THE STONE-WEIERSTRASS THEOREM AND ITS APPLICATIONS

VINAMRA DHOOT

Abstract. This paper tries to create an intuition behind the proof of the fundamental theorem–Weierstrass approximation theorem–which will be shown using Bernstein polynomials and its generalization beyond polynomials–the Stone-Weierstrass theorem. The paper will look at a few applications of the theorem in the Universal Approximation Theorem and abstract algebra.

1. Introduction

The history of the Stone-Weierstrass theorem, and how it evolved to be so, is one that is half a century long. Its roots trace back to the mathematician Karl Weierstrass.

Karl Weierstrass was a German mathematician who is often referred to as the "father of modern analysis". He had not started mathematics until he was in his late 30s or early 40s! In 1885, when Karl Weierstrass was 70, he proved one of the most important results, in approximation theory and real analysis: any function in the closed interval $[a, b]$ can be uniformly approximated by polynomials.

Today, there exist multiple different proofs of the Weierstrass approximation theorem, which vary from each other, but the most notable was by the Ukrainian mathematician, Sergei Bernstein. He provided an alternative constructive using what we now refer to as Bernstein polynomials. This offered mathematicians a method to compute the Weierstrass approximation theorem explicitly.

Before moving further, it is crucial to note that the Weierstrass approximation theorem and Taylor's theorem are different, even though they may seem similar and turns out to be a common misconception. So, first, what does Taylor's theorem tell us? Taylor's theorem gives us a local approximation for any function k times differentiable around a given point a by a polynomial of degree k, which is known as the k-th-order Taylor Polynomial.

Although the two may sound very similar, they are very different from each other. In Taylor's theorem, while we may be able to make the polynomials very close to the actual function on a point, we would require function to be highly differentiable–which means that the function has a derivative at each point in its domain–which is a very small subclass of functions [Joh21].

Nevertheless, the Weierstrass approximation theorem does not require differentiability at all. Moreover, Taylor's theorem allows for local approximations (approximations around a point) to take place, however, the Weierstrass approximation theorem guarantees for global approximations (approximations in a closed interval, so all points in the interval $[a, b]$ can be approximated within a distance ϵ). Lastly, Taylor's theorem controls errors at a singular

Date: July 11, 2025.

point using derivatives and distance, which makes it accurate for smooth functions. However, the Weierstrass approximation theorem guarantees that all functions, even those that are non-differentiable (like $|x|$), can be approximated as closely as possible using a set of polynomials.

However, the original theorem only applied to continuous functions in compact intervals in $\mathbb{R}$, and approximating them using a set of polynomials. Naturally, mathematicians expanded the question and began asking whether more general functions or topological spaces can be used to prove the same.

This was done in 1937 by Marshall H. Stone, an American mathematician, who published his generalization of Weierstrass's result by building upon work in topology. Stone considered the space $C(X)$ of real-valued continuous functions on a compact Hausdorff space X under the uniform norm. The result now called the Stone-Weierstrass theorem can be summarized as follows:

Let $\mathcal{A}$ be a subalgebra of $C(X, \mathbb{R})$, where X is a compact Hausdorff space. If $\mathcal{A}$ contains the points and separates the points of X, we can prove that $\mathcal{A}$ is uniformly dense in $C(X, \mathbb{R})$.

Stone's result shifted the focus from just approximating using polynomials to using elements of arbitrary subalgebras. The study was influential in multiple mathematical contexts but also extended to other subjects.

In this paper, we will look at the proof of the Weierstrass approximation theorem using Bernstein polynomials followed by the proof of the Stone-Weierstrass Theorem, and finally look at a few applications of the Stone-Weierstrass Theorem in the Universal Approximation Theorem and abstract algebra.

2. Notations and Key Definitions

Before looking at the Weierstrass approximation theorem, below are a few key definitions and notations that are vital for understanding the theorem.

Definition 2.1. Fields
A field is a set that contains two operations: addition and multiplication. It satisfies all of the ring axioms (from definition 5.7) but also includes multiplicative inverses (which means for any $a \in R$, there exists a^{-1} such that $a \cdot a^{-1} = 1$).

Note: We denote $\mathbb{R}$ to represent real numbers and $\mathbb{C}$ to represent complex numbers.

Example.
The real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$ are examples of a field because they are closed under addition, multiplication, and their inverses (which means that any of these operations would give a result in the same field).

Definition 2.2. Metric Space
A metric space is a set S with a function $d : S \times S \to \mathbb{R}^+$ if any of the following conditions are met:

(1) $d(a, b) \geq 0$. $d(a, b) = 0 \Leftrightarrow a = b$.

(2) $d(a, b) = d(b, a)$.

(3) *The Triangle Inequality:* $d(a,c) \leq d(a,b) + d(b,c)$.

Example. (Metric Space in Cartesian Coordinates) A simple example of this would be using Cartesian coordinates. If we have X that includes all points in the form (x, y) in the Cartesian coordinate system, then we have $d(a, b)$ that distance between two points in X, which is a valid metric.

Definition 2.3. Topology A topology, denoted as $\mathcal{T}$, over a set X is a collection of subsets of X such that:

(1) $\varnothing$ and X are in $\mathcal{T}$.

(2) The intersection of any finite number of elements in $\mathcal{T}$ is in $\mathcal{T}$.

(3) Arbitrary unions of elements are in $\mathcal{T}$.

Sets in $\mathcal{T}$ are normally referred to as open sets of X, and X is referred to as a topological space.

Definition 2.4. Uniform Continuity
For any $\epsilon > 0$, choose $\delta > 0$, such that $|x - y| < \delta$, we have $|f(x) - f(y)| < \epsilon$ [Her15].

Definition 2.5. Norm
A norm, denoted by $||x||$ for a vector x must satisfy the following properties (which are very similar to a Metric Space):

(1) $||x|| \geq 0$ and $||x|| = 0$ only if $||x||$ is a 0 vector.

(2) $||\lambda x|| = |\lambda| \cdot ||x||$

(3) *The Triangle Inequality:* $||x + y|| \leq ||x|| + ||y||$

Example. On $\mathbb{R}^n$, we define

$$||x|| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

This is known as the Euclidean norm (also known as the magnitude of a vector).

A few other key definitions that we will come across later in the paper but are important to be defined before are: bounded and closed sets, compact space, density, and closures.

Definition 2.6. Bounded sets
A set is called bounded if all of its points are said to be in a certain distance of each other.

Definition 2.7. Closed sets
Closed sets are those sets that contain all of their limit points.

Definition 2.8. Compact space
Compactness generalizes the idea of a closed and bounded subset of a Euclidean space. Similarly, a compact space is one that includes all of the limiting values or points.

Example. The interval $(0, 1)$ is open and therefore is not compact. However, the interval $[0, 1]$ is closed and therefore, includes the limiting points 0 and 1 and is compact.

Definition 2.9. Density
A subset $A \in \mathcal{T}$ is dense in $\mathcal{T}$ if every point in $\mathcal{T}$ belongs to A or is arbitrarily close to A.

Example. The rational numbers $\mathbb{Q}$ are dense in the subset of $\mathbb{R}$ because every element in $\mathbb{R}$ is either $\mathbb{Q}$ or is arbitrarily close to $\mathbb{Q}$.

Definition 2.10. Topological Closure
In topology, the closure of a subset A, denoted by $\overline{A}$, of a topological space consists of all points in A along with all of its limit points. A limit point of set A–not necessarily in set A–is the point you can get arbitrarily close to using other elements of A.

Example. Let us say that $A \subseteq \mathbb{Q}$ in the interval [0,1]. Then, we can say that $\overline{A}$ is all the real numbers in the same interval because this set contains all the rational numbers and for any real number you can find a sequence of rational numbers that get arbitrarily close to it.

Note: Sometimes, closure also refers to a subset being closed under an operation. Though this out of the scope of this paper, it is still worth noting.

3. Weierstrass Approximation Theorem

The Stone-Weierstrass theorem was a generalization of the Weierstrass Approximation theorem. Therefore, it becomes crucial to address this theorem to lay the groundwork for the Stone-Weierstrass theorem.

Theorem 3.1. (Weierstrass Approximation theorem, 1885) *[Rud76]* Let $f \in C([a,b],\mathbb{R})$ be continuous. Then, there is a sequence of polynomials $p_n(x)$ that converge uniformly to f on $[a,b]$. So, for any $\epsilon > 0$, there exists a polynomial P such that

$$|P(x) - f(x)| < \epsilon, \text{ for all } x \in [a,b].$$

In simpler terms, the theorem states that for any continuous real-valued function f in the closed interval $[a,b]$, the function can be very closely approximated arbitrarily well using polynomials p_n.

In order to prove the theorem, the Bernstein polynomial provides a constructive proof [PT96].

3.1. **Bernstein Polynomial.**

Theorem 3.2. (Bernstein Polynomial)

Described by [You06] as:

For each $n \in \mathbb{N}$, the n^{th} Bernstein Polynomial of a function $f \in C([0,1],\mathbb{R})$ is defined as

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

Note: First, we define $B_n(f)(x) := P(x)$. Second,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

is the binomial co-efficient and is also denoted as $\text{Bin}(n, x)$.

Before giving a proof of the Weierstrass approximation theorem, there are a few important properties of the Bernstein polynomials.

Lemma 3.3. *We have:*

$$\sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = 1, \text{for all } n \geq 0.$$

Proof. Using the binomial expansion, the right-hand side expands to

$$(x + (1-x))^n.$$

So for any n, we always get 1. ■

Lemma 3.4. *We have:*

$$\sum_{k=0}^{n} k \binom{n}{k} x^k (1-x)^n - k = nx.$$

Proof. We begin by applying the identity

$$k\binom{n}{k} = n\binom{n-1}{k-1}.$$

So we get

$$\begin{aligned}
\sum_{k=0}^{n} k \binom{n}{k} x^k (1-x)^{n-k} &= \sum_{k=1}^{n} n \binom{n-1}{k-1} x^{(k-1)+1} (1-x)^{(n-1)-(k-1)} \\
&= nx \sum_{k=1}^{n} \binom{n-1}{k-1} x^{k-1} (1-x)^{(n-1)-(k-1)} \\
&= nx(1) \quad \text{(From lemma 3.3).}
\end{aligned}$$

■

3.2. **Proof of the Weierstrass Approximation Theorem using the Bernstein Polynomial.**

Proof of the theorem. Since f is continuous on the closed interval $[0, 1]$, it is uniformly continuous. So, by definition, given $\epsilon > 0$, choose $\delta > 0$ such that:

$$|f(x) - f(y)| < \frac{\epsilon}{2} \text{ whenever } |x - y| < \delta, \text{ for all } x, y \in [0, 1]. \tag{3.1}$$

Note: By an affine (linear) change of variables, any closed interval $[a, b]$ can be mapped $[0, 1]$, so it suffices to prove the theorem in $[0, 1]$.

The Bernstein Polynomial can be represented as the expected value of $f\left(\frac{k}{n}\right)$. In other words,

$$B_n(f)(x) = \mathbb{E}\left[f\left(\frac{k}{n}\right)\right]. \tag{3.2}$$

This is because the weights $\binom{n}{k}x^k(1-x)^{n-k}$ correspond exactly to the probability mass function of the binomial distribution with parameters n and x. In other words, if $k \sim \text{Bin}(n, x)$ then

$$B_n(f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\mathbb{P}(k) = \mathbb{E}\left[f\left(\frac{k}{n}\right)\right]$$

and therefore, we get the equation 3.2.

Therefore,

$$|B_n(f)(x) - f(x)| = \left|\mathbb{E}\left[f\left(\frac{k}{n}\right) - f(x)\right]\right|.$$

Note: that this expectation can be written as follows

$$\mathbb{E}\left[f\left(\frac{k}{n}\right)\right] = \left|\sum_{k=0}^{n} f\left(\frac{k}{n}\right) - f(x)\binom{n}{k}x^k(1-x)^{n-k}\right|.$$

Using linearity of summation and applying the triangle inequality (from definition 2.2, item 3),

$$|B_n(f)(x) - f(x)| \leq \mathbb{E}\left[\left|f\left(\frac{k}{n}\right) - f(x)\right|\right]. \tag{3.3}$$

For any $\delta > 0$, we the expectation is split into two:

- The near points: $\left|\frac{k}{n} - x\right| < \delta$,
- The far away points: $\left|\frac{k}{n} - x\right| \geq \delta$,

such that,

$$\mathbb{E}\left[\left|f\left(\frac{k}{n}\right) - f(x)\right|\right] = A + B,$$

where we call these two parts as,

$$A := \sum_{\left|\frac{k}{n}-x\right|<\delta} \left|f\left(\frac{k}{n}\right) - f(x)\right| \binom{n}{k}x^k(1-x)^{n-k}$$

$$B := \sum_{\left|\frac{k}{n}-x\right|\geq\delta} \left|f\left(\frac{k}{n}\right) - f(x)\right| \binom{n}{k}x^k(1-x)^{n-k}.$$

$A < \frac{\epsilon}{2}$ through uniform continuity (definition 3.1)

Similarly, we need to prove that $B < \frac{\epsilon}{2}$ for some large n. Choose $M > 0$, such that

$$\left| f\left(\frac{k}{n}\right) - f(x) \right| \leq M + M = 2M.$$

Therefore, we have

$$B \leq 2M \sum_{\left|\frac{k}{n}-x\right|} \binom{n}{k} x^k (1-x)^{n-k}.$$

We want

$$2M \sum_{\left|\frac{k}{n}-x\right|} \binom{n}{k} x^k (1-x)^{n-k} < \frac{\epsilon}{2}. \tag{3.4}$$

We can observe that the sum is a tail probability of a binomial distribution. If we let k be a random variable with the parameters being n and x, we can represent the sum as

$$\sum_{\left|\frac{k}{n}-x\right|} \binom{n}{k} x^k (1-x)^{n-k} = P\left(\left|\frac{k}{n} - x\right| \geq \delta\right).$$

The variance (which is the measure of how spread out a set of numbers is) of $\frac{k}{n}$ is

$$Var\left(\frac{k}{n}\right) = \frac{x(1-x)}{n}.$$

An important definition to move on to the next step is the Chebyshev Inequality.

Definition 3.5. Chebyshev Inequality
The Chebyshev Inequality is as follows:

$$P(|x - \mathbb{E}[x]| \geq a) = \frac{Var(x)}{a^2}.$$

So, applying Chebyshev inequality, we get,

$$P\left(\left|\frac{k}{n} - x\right| \geq \delta\right) = \frac{Var\left(\frac{k}{n}\right)}{\delta^2} = \frac{x(1-x)}{\delta^2 n}.$$

Notice that $x(1-x) \leq \frac{1}{4}$ because the vertex of $x(1-x)$ is $\frac{1}{4}$. Therefore,

$$P\left(\left|\frac{k}{n} - x\right| \geq \delta\right) \leq \frac{1}{4\delta^2 n}.$$

On multiplying both sides by $2M$, we get

$$2M \sum_{\left|\frac{k}{n}-x\right|} \binom{n}{k} x^k (1-x)^{n-k} < \frac{M}{2\delta^2 n}.$$

So, for n large enough, specifically

$$n > \frac{M}{\delta^2 \epsilon},$$

we have,

$$B \leq 2M \sum_{\left|\frac{k}{n} - n\right|} \binom{n}{k} x^k (1-x)^{n-k} < \frac{\epsilon}{2}.$$

From this we get,

$$B < \frac{\epsilon}{2}.$$

As a result,

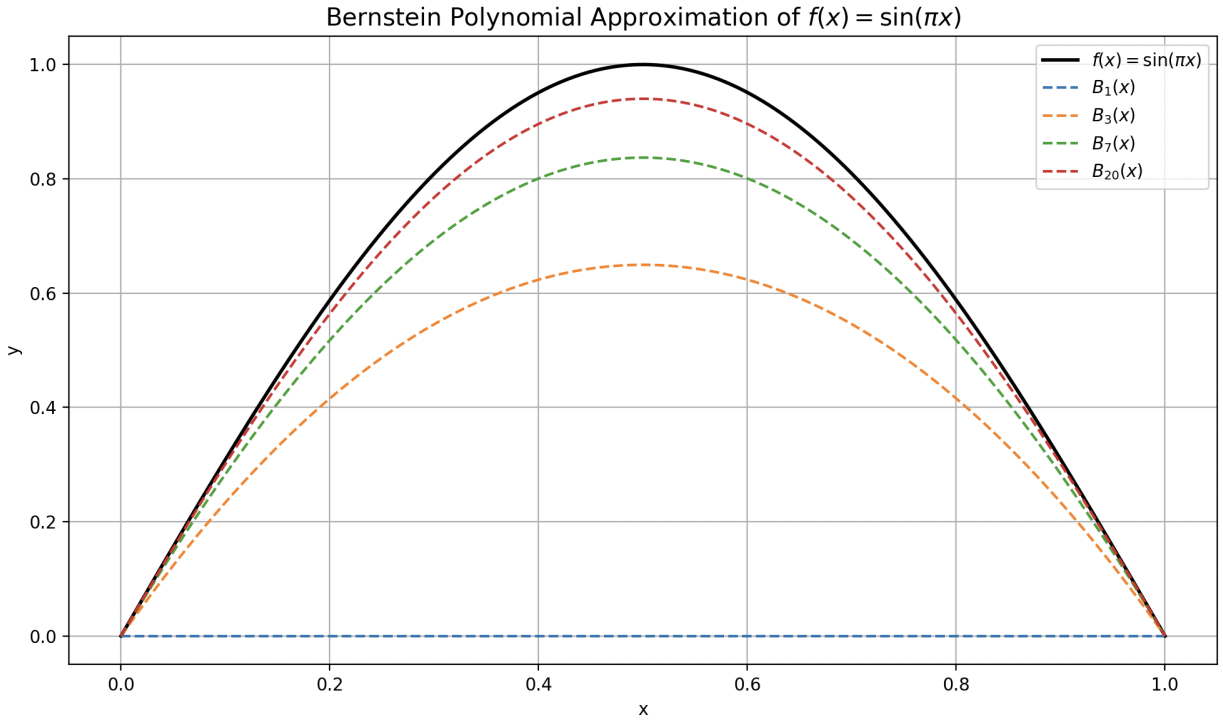
$$|B_n(f)(x) - f(x)| \leq A + B < \frac{\epsilon}{2} + \frac{\epsilon}{2} < \epsilon.$$

So,

$$|B_n(f)(x) - f(x)| < \epsilon. \qquad \blacksquare$$

Example. To create intuition, we shall create a visual proof.

We consider the function $f(x) = \sin(\pi x)$. We will approximate this using Bernstein polynomials in the interval $[0, 1]$.



Each dashed curve represents the Bernstein polynomial of different degrees. Each cruve makes use of a different degree n. As the value of n increases from 1 to 20, we see that the approximation becomes far more accurate and closer to the original curve. This proves that as

$$n \to \infty,$$

the Bernstein polynomial converges to $f(x)$ proving that for any function $f(x)$, we can find a set of polynomials $B_n(f)(x)$ such that they are in a very small $\epsilon > 0$.

4. The Stone-Weierstrass Theorem

Before moving on to the theorem we shall define a few key terms that will be important to understand the theorem.

Definition 4.1. Supremum norm
Let $f \in C(X)$, where X is compact. Then the sup norm is defined as:

$$||f||_\infty = \sup_{x \in X} |f(x)|.$$

Definition 4.2. Hausdorff Space
Consider two points $x, y \in X$, where X is a topological space. The two points can be separated by neighborhoods if there exists a neighborhood U of x and a neighborhood V of y such that they are disjoint ($U \cap V = \varnothing$). Therefore, X would be a Hausdorff space if any two distinct points in X can be separated by their neighborhoods.

Theorem 4.3. (Stone-Weierstrass Theorem) Let X be a compact (recall from definition 2.8) metric space, and let $\mathcal{A} \subseteq C(X, \mathbb{R})$ be a sub-algebra which separates points of X. Then $\mathcal{A}$ is dense (recall from definition 2.9) in $C(X, \mathbb{R})$ *[Puk17].*

Note: All metric spaces here are Hausdorff. The more general version of the Stone-Weierstrass theorem is stated for compact Hausdorff spaces.

Proof. Fix $\epsilon > 0$. For any $f \in C(X, \mathbb{R})$, we want to show that there exists $g \in \overline{\mathcal{A}}$ such that

$$||f - g||_\infty < \sup_{x \in X} |f(x) - g(x)| < \epsilon \tag{4.1}$$

which implies that $f \in \overline{\mathcal{A}}$, and so $\overline{\mathcal{A}} = C(X)$, which proves density (recall the definition of closures from definition 2.10).

Through affine (linear) interpolation, given any two distinct points $x, y \in X$, and any real numbers $\alpha, \beta \in \mathbb{R}$, we say that there exists $g \in \mathcal{A}$ such that $g(x) = \alpha$ and $g(y) = \beta$. This is true because we know that since $\mathcal{A}$ separates points, there exists $h \in \mathcal{A},\ h(x) \neq h(y)$ [You06].

Therefore, we can define

$$g(z) = \alpha + (\beta - \alpha)\frac{h(z) - h(x)}{h(y) - h(x)}.$$

So, when $z = x$,

$$g(x) = \alpha + (\beta - \alpha)\frac{h(x) - h(x)}{h(y) - h(x)} = \alpha + (\beta - \alpha) \cdot 0 = \alpha,$$

and when $z = y$,

$$g(y) = \alpha + (\beta - \alpha)\frac{h(y) - h(x)}{h(y) - h(x)} = \alpha + (\beta - \alpha) \cdot 1 = \alpha + \beta - \alpha = \beta.$$

The interpolation shows that the algebra can control function values at specific points which is vital in proving the idea of density (which in simpler terms is the ability to closely approximate continuous functions, in this context).

Because X is compact, we can find a point $p \in X$ which has a neighborhood U_p on which we can approximate f closely. We can choose finitely many points $p_1, p_2, \ldots, p_n \in X$ such that their neighborhoods $U_{p_1}, U_{p_2}, \ldots, U_{p_n}$ cover all of X.

We can construct a function $g_i(x) \in \mathcal{A}$ such that,

$$g_i(x) \in [f(x) - \epsilon, f(x) + \epsilon] \text{ for all } x \in U_{p_i}.$$

So on each neighborhood, we have a function g_i that always stays very close to f.

Now we define two new functions as

$$G(x) := \min\{g_1(x), g_2(x), \ldots, g_n(x)\}$$
$$H(x) := \max\{g_1(x), g_2(x), \ldots, g_n(x)\}$$

Since we are not very sure that the functions are in A, we can safely say that

$$G, H \in \overline{\mathcal{A}}.$$

From equation 4.1 we get,

$$f(x) - \epsilon < g_i(x) \text{ and } f(x) + \epsilon > g_i(x)$$

Thus,

$$f(x) - \epsilon < G(x) \text{ and } f(x) + \epsilon > H(x).$$

Since all of $g_i(x)$ are within an ϵ of f (that means some are above and below f), we can say that

$$G(x) \leq f(x) \leq H(x)$$

and

$$\begin{aligned} H(x) - G(x) &< (f(x) + \epsilon) - (f(x) - \epsilon) \\ &< 2\epsilon. \end{aligned}$$

Therefore, we have

$$|f(x) - g(x)| \leq \frac{H(x) - G(x)}{2} < \epsilon.$$

So,

$$\overline{\mathcal{A}} = C(X). \qquad \blacksquare$$

Since $\overline{\mathcal{A}}$ contains all the functions that can be uniformly approximated by elements of $\mathcal{A}$, $\overline{\mathcal{A}} = C(X)$ shows that $\mathcal{A}$ is dense in $C(X)$.

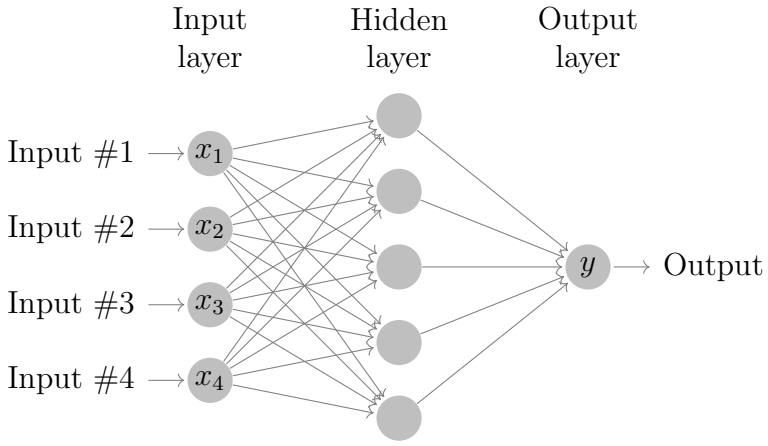
5. Applications of the Stone-Weierstrass Theorem

The Stone-Weierstrass theorem is not just limited to its generalization of polynomial approximation but rather it extends and plays a vital role in other mathematical setups too. The theorem extends to non-classical and abstract settings. Under this section, we will look at the following applications of the Stone-Weierstrass theorem: we shall look at how the Stone-Weierstrass Approximation theorem is used in the Universal Approximation theorem which will be followed by an f-ring interpretation of the theorem, under abstract algebra.

5.1. **Universal Approximation Theorem.**

In this subsection, under the Universal Approximation Theorem, we will look at neural networks. Therefore, it is pertinent to understand what we mean by neural networks. A neural network is a model that is inspired by the structure and function of the human brain. A neural network is divided into three sections: an input layer, followed by hidden layers, and an output layer.

The diagram below shows a very simplified version of a neural network with a single hidden layer.



A one-hidden layer neural network was defined by [Cyb89]. We define the set of functions in a one-hidden layer neural network (also known as a shallow feedforward neural network), as:

$$\mathcal{F} := \left\{ f : R^n -> R \mid f(x) = \sum_{j=1}^{N} \alpha_j \cdot G(w_j^T x + b_j) \right\}$$

where:

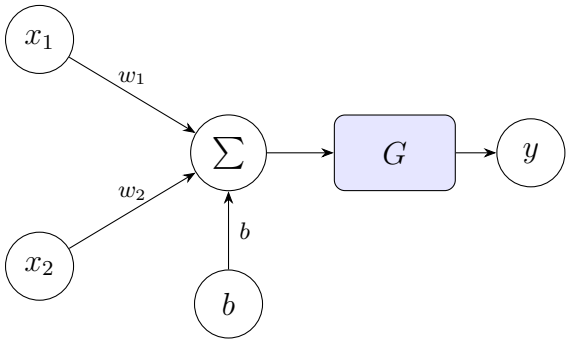
- $x \in R^n$ are the input vectors,
- Each $G(w_j^T x + b_j)$ is a neuron with weights $w_j \in R^n$ and bias $b_j \in R^n$, and activation function G,
- $\alpha \in R^n$ is the output vector.

This leads us to define a few key terms used above.

Definition 5.1. Weights
These are the numerical values that determine the influence an input has on a neuron's output.

Definition 5.2. Bias
A bias is a constant added to the neuron's weighted input which shifts the activation graph horizontally.

Using the new definitions, we can draw an even more accurate neural network that includes weights, biases, summation, and an activation function.



Now we shall move on to the Universal Approximation theorem. We shall begin by defining it formally.

Theorem 5.3. Universal Approximation Theorem [HSW89]. Let $\sigma : R \to R$ be a nonconstant, bounded, and continuous function. Let $C(K)$ be the space of real-valued continuous functions on a compact subset $K \subset \mathbb{R}^n$. Then for every, function $f \in C(K)$ and $\epsilon > 0$, there exists a neural network function:

$$f_N(x) = \sum_{j=1}^{N} \alpha_j \cdot \sigma(w_j^T x + b_j)$$

such that,

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

To simplify this and understand more intuitively, if we have an activation function, then even a single hidden layer neural network using the activation functions, is powerful enough to approximate any continuous function defined on a compact space with arbitrary precision as you add more neurons (N) to the approximation.

Note: an activation function can be of different sorts. The ones that will be considered in this paper are tanh and sigmoid curves. A sigmoid curve is defined as:

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

and a tanh curve is defined as:

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

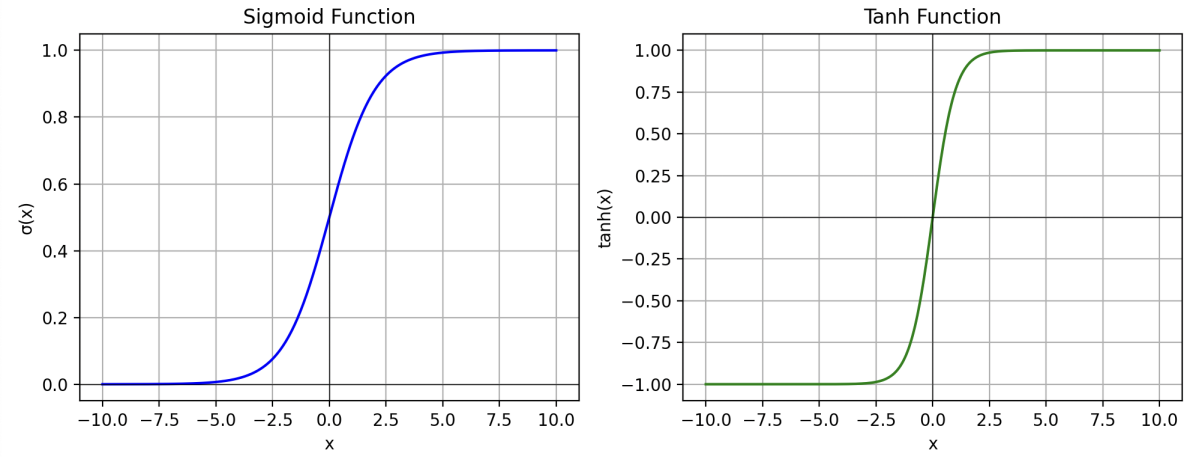
As

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

Similarly,

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

The figure below draws the two functions and represents the information mentioned above.



Note: for this version of the Universal Approximation theorem, that makes use of the Stone-Weierstrass Theorem, we must use functions like sigmoid and tanh because they are continuous (they behave nicely under limits), bounded (so their outputs are under control), and nonconstant (allowing them to have a rich enough approximations).

The Stone-Weierstrass interpretation of the Universal Approximation theorem cannot be applied in other activation functions like ReLU (Rectified Linear Unit) because they are not bounded. The ReLU functions output 0 whenever there is a negative input. However, for positive values, it can grow infinitely large and, therefore, they are unbounded and cannot be used by the Stone-Weierstrass interpretation of the Universal Approximation Theorem. However, keep in mind that ReLU functions can be approximated by the Universal Approximation theorem just not using the Stone-Weierstrass theorem.

Returning to theorem 5.3, we can recognize from theorem 4.3, that

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

The goal now becomes to prove that $\mathcal{F}$ satisfies the conditions of the Stone-Weierstrass theorem and is therefore dense in $C(X)$ which would allow us to prove that any continuous function can be approximated by functions in $\mathcal{F}$.

The function set $\mathcal{F}$ is closed under addition, scalar multiplication, and (approximately) under function multiplication. For detailed formal proofs, refer to [HSW89]. In this paper, we shall look at the proof for point separation and show that the function set contains constants.

First, we prove that $\mathcal{F}$ separates points.

Proof. If we have $x \neq y$, we can choose a w and b such that:

$$w^T x + b \neq w^T y + b.$$

Note: we are not trying to prove that $x \neq y$ because that is a given and we want to prove instead that whenever we choose two distinct points $x, y \in \mathcal{F}$, we can find a function in $\mathcal{F}$ such that we get different outputs at those two points.

Since G is injective (a on-to-one function), we get that

$$G(w^T x + b) \neq G(w^T y + b)$$

.

So by choosing the right biases and weights, the neurons outputs differ at the distinct points x and y. Therefore, we can say that $\mathcal{F}$ separates points. ■

Second, we shall show that $\mathcal{F}$ contain constants. Why is showing constants crucial? That is because we must be able to approximate constants of a function.

Example. So, for instance, let us take a function like $\sin(x) + 2$. Our approximation must be able to approximate the $+2$ aspect of the function which are intuitively important to build the target function.

Proof. We can show that $\mathcal{F}$ contains constants by taking

$$w = 0 \in \mathbb{R}^n$$

in order to eliminate the the dependence on x.

Then, we are able to get

$$f(x) = \alpha \cdot G(b),$$

which is a constant.

Therefore, we are able to prove that $\mathcal{F}$ contains constants. ■

As a result, we are able to prove that $\mathcal{F}$ is an algebra. We are able to invoke the Stone-Weierstrass theorem, which tells us that if $\mathcal{F} \subseteq C(X, \mathbb{R})$ separates points and contains constants, we can say that $\mathcal{F}$ is dense in $C(X, \mathbb{R})$.

Therefore, using the Stone-Weierstrass theorem we are able to prove that a shallow feedforward neural network, can approximate any continuous function arbitrarily well, leading us directly to the Universal Approximation Theorem.

Overall, the Universal Approximation theorem serves as a crucial theorem in the study of neural networks. The theorem allows us to prove that a single hidden layer is sufficient enough to approximate any continuous function, defined on a compact subset of $\mathbb{R}^n$ just by using suitable activation functions, to arbitrary precision. This idea allows the theorem to be further applied in deep learning, where it plays a very crucial role. Therefore, neural networks

can learn to perform complex tasks like image recognition, natural language processing (NLP), among others, by approximating the functions that define these tasks.

However, a key limitation of this is that it is not something that can be applied practically simply because it does not tell us the number of neurons required for a specific task and does not prescribe an architecture for doing so either, but rather just tells us that there is a structure like such that exists–even though this is a crucial result for neural networks.

5.2. f-Ring interpretation of the Stone-Weierstrass theorem.

First, let us begin by understanding what an f-ring is. Consider R to be a commutative l-ring (which is a commutative poring with a lattice). R is an f-ring if for all $x, y, z \in R$ we have

$$x \wedge y = 0 \text{ and } 0 \leq z, \text{ such that } zx \wedge y = 0 \Rightarrow xz \wedge y = 0. \tag{5.2.1}$$

It is important to to define a few key terms here. We must first begin by understanding what we mean by a lattice.

Definition 5.4. Lattice
A lattice can be defined as a type of poset such that every pair of elements has an infimum (the meet, denoted by $a \vee b$) and a supremum (the join, denoted by $a \wedge b$).

To understand what porings are we need to understand the partial orders, posets, and rings.

Definition 5.5. Partial Orders
It is a relation (or in simpler terms a rule for comparing things) that satisfies the following three properties:

- *Reflexivity:* Every item is related to itself. $a \leq a$.
- *Antisymmetric:* If $a \leq b$ and $b \leq a$, then $a = b$.
- *Transitive:* If $a \leq b$ and $b \leq c$, then $a \leq c$.

Definition 5.6. Partially Ordered Sets (Posets)
A poset is a pair $(X, \leq)$, where X is a set of elements with a partial order, in this case $\leq$ in X, such that the elements satisfy the three conditions stated in a partial order 5.5.

Definition 5.7. Rings
A ring is a set R with the two binary operations, addition $(+)$ and multiplication $(\cdot)$ such that it satisfies the following axioms:

(1) R is an abelian group under addition, which means that there is:
   - *Associativity under addition:* $(a + b) + c = a + (b + c)$ for all $a, b, c \in R$,
   - *Commutativity:* $a + b = b + a$.
   - *Additive Inverse:* For each $a \in R$, there exists $-a \in R$, such that $a + (-a) = 0$.
   - *Additive identity:* this states that for all $a \in R$, $a + 0 = a$ and therefore 0 is the additive identity because it leaves any element unchanged under addition.

(2) R is monoid under multiplication, which means that there is:

- *Associativity under multiplication:* $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in R$,
- *Multiplicative identity:* 1 is the multiplicative identity (an element that leaves an element unchanged under multiplication) because $a \cdot 1 = a$ and similarly, $1 \cdot a = a$ for all $a \in R$.

(3) Multiplication is distributive under addition, which means that:

- $a \cdot (b+c) = (a \cdot b) + (a \cdot c)$ for all $a, b, c \in R$ (this is often called left distributivity),
- $(b + c) \cdot a = (b \cdot a) + (c \cdot a)$ for all $a, b, c \in R$ (similarly, this is called right distributivity).

Having looked at all the things required for understanding porings, let us define porings.

Definition 5.8. Partially Ordered Rings (Porings)
A poring is a ring R with a partial order, let's say $\leq$, such that:

- $(R, \leq)$ is a poset;
- $x \leq y$ implies that $x + z \leq y + z$ for all $x, y, z \in R$;
- if $0 \leq x$ and $0 \leq y$, then $0 \leq xy$.

The equation we have used to understand what f-rings (equation 5.2.1) makes use of two notations $\wedge$ and $\vee$ which are known as meet and join respectively. Their definitions are given below.

Definition 5.9. Meet and Join (in Lattice Structures)
The meet of a subset S of a poset P is the infimum (greatest lower bound) of S and is denoted by $\wedge S$. Similarly, a join of a subset S of a poset P is the supremum (least upper bound) of S and is denoted by $\vee S$.

To simplify equation 5.2.1, if the meet of two elements is 0, then even multiplying by a non-negative element will not introduce any overlap.

The Stone-Weierstrass theorem applies directly to sub-f-rings in $C(X)$. [Ban01], in his 2001 paper shows the correlation between the Stone-Weierstrass theorem and f-rings.

He redefines the Stone-Weierstrass theorem by slightly tweaking it to be applicable under f-ring structures.

Theorem 5.10. Let X be a compact Hausdorff space. Suppose $\mathcal{R} \subset C(X)$ is a subring satisfying the following:

- $\mathcal{R}$ separates points of X: for any $x \neq y \in X$, there exists $f \in \mathcal{R}$ such that $f(x) \neq f(y)$,
- $\mathcal{R}$ is closed under pointwise minimum and maximum; that is, for all $f, g \in \mathcal{R}$, we have $\min(f, g), \max(f, g) \in \mathcal{R}$,

then we know that $\mathcal{R}$ is uniformly dense in $C(X)$ and $||f - g||_\infty < \epsilon$.

The importance of the use of the f-ring interpretation of the Stone-Weierstrass theorem, stems from the fact that the classical Stone-Weierstrass theorem 4.3, does not necessarily guarantee that the approximation will always stay positive, bounded, and maintain the order (maintaining increasing or decreasing behavior). These three conditions may be important in certain applications (such as, probability or ordered function spaces). However, with the f-rings interpretation of the Stone-Weierstrass theorem, we are able to preserve the three conditions widening its applications.

To further understand this concept, we will take a relatively simple and intuitive example. For instance, let $X = [0, 1]$, and let $\mathcal{R}$ be a set of piecewise linear functions (that is functions pieced together from multiple linear segments) on $[0, 1]$. This set forms a sub-f-ring of $C([0, 1])$ since it is a ring under pointwise minimum and maximum, it separates points of $[0, 1]$, and it's closed under min and max (because it takes pieces from the functions inside the ring and, therefore, it stays inside the ring).

6. Conclusion

In this paper, we began with the Weierstrass approximation theorem which shows that any continuous function on a closed interval can be uniformly approximated by a set of polynomials. The proof of the Weierstrass approximation theorem was given by the Bernstein polynomials because they are explicit, constructive, and allow mathematicians to compute these functions.

Then, the paper looked at the Stone generalization of the Weierstrass approximation theorem to broader subalgebras of continuous functions and compact Hausdorff spaces. We looked at a few key strategies to prove the theorem, like interpolation, compactness, and min and max operations to construct uniform global approximations.

It also looks at a few of the key applications of the Stone-Weierstrass theorem to the Universal Approximation theorem and neural networks. After that, the f-ring interpretation of the Stone-Weierstrass theorem was also briefly discussed.

The applications of the Stone-Weierstrass theorem can be further developed into other more abstract and complex function spaces, like Banach function spaces or C^*-algebras.

Overall, this paper tries to intuitively explain the two theorems and show how the 19th century analysis still continues to define the 21st century's mathematical and computational practices.

7. Acknowledgments

The author thanks Ophir Horowitz and Simon Rubinstein-Salzedo for the consistent guidance and mentoring throughout this paper. Without their insights, this paper would not have been possible.

References

[Ban01] B. Banaschewski. f-Rings and the Stone-Weierstrass Theorem. *Order*, 18(2):105–117, 2001.

[Cyb89] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

[Her15] Cesar A. Hernández. Epsilon-delta proofs and uniform continuity. *Lecturas Matemáticas*, 36(1):23–32, 2015. Demostraciones de límites y continuidad usando sus definiciones con epsilon y delta y continuidad uniforme.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, 1989.

[Joh21] P. Sam Johnson. Weierstrass Theorem and Some Generalizations. Monthly Seminar, National Institute of Technology Karnataka, Surathkal, 2021. Accessed via sam.nitk.ac.in.

[PT96] G.M. PHILLIPS and P.J. TAYLOR. Chapter 5 - 'best' approximation. In G.M. PHILLIPS and P.J. TAYLOR, editors, *Theory and Applications of Numerical Analysis (Second Edition)*, pages 86–130. Academic Press, London, second edition edition, 1996.

[Puk17] Liam Puknys. The Stone–Weierstrass Theorem. Research experience for undergraduates report, University of Chicago, 2017.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[You06] Matt Young. The Stone-Weierstrass Theorem, 2006.

Email address: vinamradhoot2009@gmail.com