BAIRE CATEGORIES, CONTINUOUS FUNCTIONS, AND APPROXIMATIONS

ANDREW CHEN

Abstract. This paper is an exploration into continuous functions and how they can be approximated. We first gain insight into the nature of continuous functions and demonstrate why most continuous functions are nowhere differentiable using the Baire Category Theorem. We then discuss two methods of approximating continuous functions: the first of which is using polynomials via the Weierstrass Approximation Theorem, and the second of which is using neural networks via the Universal Approximation Theorem.

Contents

1. Introduction

For many years, it was widely thought by mathematicians that most (if not all) continuous functions are differentiable, due to intuitive notions of continuity and differentiability. While calculus was invented in the late 1600s, the first known example of a non-differentiable continuous function was not discovered until around year 1830, by Bernard Bolzano, and the first published example was not presented until 1872, by Karl Weierstrass [9]. Due to later advances in the field of topology, we now know that *most* continuous functions are in fact nowhere differentiable. Section 2 will explore a proof of this, which utilizes an important theorem in topology called the Baire Category Theorem. We will then explore methods of approximating continuous functions, even those that are nowhere differentiable. Section 3 will explore a probabilistic proof of the Weierstrass Approximation Theorem, which will show that all continuous functions $f : [0,1] \to \mathbb{R}$ can be approximated arbitrarily well using polynomials. The fourth and final section will then explore the ability of machine learning to approximate continuous functions. The main result is the Universal Approximation Theorem, which states that a feedforward neural network used in machine learning can approximate continuous functions arbitrarily well.

2. Continuous Functions and Differentiability

The goal of this section is to use analysis to explore the nature of continuous functions. In particular, we will apply the Baire Category Theorem to the space of real-valued continuous functions $\mathcal{C}(K)$ on a compact set K to show that most continuous functions are nowhere differentiable. The definitions in this section can be found in most real analysis textbooks such as [8], and the specific proof presented here was inspired by [5].

2.1. **Introducing the Baire Category Theorem.** This section will assume that the reader has a general understanding of metric spaces and their usual topology. As a reminder, a metric space is a set X with some notion of distance between any two elements of the set, given by a distance function $d : X \times X \to [0, \infty)$ that is symmetric, satisfies the triangle inequality, and has the property that $d(x, y) = 0 \iff x = y$.

Definition 2.1. For an element p in a metric space X, and a real number $r > 0$, we define the **open ball** $B(p, r)$ as

$$B(p, r) = \{q \in X \mid d(p, q) < r\}.$$

Similarly, we define the **closed ball** $\overline{B}(p, r)$ as

$$\overline{B}(p, r) = \{q \in X \mid d(p, q) \leq r\}.$$

Definition 2.2. A set D is **dense** in a metric space X if and only if for every element $p \in X$ and $\epsilon > 0$, $D \cap B(p, \epsilon) \neq \varnothing$. Alternatively, D is dense in X iff its closure $\overline{D} = X$.

Example 2.3. The rational numbers $\mathbb{Q}$ are dense in the real numbers $\mathbb{R}$.

Proof. Between any two real numbers a and b with $a < b$, there exists a rational number q such that $a < q < b$. Thus the intersection of the rationals with any open ball in the reals will always be nonempty. □

Definition 2.4. A set E is **nowhere dense** in a metric space X if $\left(\overline{E}\right)^{\circ} = \varnothing$.

Example 2.5. The integers $\mathbb{Z}$ are nowhere dense in the real numbers $\mathbb{R}$.

Proof. $\mathbb{Z}$ has no limit points, so we know that $\overline{\mathbb{Z}} = \mathbb{Z}$. Then assume for contradiction that the interior of $\mathbb{Z}$ is not empty; in other words, assume there exists some nonempty open ball $U \subset \mathbb{R}$ such that $U \subset \mathbb{Z}$. Since $\mathbb{Z}$ is countable, U must also be countable. However, this is a contradiction because all open subsets of $\mathbb{R}$ are uncountable. Therefore U does not exist, so the interior of $\mathbb{Z}$ is empty. □

Remark 2.6. A set can be not dense, but also not nowhere dense. For example, consider the interval $(0, 1) \subset \mathbb{R}$. $\overline{(0, 1)} = [0, 1] \neq \mathbb{R}$, so $(0, 1)$ is not dense in $\mathbb{R}$. However, $[0, 1]^{\circ}$ is also nonempty in $\mathbb{R}$ since it contains $(0, 1)$, so the interval is also not nowhere dense in $\mathbb{R}$. We can see that nowhere dense is a much stronger property than simply being not dense; in a sense, it means a set is not dense everywhere.

Lemma 2.7. *The complement of a closed nowhere dense set is an open dense set.*

Proof. Let A be a closed nowhere dense set. $X \setminus (A^{\circ}) = \overline{X \setminus A}$, but A has empty interior, so $X = \overline{X \setminus A}$. Therefore $X \setminus A$ is open and dense. □

We are ready to state the major result of this section.

Theorem 2.8 (Baire Category Theorem). *If X is a nonempty complete metric space, and $(G_n)_{n=1}^{\infty}$ is a countable collection of dense open subsets of X, then $\bigcap_{n=1}^{\infty} G_n$ is dense in X.*

Proof. Let U be an arbitrary open ball in X. Since G_1 is dense in X, we know it has nonempty intersection with U. Pick $x_1 \in G_1 \cap U$ and $0 < r_1 < 1$ such that $\overline{B}(x_1, r_1) \subseteq G_1 \cap U$. This is possible since both

G_1 and U are open. Since G_2 is dense in X, we know that $G_2 \cap B(x_1, r_1)$ is nonempty. We repeat the same process, picking $x_2 \in G_2 \cap B(x_1, r_1)$ and $0 < r_2 < \frac{1}{2}$ such that $\overline{B}(x_2, r_2) \subseteq G_2 \cap B(x_1, r_1)$. Since every G_n is dense, we continue repeating the process to get a nested sequence of closed balls, such that $\overline{B}(x_n, r_n) \subseteq G_n \cap B(x_{n-1}, r_{n-1})$, and $0 < r_n < \frac{1}{n}$. Note that $x_n \in B(x_m, r_n)$ for all $n > m$, or in other words, $|x_m - x_n| < r_n < \frac{1}{n}$. Thus, for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$, and for $n, m \geq N$, $|x_m - x_n| < r_n < \epsilon$. Since X is a complete metric space, $x_n \to x \in X$ as $n \to \infty$, for some x. By definition of closure, we know that $x \in \overline{B}(x_n, r_n)$. Since our sequence of balls is nested, and $\overline{B}(x_n, r_n) \subseteq G_n$, we know that $x \in G_n$ for all n. Additionally, since $\overline{B}(x_1, r_1) \subseteq U$, we know that $x \in U$. Therefore, we have shown that $(\bigcap G_n) \cap U$ contains x and is thus nonempty. Since U is arbitrary, we have shown that $\bigcap G_n$ is dense in X. □

To understand why this is called the Baire Category Theorem, we introduce the following definition.

Definition 2.9. In a metric space X, a set $E \subseteq X$ is called **meager** if E can be written as a countable union of nowhere dense sets:

$$E = \bigcup_{n=1}^{\infty} E_n,$$

where each E_n is nowhere dense. Meager sets are also known as **first category**. E is called **non-meager**, or **second category**, if it is not first category. Two important properties of meager sets are that a subset of a meager set is also meager, and the union of meager sets is also meager.

Remark 2.10. The usage of the word "category" in this context is unrelated to category theory.

Corollary 2.11. *If X is a nonempty complete metric space, then X is non-meager in itself.*

Proof. Let $\{E_i\}$ be an arbitrary countable collection of closed nowhere dense subsets of X. Then by Lemma 2.7, $V_i = X \setminus E_i$ is an open dense subset of X for every i. By Theorem 2.8, $\bigcap_i V_i$ is dense in X and nonempty. Since $\bigcap_i V_i$ is disjoint with $\bigcup_i E_i$, we know that $X \neq \bigcup_i E_i$ since there are elements in $\bigcap_i V_i \subset X$ that are not in $\bigcup_i E_i$. If $\{E_i\}$ is instead an arbitrary countable collection of open nowhere dense subsets of X, then each E_i must be the empty set, and thus $\bigcup_i E_i = \varnothing \neq X$. Therefore, X cannot be written as a countable union of nowhere dense subsets. □

Example 2.12. $\mathbb{Z}$ and $\mathbb{Q}$ are meager in $\mathbb{R}$, but the set of irrationals $\mathbb{R} \setminus \mathbb{Q}$ is non-meager in $\mathbb{R}$.

Proof. $\mathbb{Z}$ and $\mathbb{Q}$ are both countable, so they can be expressed as the countable union of singleton sets. Singleton sets are nowhere dense in $\mathbb{R}$, so $\mathbb{Z}$ and $\mathbb{Q}$ are meager in $\mathbb{R}$. Assume for contradiction $\mathbb{R} \setminus \mathbb{Q}$ is also meager in $\mathbb{R}$. Then $\mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q}) = \mathbb{R}$ is meager in $\mathbb{R}$, but $\mathbb{R}$ is non-meager by Corollary 2.11, so our assumption is false and $\mathbb{R} \setminus \mathbb{Q}$ is non-meager. □

Remark 2.13. This example helps show why we can think of meager sets as being small relative to non-meager sets *in the same topological space*. However, it is important to note that not all non-meager sets are of the same "size", and that meagerness and non-meagerness have nothing to do with the "absolute" or "objective" size of a set. For example, $\mathbb{Z}$ is meager in $\mathbb{R}$, so it is small relative to $\mathbb{R}$. This matches our intuitive notion of their relative sizes, since $\mathbb{Z}$ is countable and $\mathbb{R}$ is uncountable. However, $\mathbb{Z}$ also forms a complete metric space, so by the Baire Category Theorem, it is non-meager in itself. This is because singleton sets in the discrete topology have nonempty interior, and thus are not nowhere dense.

2.2. **Applying the Baire Category Theorem to the Set of Continuous Functions.** We now apply the Baire Category Theorem to show that most continuous functions are nowhere differentiable. We can do this by showing the set of differentiable functions is meager in the set of continuous functions. Let $K = [0, 1] \subset \mathbb{R}$ and let $\mathcal{C}(K)$ be the set of all continuous functions $f : K \to \mathbb{R}$.

Definition 2.14. For $f \in \mathcal{C}(K)$ we define $\|f\|$, the **supremum norm** of f, by

$$\|f\|_{\sup} \coloneqq \sup\{|f(x)| : x \in K\}.$$

Remark 2.15. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions. Then (f_n) converges to f under $\|\cdot\|_{\sup}$ iff (f_n) converges to f uniformly.

Proposition 2.16 (Uniform Limit Theorem)**.** *Let (f_n) be a sequence of functions in $\mathcal{C}(K)$. If (f_n) converges uniformly to f, then $f \in \mathcal{C}(K)$.*

Proof. Let $\epsilon > 0$. By uniform convergence of (f_n), we know that there exists $N \in \mathbb{N}$ such that if $n \geq N$, then $|f_n(x) - f(x)| < \frac{\epsilon}{3}$ for all $x \in K$. Consider the case of $n = N$ and let $x_0 \in K$. Since f_N is continuous at x_0, we know there is $\delta > 0$ such that if $|x - x_0| < \delta$, then $|f_N(x) - f_N(x_0)| < \frac{\epsilon}{3}$. Then, using the triangle inequality and utilizing the two inequalities above, we have that

$$|f(x) - f(x_0)| \leq |f(x) - f_N(x)| + |f_N(x) - f_N(x_0)| + |f_N(x_0) - f(x_0)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}.$$

Thus, if $|x - x_0| < \delta$, we have that $|f(x) - f(x_0)| < \epsilon$, proving that f is continuous on K. □

Theorem 2.17. $\mathcal{C}(K)$ *with the supremum norm is a complete metric space.*

Proof. Let (f_n) be an arbitrary Cauchy sequence in $\mathcal{C}(K)$. Let $\epsilon > 0$, and since (f_n) is Cauchy with respect to $\|\cdot\|_{\sup}$, there exists $N_1 \in \mathbb{N}$ such that for $m, k \geq N_1$, $\|f_m - f_k\|_{\sup} < \epsilon$. By definition of supremum norm, this implies that

$$|f_m(x) - f_k(x)| < \epsilon \text{ for all } x \in K, \text{ and } m, k \geq N_1. \tag{2.18}$$

Now fix arbitrary $x \in K$, and consider the sequence of real numbers $(f_n(x))$. By Equation 2.18, there exists N_1 such that this sequence fulfills Cauchy's criterion in $\mathbb{R}$, so $(f_n(x))$ converges to some $p_x \in \mathbb{R}$. By definition of sequence convergence in $\mathbb{R}$, there exists $N_2 \in \mathbb{N}$ such that for all $p \geq N_2$, $|f_p(x) - p_x| < \epsilon$. Let us define $f : K \to \mathbb{R}$ as $f(x) = p_x$, and let $N = \max\{N_1, N_2\}$. Then if $n \geq N$, we have that $|f_n(x) - p_x| = |f_n(x) - f(x)| < \epsilon$ for all $x \in K$. We thus have that $\|f_n - f\|_{\sup} < \epsilon$ for all $n \geq N$, so (f_n) converges uniformly to f. By Proposition 2.16, we know that $f \in \mathcal{C}(K)$, so $\mathcal{C}(K)$ is complete. □

Corollary 2.19. $\mathcal{C}(K)$ *is non-meager in itself.*

Proof. Since $\mathcal{C}(K)$ is a complete metric space, it is non-meager in itself by Corollary 2.11. □

Now we will show that the set of differentiable functions is meager in $\mathcal{C}(K)$. Let D be the set of differentiable functions $D = \{f \in \mathcal{C}(K) \mid f \text{ is differentiable at } x \text{ for some } x \in K\}$, and let $A_{n,m}$ be the set $A_{n,m} = \left\{f \in \mathcal{C}(K) \mid \text{there is } x \in K \text{ such that } \left|\frac{f(t)-f(x)}{t-x}\right| \leq n \text{ if } 0 < |t - x| < \frac{1}{m}\right\}$.

Definition 2.20. A function $p : K \to \mathbb{R}$ is **piecewise-linear** if there is a partition $0 = a_0 < a_1 < \ldots < a_n = 1$ of $[0, 1]$, such that p is linear on the interval $[a_i, a_{i+1}]$, for $i = 0, \ldots, n$. Let $PL(K) \subseteq \mathcal{C}(K)$ be the set of piecewise linear continuous functions on K.

Theorem 2.21. *$PL(K)$ is dense in $\mathcal{C}(K)$.*

Proof. Suppose $f \in \mathcal{C}(K)$ and $\epsilon > 0$. We know that f is uniformly continuous on K since K is compact. Therefore, there is $\delta > 0$ such that for $x, y \in K$, if $|x - y| < \delta$, then $|f(x) - f(y)| < \frac{\epsilon}{2}$.

Let $0 = a_0 < a_1 < \ldots < a_n = 1$ be a partition of $[0, 1]$ such that $|a_{i+1} - a_i| < \delta$ for $i = 0, \ldots, n$. Let $p : K \to \mathbb{R}$ be the piecewise linear function such that $p(a_i) = f(a_i)$, p is linear on each $[a_i, a_{i+1}]$, and p is continuous. For all $x \in K$, there is an i such that $a_i \leq x \leq a_{i+1}$. Then since $|x - a_i| < \delta$, we know that

$$|p(x) - f(a_i)| = |p(x) - p(a_i)| \leq |p(a_{i+1}) - p(a_i)| = |f(a_{i+1}) - f(a_i)| < \frac{\epsilon}{2},$$

and since $|x - a_{i+1}| < \delta$, we know that

$$|f(x) - f(a_i)| < \frac{\epsilon}{2}.$$

Combining the two inequalities, we thus have that $|p(x) - f(x)| < \epsilon$ and therefore $p \in B(f, \epsilon)$. Since $PL(K)$ has nonempty intersection with arbitrary open ball $B(f, \epsilon)$, we have shown that $PL(K)$ is dense in $\mathcal{C}(K)$. □

Remark 2.22. This also shows we can approximate continuous functions arbitrarily well using piecewise linear functions.

Lemma 2.23. *$A_{n,m}$ is closed for all n, m.*

Proof. Since complete subspaces are closed, we will show that every $A_{n,m}$ is a complete subspace. Let (f_i) be a Cauchy sequence in an arbitrary $A_{n,m}$, and assume it converges to some $f \in \mathcal{C}(K)$. By construction of $A_{n,m}$, for each i we can find $x_i \in K$ such that

$$\left| \frac{f_i(t) - f_i(x_i)}{t - x_i} \right| \leq n \text{ for all } 0 < |t - x_i| < \frac{1}{m}.$$

We know that (x_i) is a bounded sequence since K is compact, so (x_i) has a convergent subsequence (x_{i_k}) by the Bolzano-Weierstrass Theorem. Replace the sequence (f_i) with the corresponding subsequence (f_{i_k}), and without loss of generality suppose (x_{i_k}) converges to x. Then if $0 < |x - t| < \frac{1}{m}$, we have

$$\left| \frac{f(t) - f(x)}{t - x} \right| = \lim_{k \to \infty} \left| \frac{f_k(t) - f_k(x_k)}{t - x_k} \right| \leq n.$$

Thus, $f \in A_{n,m}$, so $A_{n,m}$ is complete. □

Lemma 2.24. *Each $A_{n,m}$ is nowhere dense in $\mathcal{C}(K)$.*

Proof. Since each $A_{n,m}$ is already closed by the previous lemma, we only have to show that the interior of $A_{n,m}$ is empty; in other words, we show that $A_{n,m}$ does not contain an open ball. Let $f \in \mathcal{C}(K)$ be arbitrary and $\epsilon > 0$, so we can form the arbitrary open ball $B(f, \epsilon)$. We want to show $B(f, \epsilon) \not\subseteq A_{n,m}$, so we find $g \in B(f, \epsilon)$ such that $g \notin A_{n,m}$. By Theorem 2.21, we can find a piecewise linear function $p(x)$ such that $\|f - p\|_{\sup} < \frac{\epsilon}{2}$. Since p is piecewise linear, we know that p is differentiable at all but finitely many points, and we can pick a M such that $|p'(x)| \leq M$ for all x where p is differentiable.

Now pick a number k such that $k > \frac{2(M+n)}{\epsilon}$, and let $\varphi(x)$ be a continuous piecewise linear function such that $|\varphi(x)| \leq 1$ for all $x \in K$, and $\varphi'(x) = \pm k$ for all x where p is differentiable. Let $g(x) = p(x) + \frac{\epsilon}{2}\varphi(x)$. Since $\|f - p\|_{\sup} < \frac{\epsilon}{2}$ and $\|g - p\|_{\sup} < \frac{\epsilon}{2}$, we have that $\|f - g\|_{\sup} < \epsilon$, so $g \in B(f, \epsilon)$.

We claim that $g \notin A_{n,m}$. Let $x \in K$ be such that p and φ are differentiable at x. Then g is differentiable at x, and $|g'(x)| = |p'(x) \pm \frac{k\epsilon}{2}|$. Since $|p'(x)| \leq M$, and $|\frac{k\epsilon}{2}| > |M + n|$, it must be that $|g'(x)| > n$. Thus if $0 < |x - t| < \frac{1}{m}$, then

$$\left| \frac{g(t) - g(x)}{t - x} \right| > n,$$

so $g \notin A_{n,m}$. Thus, $B(f, \epsilon) \nsubseteq A_{n,m}$, and our ball was arbitrary, so $A_{n,m}$ is nowhere dense in $\mathcal{C}(K)$. □

Theorem 2.25. *D is meager in $\mathcal{C}(K)$.*

Proof. Let $A = \bigcup_{m=1}^{\infty} \bigcup_{n=1}^{\infty} A_{n,m}$. We first show that $D \subseteq A$. Take arbitrary $f \in \mathcal{C}(K)$ and suppose it is differentiable at x. Choose n such that $|f'(x)| < n$. Since $f'(x) = \lim_{x \to t} \frac{f(t)-f(x)}{t-x}$, we know there is $\delta > 0$ such that if $0 < |t - x| < \delta$, then $\left|\frac{f(t)-f(x)}{t-x}\right| < n$. Choose m such that $\frac{1}{m} < \delta$. Then $f \in A_{n,m}$. Because each $A_{n,m}$ is nowhere dense by Lemma 2.24, we know that A is meager in $\mathcal{C}(K)$. Since $D \subseteq A$, D is meager in $\mathcal{C}(K)$. □

Thus, we can see that the set of differentiable functions is small compared to the set of continuous functions.

3. Weierstrass Approximation Theorem

Now we can show that even though most functions in $\mathcal{C}(K)$ are nowhere differentiable, all functions in $\mathcal{C}(K)$ can be approximated arbitrarily well by a polynomial, which is everywhere differentiable. This is known as the Weierstrass Approximation Theorem, and several known proofs exist. Here, we will demonstrate one of the most common proofs, which is a constructive probabilistic proof first presented by Sergei Bernstein in 1912. This section assumes the reader is familiar with basic probability terminology, and the probability definitions presented in this section can be found in textbooks such as [7]. The method of proof is inspired by Bernstein's original proof, as transcribed in [4].

In this section, let Ω be an arbitrary sample space with a probability function $\Pr : \Omega \to [0, 1]$.

Definition 3.1. A **random variable** is a function $X : \Omega \to \mathbb{R}$. If $X(\Omega) \subset \mathbb{R}$ is countable, then X is **discrete**.

Definition 3.2. Let X be a discrete random variable. The **probability mass function** $p_X(a)$ of X is given by $p_X(a) := \Pr\{X = a\}$.

Proposition 3.3. *Let there be two mutually exclusive outcomes in a trial, success and failure. Let n trials occur, let x be the probability of success for any one trial, and let k be the number of successes. Then the probability mass function of a binomial random variable with parameters $(n, x; k)$, also known as a **binomial distribution**, is given by*

$$p(k) = \binom{n}{k} x^k (1 - x)^{n-k}.$$

*A random variable with the binomial distribution as its probability mass function is called a **binomial random variable**.* [4]

Definition 3.4. The **expected value** of a discrete random variable X is

$$E(X) := \sum_{\omega \in \Omega} X(\omega) \Pr(\omega).$$

Note that E is linear: for $a, b \in \mathbb{R}$ and random variables X, Y, we have that

$$E(aX + bY) = aE(X) + bE(Y).$$

Example 3.5. Consider an unfair coin with $\frac{3}{5}$ probability of heads (success) and $\frac{2}{5}$ probability of tails (fail) on a single flip, and let us conduct an experiment tossing the coin 3 times. Let H be the discrete binomial random variable (with parameters $(3, \frac{3}{5}; k)$) representing the number of heads. Then H can take on 4 possible values, and we can calculate the probability of each possible value using the binomial distribution:

$$H = \begin{cases} 0 & p_H(0) = \binom{3}{0}\frac{3}{5}^0\frac{2}{5}^3 = \frac{2}{5}^3 = \frac{8}{125} \\ 1 & p_H(1) = \binom{3}{1}\frac{3}{5}^1\frac{2}{5}^2 = 3\cdot\frac{3}{5}\cdot\frac{2}{5}^2 = \frac{36}{125} \\ 2 & p_H(2) = \binom{3}{2}\frac{3}{5}^2\frac{2}{5}^1 = 3\cdot\frac{3}{5}^2\cdot\frac{2}{5} = \frac{54}{125} \\ 3 & p_H(3) = \binom{3}{3}\frac{3}{5}^3\frac{2}{5}^0 = \frac{3}{5}^3 = \frac{27}{125} \end{cases}$$

We can also calculate the expected value of H by applying the definition:

$$E(H) = 0\cdot\frac{8}{125} + 1\cdot\frac{36}{125} + 2\cdot\frac{54}{125} + 3\cdot\frac{27}{125} = \frac{225}{125} = 1.8.$$

Definition 3.6. Let X be a discrete random variable with finite mean μ. Then the **variance** of X, which is also denoted by $\mathrm{Var}(X)$ or σ_X^2, is defined as $E((X-\mu)^2)$. The square root of the variance, denoted by σ_X, is called the **standard deviation**.

Proposition 3.7. *The mean of a binomial random variable is nx and the standard deviation of a binomial random variable is $\sqrt{nx(1-x)}$.*

Proof. See [7]. □

Theorem 3.8 (Markov's Inequality)**.** *Let $X \geq 0$ be a random variable with positive value, and let $\alpha > 0$ be arbitrary. Then $\Pr(X \geq \alpha) \leq \frac{1}{\alpha}E(X)$.*

Proof. $E(X) = \sum_{k=1}^{\infty}\omega_k \Pr(X = \omega_k) \geq \sum_{\omega_k \geq \alpha}\omega_k \Pr(X = \omega_k) \geq \alpha \sum_{\omega_k \geq \alpha}\Pr(X = \omega_k) = \alpha\Pr(X \geq \alpha)$. □

Theorem 3.9 (Chebyshev's Inequality)**.** *If X is a discrete random variable with with finite mean μ and standard deviation σ_X, then for any $k > 0$,*

$$\Pr(|X-\mu| \geq \sigma_X k) \leq \tfrac{1}{k^2}.$$

Proof. Since $(X-\mu)^2$ is a non-negative random variable, we know from Markov's inequality (with $\alpha = k^2\sigma_X^2$) that

$$\Pr((X-\mu)^2 \geq \sigma_X^2 k^2) \leq \frac{E\big((X-\mu)^2\big)}{\sigma_X^2 k^2}.$$

Since $(X-\mu)^2 \geq \sigma_X^2 k^2$ is equivalent to $|X-\mu| \geq \sigma_X k$ and $E\big((X-\mu)^2\big) = \sigma_X^2$ by definition, we have that

$$\Pr(|X-\mu| \geq \sigma_X k) \leq \tfrac{1}{k^2}.$$

□

Definition 3.10. A **Bernstein polynomial** of degree n for a function f is defined as

$$B_n^f(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}, \text{ for } n \in \mathbb{N}.$$

Theorem 3.11 (Weierstrass Approximation Theorem)**.** *If f is a continuous bounded function on the interval $[0,1]$, then its sequence of Bernstein polynomials $(B_n^f)_{n\in\mathbb{N}}$ converges uniformly to f.*

Proof. Let $\epsilon > 0$ and $f \in \mathcal{C}(K)$ be arbitrary. We want to show that there exists some $N \in \mathbb{N}$ such that $\|f - B_n^f\|_{\sup} < \epsilon$ for all $n \geq N$. First we show that $B_n^f(x)$ is equivalent to the expected value of a random variable. For $x \in K$ and $n \in \mathbb{N}$, let $(F_{n,x})$ be a sequence of binomial random variables with parameters

$(n, x; k)$. We then define the binomial random variable $F_{n,x}(k) \coloneqq f\left(\frac{k}{n}\right)$, such that if k successes occur and $F_{n,x} = k$, then $F_{n,x}(k) = f(\frac{k}{n})$. The expected value of $F_{n,x}(k)$ is

$$E(F_{n,x}(k)) = \sum_{k=0}^{n} F_{n,x}(k)P(k) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k} = B_n^f(x). \tag{3.12}$$

We then show that $B_n^f(x)$ converges uniformly to f on K. f is bounded on K because it's continuous and K is compact, so we can pick M such that $|f(x)| \leq M$ for all $x \in K$. Then we also know that $|f(x) - f(y)| \leq 2M$ for $x, y \in K$. Additionally, we know that f is uniformly continuous on K by the Heine-Cantor Theorem, so there exists $\delta > 0$ such that if $|x - y| < \delta$, then $|f(x) - f(y)| < \frac{\epsilon}{2}$. Let us pick $j \in \mathbb{N}$ such that $\frac{2M}{j^2} < \frac{\epsilon}{2}$, and then pick $N \in \mathbb{N}$ such that $\frac{j}{2\sqrt{N}} < \delta$. Now, let $n \geq N$, and using the linearity of expected value, we have

$$\begin{aligned} \Big|f(x) - E(F_{n,x}(k))\Big| = \Big|E\big(f(x) - F_{n,x}(k)\big)\Big| &= \left|\sum_{k=0}^{n}\left(f(x) - f\left(\frac{k}{n}\right)\right)\binom{n}{k}x^k(1-x)^{n-k}\right| \\ &\leq \sum_{k=0}^{n}\left|\left(f(x) - f\left(\frac{k}{n}\right)\right)\binom{n}{k}x^k(1-x)^{n-k}\right|. \end{aligned} \tag{3.13}$$

Let us break this sum up into two sections, the first where $|x - \frac{k}{n}| < \frac{j}{2\sqrt{n}}$ and $k \leq n$, and the second (its complement) where $|x - \frac{k}{n}| \geq \frac{j}{2\sqrt{n}}$ and $k \leq n$:

$$= \sum_{|x-\frac{k}{n}|<\frac{j}{2\sqrt{n}}}\left|\left(f(x) - f\left(\frac{k}{n}\right)\right)\binom{n}{k}x^k(1-x)^{n-k}\right| + \sum_{|x-\frac{k}{n}|\geq\frac{j}{2\sqrt{n}}}\left|\left(f(x) - f\left(\frac{k}{n}\right)\right)\binom{n}{k}x^k(1-x)^{n-k}\right|. \tag{3.14}$$

Since $|x - \frac{k}{n}| < \frac{j}{2\sqrt{n}} \leq \frac{j}{2\sqrt{N}} < \delta$, we apply uniform continuity:

$$< \frac{\epsilon}{2}\sum_{|x-\frac{k}{n}|<\frac{j}{2\sqrt{n}}}\left|\binom{n}{k}x^k(1-x)^{n-k}\right| + \sum_{|x-\frac{k}{n}|\geq\frac{j}{2\sqrt{n}}}\left|\left(f(x) - f\left(\frac{k}{n}\right)\right)\binom{n}{k}x^k(1-x)^{n-k}\right|. \tag{3.15}$$

We then utilize the fact that

$$\sum_{|x-\frac{k}{n}|<\frac{j}{2\sqrt{n}}}\left|\binom{n}{k}x^k(1-x)^{n-k}\right| \leq 1, \text{ which we know because } \sum_{k=0}^{n}\left|\binom{n}{k}x^k(1-x)^{n-k}\right| = 1.$$

We can also use the fact that we picked M such that $|f(x) - f(\frac{k}{n})| \leq 2M$, and continue Inequality 3.15 as follows:

$$\leq \frac{\epsilon}{2} + 2M\sum_{|x-\frac{k}{n}|\geq\frac{j}{2\sqrt{n}}}\left|\binom{n}{k}x^k(1-x)^{n-k}\right|. \tag{3.16}$$

Since $x \in K = [0, 1]$, we know that $x(1-x) \leq \frac{1}{2}$, giving us that $|x - \frac{k}{n}| \geq \frac{j}{2\sqrt{n}} \geq k\sqrt{\frac{x(1-x)}{n}}$. We know that $|x - \frac{k}{n}| \geq k\sqrt{\frac{x(1-x)}{n}} \iff |nx - k| \geq k\sqrt{nx(1-x)}$, so we can apply Theorem 3.9 in the following manner:

$$2M\sum_{|x-\frac{k}{n}|\geq\frac{j}{2\sqrt{n}}}\left|\binom{n}{k}x^k(1-x)^{n-k}\right| \leq 2M\sum_{|nx-k|\geq k\sqrt{x(1-x)}}\left|\binom{n}{k}x^k(1-x)^{n-k}\right| \leq \frac{2M}{j^2} < \frac{\epsilon}{2}. \tag{3.17}$$

Substitute the result of Inequality 3.17 into Inequality 3.16 and we arrive at what we want to show: beginning at Equation 3.13, we have demonstrated that there exists N such that if $n \geq N$, then

$$|f(x) - E(F_{n,x}(k))| = |f(x) - B_n^f(x)| < \epsilon.$$

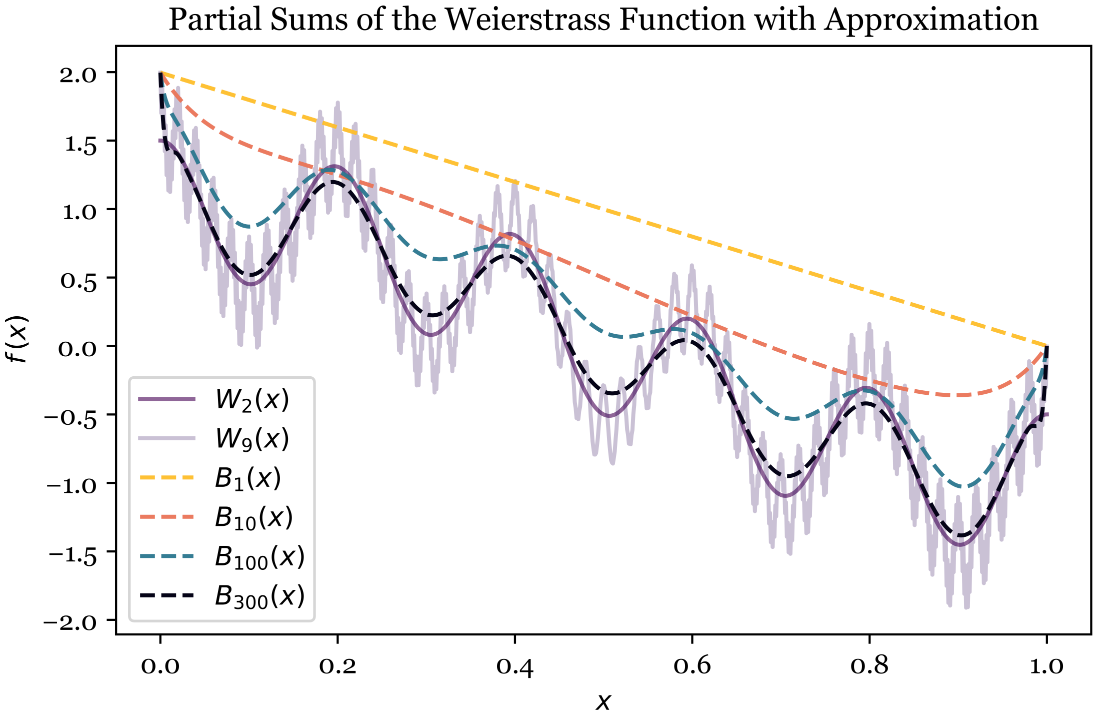

FIGURE 1. The Weierstrass Function with approximation using the partial sums of its Bernstein Polynomial.

Since our x was arbitrary, we have shown that $\|f - B_n^f\|_{\sup} < \epsilon$ for all $n \geq N$, and thus B_n^f converges uniformly to f. □

Figure 1 is a demonstration of the Weierstrass Approximation Theorem using Bernstein polynomials. The function $W(x)$ being approximated is the Weierstrass function, which Karl Weierstrass presented in 1872 as an example of a real-valued function that is continuous but nowhere differentiable (the concept of such functions were radical at the time). The function is expressed as a Fourier series, which has partial sums given by

$$W_n(x) = \sum_{k=1}^{n} a^k \cos(b^k \pi x),$$

where $0 < a < 1$, b is a positive odd integer, and $ab > 1 + \frac{3}{2}\pi$. For the approximation shown, I picked $a = \frac{1}{2}$ and $b = 10$. The degree m Bernstein polynomial approximation of $W(x)$ is given by

$$B_m(x) = \sum_{j=1}^{m} \left(\sum_{k=1}^{9} \frac{1}{2^k} \cos\left(10^k \pi \frac{j}{m}\right)\right) \binom{m}{j} x^j (1-x)^{m-j}.$$

Although the Bernstein polynomial can approximate any function in $\mathcal{C}(K)$, we can see that a "good" approximation may require a polynomial of very high degree.

4. The Universal Approximation Theorem

Thanks to the Weierstrass Approximation Theorem, we have shown it is possible to approximate any given continuous real-valued function over $K \subset \mathbb{R}$ using polynomials. This final section of the paper will

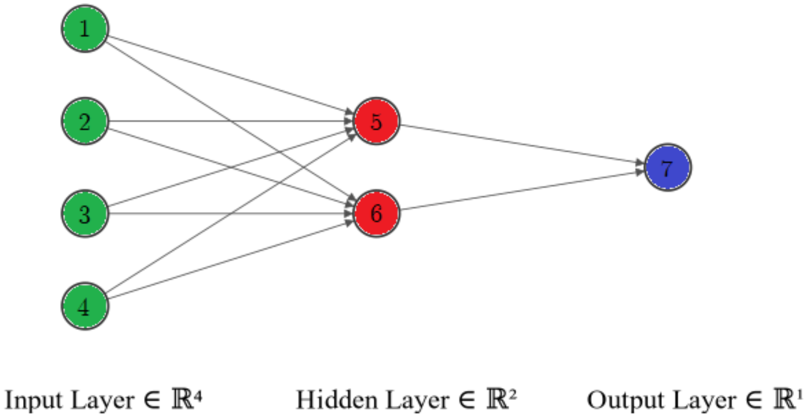

Figure 2. A diagram of an extremely simple feed forward neural network with 4 inputs, 1 output, and a single hidden layer. Made using [3] and modified by author.

explore another method of approximating continuous real-valued functions, which utilizes neural networks and machine learning. Specifically, we will discuss feedforward neural networks, and prove the Universal Approximation Theorem for continuous sigmoid activation functions. The definitions and theorems utilized in the proof of the Universal Approximation Theorem can be found in functional analysis textbooks such as [8], and the proof mimics George Cybenko's original proof found in [1] and [2]. First, we must understand the general process of how neural networks function.

Figure 2 diagrams an example of a very simple feedforward neural network. We refer to each circle as a **node**, and each arrow represents a connection between two nodes. In theory, there is no limit to the number of nodes in each layer, or to the number of layers in the network. In practice, having more layers and nodes require more computational power. The network is called feedforward because each node feeds an input "forward" into the next node, and no recurrence is present; there are many other types of network structures used today, but feedfoward neural networks provide an easy introductory example for us to understand.

To give an example of the utility of such a network, let us discuss the concrete example of image recognition, a typical task that can utilize machine learning. Imagine our neural network is given a 10×10 pixel greyscale image with a single digit drawn on it, and we want the network to recognize the digit displayed on the image. One way to set up the network is to have 100 input nodes, with each input node taking on the greyscale value (perhaps between 0 and 1, with 0 being white and 1 being black) of a single pixel in the image. We can give the network some number of hidden layers, which the network will use to learn. We could then set up 10 nodes in the output layer of this neural network, where each node corresponds to a digit from 0 to 9. Under this setup, if we input pixel values corresponding to an image of a 6, then we ultimately hope to see a value of 1 for the output node representing the number 6, and a 0 for the remaining output nodes. In this example, the purpose of our neural network is to approximate an unknown function $F : K^{100} \to K^{10}$ that takes in the 100 pixel input values, and outputs the vector representing the digit that is on the image with perfect accuracy. In general, finding this exact F is extremely challenging and usually impossible in practice. However, even if we don't know how to explicitly calculate F, the Universal Approximation Theorem will show us that we can use a neural network to approximate F arbitrarily well. For our purposes, we will make the assumption that F is a continuous function. Even if F is discontinuous, a continuous approximation is

Figure 3. Examples of training images for a number recognition neural network, from [6].

often good enough [6].

To make a reasonably good approximation of F, the network first needs to undergo a machine learning process. Initially, the network will be highly inaccurate, and will output almost random results. To improve the accuracy of the network, we train it on a dataset where the inputs and correct outputs are known. During the training process, the network will keep comparing its own answers with a set of correct answers, and find a way to minimize the ℓ^2-error between its own outputs and the correct outputs. This is done by adjusting the value of the nodes, and utilizes techniques known as gradient descent and backpropagation. The exact details of these techniques won't be presented here, since it's not directly relevant to the mathematics of the Universal Approximation Theorem. However, interested readers can learn more about these methods, and the process of machine learning in general, in [6].

We will move on and discuss how the value of each node is calculated, which will directly lead into the Universal Approximation Theorem.

Definition 4.1. Let j be the index of the input layer, and let $j > 1$ be the index of a non-input layer. Let there be p nodes in layer j, and let $n_{j,k}$ be the value of the kth node in the jth layer. Then, the value of the rth node in the $(j+1)$th layer is

$$n_{j+1,r} = f\bigg(\sum_{k=1}^{p}(w_{j,k}^{j+1,r} n_{j,k}) + \theta_{j+1,r}\bigg). \tag{4.2}$$

In the above equation, $f : \mathbb{R} \to \mathbb{R}$ is known as an **activation function**, and $\theta_{j+1,r}$ is known as the **bias**. Note also that we can represent the sum above as the dot product of two vectors in $\mathbb{R}^p$. In that case, let

$x_{j+1,r} \in \mathbb{R}^p$ be the vector denoting the values of the nodes in the previous layer:

$$x_{j+1,r} = (n_{j,1}, n_{j,2}, \ldots, n_{j,p}).$$

Then, let $y_{j+1,r} \in \mathbb{R}^p$ be the vector denoting the weight corresponding to how much each node in the previous layer affects the current node:

$$y_{j+1,r} = \left(w_{j,1}^{j+1,r}, w_{j,2}^{j+1,r}, \ldots, w_{j,p}^{j+1,r}\right).$$

Therefore, Equation 4.2 can be re-written as

$$n_{j+1,r} = f(y \cdot x + \theta_{j+1,r}). \tag{4.3}$$

Example 4.4. Consider the neural network in Figure 2 with clearly labeled nodes. Let f be the activation function for the network. For convenience let us assume that the green input nodes have a value equal to their label, e.g. node 1 has value 1. Then the value of nodes 5 and 6, the first and second nodes in the second layer, respectively, are given by

$$n_{2,1} = f\big(w_{1,1}^{2,1} + 2w_{1,2}^{2,1} + 3w_{1,3}^{2,1} + 4w_{1,4}^{2,1} + \theta_{2,1}\big), \tag{4.5}$$

$$n_{2,2} = f\big(w_{1,1}^{2,2} + 2w_{1,2}^{2,2} + 3w_{1,3}^{2,2} + 4w_{1,4}^{2,2} + \theta_{2,2}\big). \tag{4.6}$$

Similarly, the value of node 7, the output node, is given by

$$n_{3,1} = f\big(w_{2,1}^{3,1} n_{2,1} + w_{2,2}^{3,1} n_{2,2} + \theta_{3,1}\big).$$

Definition 4.7. For an activation function $f : \mathbb{R} \to \mathbb{R}$, we define $\Sigma_n f$, the set of all functions that can be calculated by a neural network with a single hidden layer with p nodes and activation function f, as

$$\Sigma_n f \coloneqq \operatorname{Span}_{\mathbb{R}}\{f(y \cdot x + \theta) \mid y \in \mathbb{R}^p, \theta \in \mathbb{R}\}.$$

There are many different activation functions used in machine learning, but we will introduce now one of the most commonly used functions, which is also the specific example proven to be a universal approximator by Cybenko.

Definition 4.8. A bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ is called a **sigmoid function** if

$$\lim_{x \to -\infty} f(x) = 0 \text{ and } \lim_{x \to \infty} f(x) = 1.$$

A common continuous sigmoid activation function is the **logistic function**, which is given in its most basic form as

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

We now turn our attention to the Universal Approximation Theorem. To help prove the theorem, we first introduce a corollary of the Hahn-Banach Theorem, the Riesz Representation Theorem, and some necessary definitions.

Corollary 4.9. *Let V be a normed vector space over $\mathbb{R}$ and $U \subset V$ be a linear subspace such that $\overline{U} \neq V$. Then there exists a continuous linear functional $F : V \to \mathbb{R}$ with $F(x) = 0$ for any $x \in U$, and $F \not\equiv 0$.*

Proof. See [2]. □

Theorem 4.10 (Riesz Representation Theorem)**.** *Let Ω be a subset of $\mathbb{R}^n$ and $F : \mathcal{C}(\Omega) \to \mathbb{R}$ be a linear functional. Then there exists a signed Borel measure ν on Ω such that for any $f \in \mathcal{C}(\Omega)$, we have that*

$$F[f] = \int_\Omega f(x) d\nu(x).$$

Proof. See [8]. □

Definition 4.11. Let $J \subset \mathbb{R}^n$. Then $L^\infty(J) \coloneqq L^\infty(J, \mu)$ is the class of all functions on J that are bounded μ-almost everywhere.

Definition 4.12. Let Ω be a metric space. A neural network with activation function $f : \mathbb{R} \to \mathbb{R}$ is a **universal approximator** on Ω if $\Sigma_n f$ is dense in $\mathcal{C}(\Omega)$.

Let $K^n \subset \mathbb{R}^n$ be the closed rectangle $[0, 1] \times [0, 1] \ldots \times [0, 1] \subset \mathbb{R}^n$. We now define a class of functions that we will show give many examples of universal approximators over $\mathcal{C}(K^n)$.

Definition 4.13. Let n be a natural number. An activation function $f : \mathbb{R} \to \mathbb{R}$ is ***n*-discriminatory** if the only signed Borel measure ν such that $\int f(y \cdot x + \theta) d\nu(x) = 0$ for all $y \in \mathbb{R}^n, \theta \in \mathbb{R}$ is the zero measure. An activation function is **discriminatory** if it is n-discriminatory for all n.

Theorem 4.14. *If f is a continuous discriminatory function, then a neural network with f as the activation function is a universal approximator.*

Proof. Assume for contradiction that $\Sigma_n f$ is not dense in $\mathcal{C}(K^n)$. Then $\overline{\Sigma_n f} \neq \mathcal{C}(K^n)$, so by Corollary 4.9 there exists some continuous linear functional $G : \mathcal{C}(K^n) \to \mathbb{R}$ such that $G \neq 0$ but $G[w] = 0$ for any $w \in \overline{\Sigma_n f}$. Additionally, by Theorem 4.10 we know that there exists a Borel measure ν such that

$$G(w) = \int_{K^n} w(x) d\nu(x) \text{ for all } w \in \mathcal{C}(K^n).$$

For any $\theta \in \mathbb{R}$ and $y \in \mathbb{R}^n$, we know that $f(x \cdot y + \theta) \in \overline{\Sigma_n f}$ by construction, and we know from above that $G[f] = \int_{K^n} f(x \cdot y + \theta) d\nu(x) = 0$. Since f is discriminatory, this means that $\nu = 0$, and thus we have that $G[w] = \int_{K^n} w(x) d\nu(x) = 0$ for any $w \in \mathcal{C}(K^n)$. However, this contradicts Corollary 4.9 since $G \not\equiv 0$ for all $w \in \mathcal{C}(K^n)$, so the opposite of our assumption is true, and $\Sigma_n f$ is dense in $\mathcal{C}(K^n)$. □

We are now ready to prove the Universal Approximation Theorem.

Theorem 4.15. *All bounded, measurable sigmoid functions are discriminatory.*

Proof. Let f be a bounded and measurable sigmoidal function, and let $\nu \in M(K^n)$ be an arbitrary measure. We assume

$$\int f(y \cdot x + \theta) d\nu(x) = 0 \text{ for all } y \in \mathbb{R}^n, \theta \in \mathbb{R},$$

and show that this implies $\nu = 0$. To do so, let $\lambda, \psi \in \mathbb{R}$ and let us construct the function γ:

$$\gamma(x) = \lim_{\lambda \to \infty} f(\lambda(y \cdot x + \theta) + \psi) = \left\{ \begin{array}{ll} 1 & \text{if } y \cdot x + \theta > 0 \\ 0 & \text{if } y \cdot x + \theta < 0 \\ f(\psi) & \text{if } y \cdot x + \theta > 0 \end{array} \right\} . \tag{4.16}$$

By the Dominated Convergence Theorem, we have that

$$\int_{K^n} \gamma(x) d\nu(x) = \lim_{\lambda \to \infty} \int_{K^n} f(\lambda(y \cdot x + \theta) + \psi) d\nu(x) = 0.$$

We apply the piece-wise function definition to evaluate this integral. Let $P_{y,\theta}$ be the hyperplane defined by $\{x \in K^n \mid y \cdot x + \theta = 0\}$, let $H^+_{y,\theta}$ be the open positive half-space defined by $\{x \in K^n \mid y \cdot x + \theta > 0\}$, and let $H^-_{y,\theta}$ be the open negative half-space defined by $\{x \in K^n \mid y \cdot x + \theta < 0\}$. Then, Equation 4.16 gives us that

$$\begin{aligned}\lim_{\lambda \to \infty} \int_{K^n} f(\lambda(y \cdot x + \theta) + \psi) d\nu(x) &= \int_{H^+_{y,\theta}} 1 \, d\nu(x) + \int_{H^-_{y,\theta}} 0 \, d\nu(x) + \int_{P_{y,\theta}} f(\psi) d\nu(x) \\ &= \nu(H^+_{y,\theta}) + f(\psi)\nu(P_{y,\theta}) = 0.\end{aligned}$$

This is true for all ψ, θ, y. Now fix y, let h be a bounded measurable function, and let $J = y \cdot K^n \subset \mathbb{R}$. Let $F : L^\infty(J) \to \mathbb{R}$ be the linear functional defined as

$$F[h] = \int_{K^n} h(y \cdot x) d\nu(x),$$

and note that F is a bounded, well-defined functional on $L^\infty(J)$, since ν is a finite signed measure.

Now, specify h to be the indicator function of the interval $[\theta, \infty)$, such that $h(u) = 1$ if $u \geq \theta$ and $h(u) = 0$ if $u < \theta$. Then we have that

$$F[h] = \int_{K^n} h(y \cdot x) d\nu(x) = \nu(H^+_{y,-\theta}) + \nu(P_{y,-\theta}) = 0.$$

By linearity, we have that $F[h] = 0$ for the indicator function h of any interval, and it follows that $F(h) = 0$ for any simple function[1] h. Since simple functions are dense in $L^\infty(J)$, it must be that $F = 0$. Since $S(x) = \sin(y \cdot x)$ and $C(x) = \cos(y \cdot x)$ are elements of $L^\infty(J)$, we can use

$$F[C] + iF[S] = \int_{K^n} (\cos(y \cdot x) + i \sin(y \cdot x)) d\nu(x) = \int_{K^n} e^{iy \cdot x} d\nu(x) = 0.$$

This is true for all $y \in \mathbb{R}^n$, so the Fourier transform of ν is 0. We have shown that $\nu = 0$, so our f must have been discriminatory. □

Corollary 4.17 (Cybenko's Universal Approximation Theorem)**.** *Neural networks with continuous sigmoid activation functions are universal approximators.*

Proof. Using the fact that bounded, measurable sigmoid functions are continuous, we combine Theorems 4.14 and 4.15 to arrive at our result. □

Thus, we can see that even though most continuous functions are nowhere differentiable, they can still be approximated arbitrarily well using piecewise linear functions, polynomials or neural networks.

Acknowledgments

I would first and foremost like to thank my mentor, Calder Sheagren, for providing guidance and advice at every stage of this paper. I would then like to thank Peter Huxford for providing feedback on a draft of this paper. I also want to thank Professor Peter May for hosting the REU and creating this wonderful opportunity. Finally, I would like to thank the authors I've cited in my references below, since their work has taught me a lot throughout this research process.

[1] A simple function is a linear combination of indicator functions on different intervals.

References

[1] George Cybenko, *Approximation of superpositions by a sigmoidal function*, Mathematics of Control, Signals, and Systems (1989).

[2] Leonardo F. Guilhoto, *An overview of artificial neural networks for mathematicians* (2018).

[3] Alex Lenail, *Nn-svg*.

[4] Kenneth M. Levasseur, *A probabilistic proof of the weierstrass approximation theorem*, The American Mathematical Monthly **91** (1984), no. 4, 249–250.

[5] David Marker, *Most continuous functions are nowhere differentiable*, 2004.

[6] Michael A. Nielsen, *Neural networks and deep learning*, Determination Press, 2015.

[7] Sheldon M. Ross, *A first course in probability*, Fifth, Prentice Hall, 1997.

[8] Walter Rudin, *Real and complex analysis*, Third, McGraw-Hill, 1987.

[9] Johan Thim, *Continuous nowhere differentiable functions*, 2003.