)

We begin by showing that if $x \in \mathcal{X}\backslash\mathcal{M}$, f can be extended to a linear functional g on $\mathcal{M} + \mathbb{R}x$ satisfying $g(y) \leq f(y)$ where $y \in \mathcal{X}$.

In other words, assume the set up, we first show that there exits a linear functional g on $\mathcal{M} + \mathbb{R}x$ such that $g(a) \leq p(a)$ for all $a \in \mathcal{X}$ and $g|\mathcal{M} = f$.

Let $y_1, y_2 \in \mathcal{M}$ and $x \in \mathcal{X}\backslash\mathcal{M}$. We have that

$$f(y_1) + f(y_2) = f(y_1 + y_2) \leq p(y_1 + y_2) = p(y_1 - x + x + y_2) \leq p(y_1 - x) + p(x + y_2),$$

or

$$f(y_1) - p(y_1 - x) \leq p(x + y_2) - f(y_2)$$

Since $y_1, y_2 \in \mathcal{M}$ are arbitrary, we have that

$$\sup\{f(y) - p(y - x) : y \in \mathcal{M}\} \leq \inf\{p(x + y) - f(y) : y \in \mathcal{M}\}$$

Let α be the number such that

$$\sup\{f(y) - p(y - x) : y \in \mathcal{M}\} \leq \alpha \leq \inf\{p(x + y) - f(y) : y \in \mathcal{M}\}$$

Thus,

$$\alpha \leq p(x + y) - f(y) \text{for } y \in \mathcal{M}$$

$$\alpha \geq f(y) - p(y - x) \text{for } y \in \mathcal{M}$$

We now define $g : \mathcal{M} + \mathbb{R}x \to \mathbb{R}$ by $g(y + \lambda x) = f(y) + \lambda\alpha$ for $y \in \mathcal{M}$.

If $\lambda > 0$ and $y \in \mathcal{M}$, then

$$g(y + \lambda x) = \lambda[f(y/\lambda) + \alpha] \leq \lambda[f(y/\lambda) + p(x + (y/\lambda)) - f(y/\lambda)] = p(y + \lambda x)$$

If $\lambda = -\mu < 0$,

$$g(y + \lambda x) = \mu[f(y/\mu) - \alpha] \leq \mu[f(y/\mu) - f(y/\mu) + p((y/\mu)) - x)] = p(y + \lambda x)$$

Thus, $g(z) \leq p(z)$ for all $z \in \mathcal{M} + \mathbb{R}x$ and by definition of g, g is linear and $g|\mathcal{M} = f$, as desired.

Since $x \in \mathcal{X}\backslash\mathcal{M}$ is arbitrary, we can applied the above reasoning to any linear extension F of f satisfying $F \leq p$ on its domain. We know that the domain of maximal linear extension of such linear extensions must be $\mathcal{X}$. By applying Zorn's Lemma here, we can find the desired F we want. □

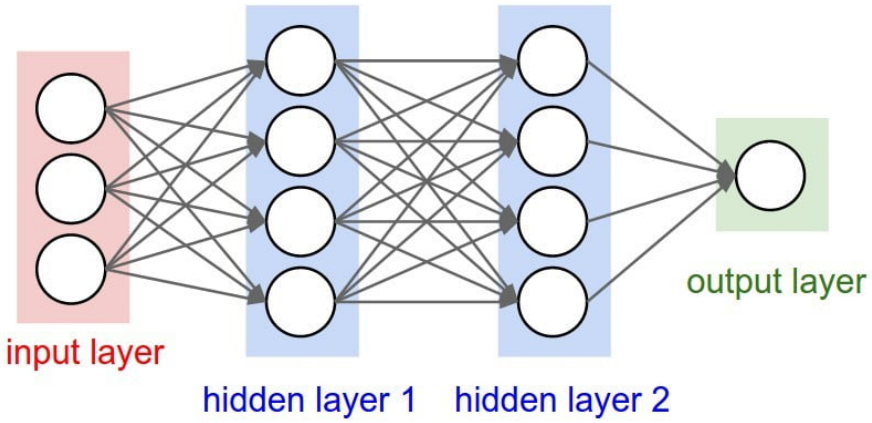

Figure 1: An example of an artificial neural network.

6 Universal Approximation Theorem

In the previous sections, we stated and proved the Riesz Representation Theorem and the Hahn-Banach Theorem. With both theorems in hand, we are now able to show the reason that artificial neural networks can approximate continuous functions to arbitrary precision by proving the Universal Approximation theorem.

6.1 Mathematical Representation of Artificial Neural Network

First, we want to give a mathematical representation of the one hidden layer neural network.

Before that, we need to have an basic understanding on what is an artificial neural network. It turns out that it is a learning algorithm that is vaguely inspired by neural networks. In examples of artificial neural network, you can find there are nodes which represents the "neurons" and they are grouped by layers and connected from each layer to the next layer. Each artificial neural network has an input layer, an output layer and some hidden layers, each layer has different number of nodes in them. By connecting each node from each layer to the nodes in the next layer, we can form a "network". It is fine that one doesn't understand the intuition of the neural network, you can basically view it as a map from $\mathbb{R}^N \to \mathbb{R}^n$ where N is the dimension of (number of nodes in) the input layer and n is the dimension of the output layer. For instance, in Figure 1, the input layer has dimension $\mathbb{R}^3$ and output layer has dimension $\mathbb{R}$ and this neural network has 2 hidden layers. For this thesis, we are only interested in artificial neural network with one hidden layer.

With an idea that an artificial neural network is a mapping, let's define this neural network with one hidden layer in a formal way.

Definition 6.1. $G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \Theta_j)$ where $x, y_j \in \mathbb{R}^n$ and $\alpha_j, \Theta \in \mathbb{R}$

Here, we regard the one hidden layer neural network as a function $G(x)$ which is a map from $\mathbb{R}^n$ to $\mathbb{R}$. N represents the number of nodes in the one hidden layer network. The input of the network is x. We call y_j weights and Θ bias which are numbers acting with the input on each neuron of the hidden layer, we called σ the activation function on each neuron. Then after applying weights of output of each neuron of the hidden layer, we sum the output to get the result of the neural network.

6.2 Universal Approximation Theorem

Since the universal approximation theorem is trying to show that the artificial neural networks with such form would be able to approximate continuous functions in arbitrary precision. We would like to let the set of our form of artificial neural networks be similar or close to the set of the continuous functions we are trying to approximate. In this section, after introducing some definitions that prepare us for the theorem, we are able to understand how this approximation works.

Definition 6.2. A function σ is discriminatory if for a measure $\mu \in M(I_n)$

$$\int_{I_n} \sigma(y^T x + \Theta) d\mu(x) = 0$$

for all $y \in \mathbb{R}^n$ and $\Theta \in \mathbb{R}$ implies that $\mu = 0$.

Theorem 6.3. *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \Theta_j)$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \text{ for all } x \in I_n$$

Proof. We prove by way of contradiction. Let S be the set of functions of the form $G(x)$. Assume that the closure of S is not all of $C(I_n)$. Then the closure of S, say R, is a closed proper subspace of $C(I_n)$.

By Hahn-Banach Theorem, there is a bounded linear functional on $C(I_n)$, call it L, with the property that $L(C(I_n)) \neq 0$ but $L(R) = 0$.

By the Riesz Representation Theorem, this bounded linear functional, L, is of the form

$$L(h) = \int_{I_n} h(x) d\mu(x)$$

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$. In particular, since $\sigma(y^T x + \Theta)$ is in R for all y and Θ, we must have that

$$\int_{I_n} \sigma(y^T x + \Theta) d\mu(x) = 0$$

for all y and Θ.

However, we assumed that σ was discriminatory so that this condition implies that $\mu = 0$ contradicting our assumption. Hence, the subspace S must be dense in $C(I_n)$. □

Lemma 6.4. *Any continuous sigmoidal function is discriminatory. (Cybenko [1])*

Theorem 6.5. *(Universal Approximation Theorem (Cybenko [1]))*

The functions of the form $G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \Theta_j)$ are dense in the space of continuous functions on the unit cube if σ is any continuous sigmoidal function.

Proof. Combine Theorem 1 and Lemma 1. □

References

[1] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[2] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.