Constructive Approximation of Functions

Benjamin Shih
Spring 2024

Abstract. These are notes from the Spring 2024 Brown University APMA DRP on the constructive approximation of functions, specifically polynomials and rational functions of one variable, under the direction of Dr. Wenjun Zhao. We closely followed the text *Approximation Theory and Approximation Practice* by Trefethen [1] and additionally engaged in material on universal approximation [2] as it relates to neural networks out of interest.

Contents

1 Week 1: 2/22/2024

Chebyshev interpolants have a very strong approximation properties, as opposed to uniformly spaced points. They are the points that correspond to the real part of equispaced points on the unit circle in the complex plane. That is, the Chebyshev points are

$$x_j = \cos\left(\frac{j\pi}{n}\right)$$

and a key property is that they "collect" near the ends of the interval in higher density. Namely, this key property is that each point is, on average, the same distance away from every other point. For the most part, we deal with approximating functions on the interval $[-1, 1]$, which any function on any interval $[a, b]$ can be scaled to.

There are connections that can be drawn between the Chebyshev, Fourier, and Laurent settings, with each being used in numerical, complex, and real analysis heavily, respectively. In the Chebyshev settings, we approximate functions $f(x), x \in [-1, 1]$ with the form

$$f(x) \approx \sum_{k=0}^{n} a_k T_k(x)$$

while using $z \in S^1 \subset \mathbb{C}$ equispaced points on the complex plane gives us the Laurent setting with Laurent polynomials

$$F(z) = F(z^{-1}) = \frac{1}{2}\sum_{k=0}^{n} a_k(z^k + z^{-k})$$

and finally using the angle $\theta \in [-\pi, \pi]$ to define $\mathcal{F}(\theta) = F(e^{i\theta}) = f(\cos(\theta))$ gives us Fourier series as

$$\mathcal{F}(\theta) \approx \frac{1}{2}\sum_{k=0}^{n} a_k(e^{ik\theta} + e^{-ik\theta})$$

Their corresponding canonical grid systems are as follow:

Chebyshev points	$x_j = \cos\left(\frac{j\pi}{n}\right),$	$0 \le j \le n$
Roots of unity (Laurent)	$z_j = e^{\frac{j\pi}{n}},$	$-n+1 \le j \le n$
Equispaced points (Fourier)	$\theta_j = \frac{j\pi}{n},$	$-n+1 \le j \le n$

1.1 Chebyshev Series and Polynomials

Definition 1.1 (k-th Chebyshev polynomial)**.** The k-th Chebyshev polynomial is the real part of the function z^k on the unit circle; i.e.

$$T_k(x) = \Re(z^k) = \frac{1}{2}(z^k + z^{-k}) = \cos(k\theta)$$

Theorem 1.2 (Existence of Chebyshev Series)**.** Suppose that f is Lipschitz on $[-1, 1]$, i.e. that there exists $C \in \mathbb{R}$ such that $|f(x) - f(y)| \le C|x - y|$ for any $x, y \in \mathbb{R}$. Then f admits a unique representation as a Chebyshev series

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x)$$

which is absolutely and uniformly convergent. The coefficients a_k are given by

$$a_k = \frac{2}{\pi} \int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}} dx$$

for $k \geq 1$, and for $k = 0$ by the same formula with a $\frac{1}{\pi}$ factor instead.

1.2 Introduction to Linear Approximation with Neural Networks

Generally, we will work with compact and convex domains $X \subset \mathbb{R}^d$ and target functions $f : X \to \mathbb{R}$ from some function space $\mathcal{F}(X)$. We denote by $f_\theta : \mathbb{R}^d \to \mathbb{R}$ a neural network with n_θ parameters. The study of the subject of neural network approximation is mainly concerned with the following three problems:

1. **Density**: When $n_\theta \to \infty$, does there exist f_θ that approximates f well?
2. **Covergence**: If $n_\theta < \infty$ is fixed, how close can f_θ be to f?
3. **Complexity**: If we want $\|f_\theta - f\| < \epsilon$, how large is n_θ?

Before discussing neural networks directly, we motivate with some background of density in polynomial approximation.

Definition 1.3 (Uniform Convergence)**.** A sequence of functions $\{f_n\}_{n=1}^{\infty}$ is said to converge uniformly to a limiting function f on a set X if for all $\epsilon > 0$, there exist $N \in \mathbb{N}$ such that $|f(x) - f_m(x)| < \epsilon$ for all $m \geq N$ and all $x \in X$ and denote this by

$$f_m \to f \quad \text{uniformly}$$

We can also equivalently define uniform convergence in terms of the supremum or infinity norm, where

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

by the condition $\|f - f_m\|_\infty \to 0$. Weierstrass gave the following result, which says that any continuous function f on a closed subinterval of the real line can be approximated arbitrarily well by an algebraic polynomial.

Theorem 1.4 (Weierstrass)**.** Given a function $f \in C([a,b])$ and $\epsilon > 0$, there exists an algebraic polynomial p such that $|f(x) - p(x)| < \epsilon$ for all $x \in [a,b]$, or equivalently, $\|f - p\|_\infty < \epsilon$.

It should be noted that a similar result exists for 2π-periodic continuous functions and trignometric polynomials; i.e., that trigonometric polynomials are dense in the class of 2π-periodic continuous functions.

2 Week 2: 2/29/2024

2.1 Interpolants and projections

Take $f(x)$ to be Lipschitz continuous on $[-1,1]$ with Chebyshev coefficients $\{a_k\}$ so that

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x)$$

Then we can approximate f in the space of n-degree polynomials by *interpolation*; that is, as

$$p_n(x) = \sum_{k=0}^{n} c_k T_k(x)$$

Another approximation of f is the *truncation* or *projection* to degree n, where the coefficientes through degree n match those of f:

$$f_n(x) = \sum_{k=0}^{n} a_k T_k(x)$$

Theorem 2.1 (Aliasing of Chebyshev polynomials). For any $n \geq 1$ and $0 \leq m \leq n$, the following Chebyshev polynomials take the same values on the $(n+1)$-point Chebyshev grid:

$$T_m, T_{2n-m}, T_{2n+m}, T_{4n-m}, T_{4n+m}, T_{6n-m}, \dots$$

Equivalently, for any $k \geq 0$, T_k takes the same value on the grid as T_m with

$$m = |(k+n-1) \pmod{2n} - (n-1)$$

which is a number in the range $0 \leq m \leq n$.

This leads to the connection between the coefficients of the polynomial approximant $\{a_k\}$ and the Chebyshev coefficients $\{c_k\}$.

Theorem 2.2 (Aliasing formula for Chebyshev coefficients). Let f be Lipschitz continuous on $[-1,1]$, and p_n be its Chebyshev interpolant in $\mathcal{P}_n$. Let $\{a_k\}$ and $\{c_k\}$ be the Chebyshev coefficients of f and p_n. Then

$$c_0 = a_0 + a_{2n} + a_{4n} + \cdots$$
$$c_n = a_n + a_{3n} + a_{5n} + \cdots$$

and for $1 \leq k \leq n-1$,

$$c_k = a_k + (a_{k+2n} + a_{k+4n} + \cdots) + (a_{-k+2n} + a_{-k+4n} + \cdots)$$

Essentially, this says that any f is indistinguishable from a polynomial interpolant of degree n on the $(n+1)$ point grid obtained by reassigning all Chebyshev coefficients $\{a_k\}$ to their aliases up to degree

n. The errors between the two different approximations are

$$f(x) - f_n(x) = \sum_{k=n+1}^{\infty} a_k T_k(x)$$
$$f(x) - p_n(x) = \sum_{k=n+1}^{\infty} a_k (T_k(x) - T_m(x))$$

where $m = |(k+n-1) \pmod{2n} - (n-1)|$ which are absolutely convergent. We note that f_n often leads to a approximation on the basis of relative error to f compared to p_n.

2.2 Barycentric Interpolation Formula

We're now interested in evaluating Chebyshev interpolants; multiple approaches exist and range from $O(n \log n)$ to $O(n)$ work. We focus on the latter, called the *barycentric interpolation formula*, which is direct and numerically stable. The formula takes the form

$$p(x) = \sum_{j=0}^{n} f_j \ell_j(x) \tag{1}$$

which is in alternative Lagrange form and is the linear combination of unique *Lagrange* or *cardinal* polynomials $\ell_j \in \mathcal{P}_n$ defined as

$$\ell_j(x_k) = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$$

In fact, we have an exact expression for ℓ_j:

$$\ell_j(x) = \frac{\prod_{k \neq j}(x - x_k)}{\prod_{k \neq j}(x_j - x_k)} \tag{2}$$

However, computational complexity wise, (2) is not so good, but we can remedy this.

Definition 2.3 (Node Polynomial)**.** The *node polynomial* $\ell \in \mathcal{P}_{n+1}$ for a given grid is

$$\ell(x) = \prod_{k=0}^{n} (x - x_k)$$

Using this, (2) then becomes

$$\ell_j(x) = \frac{\ell(x)}{\ell'(x_j)(x - x_j)}$$

Making the substitutions

$$\lambda_j = \frac{1}{\prod_{k \neq j}(x_j - x_k)}, \qquad \ell_j(x) = \ell(x) \frac{\lambda_j}{x - x_j}$$

(1) becomes the "type 1 barycentric formula":

$$p(x) = \ell(x) \sum_{j=0}^{n} \frac{\lambda_j}{x - x_j} f_j$$

2.3 Weierstrass Approximation Theorem

Chapter 6 in the text goes over the Weierstrass Approximation Theorem, which we covered in the neural network survey paper as Theorem 1.4. We note that the result has been generalized by theorem due to Runge and Mergelyan, which state that a function f defined on a compact set $K \subset \mathbb{C}$ with connected complent which is continous on K and analytic throughout K (resp. interior of K), then f can be approximated on K by polynomials.

3 Week 3: 3/7/2024

3.1 Convergence for Differentiable Functions

3.2 Density of Two Layer Neural Networks

We consider the family of two-layer feedforward networks consisting of:

- d input neurons
- r neurons in one hidden layer using the same activation $\sigma : \mathbb{R} \to \mathbb{R}$
- one output nueron with no activation nor bias.

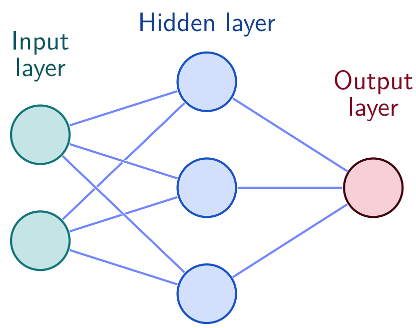

Figure 1: A feed-forward neural network with one hidden layer, where $d = 2, r = 3$.

If we let $x \in \mathbb{R}^d$ be the input to the network, then the ourput $y \in \mathbb{R}$ can be written

$$y = \sum_{i=1}^{r} W^2_{1,i} \sigma \left(\sum_{j=1}^{d} W^1_{i,j} x_j + b_i \right)$$

where $W^1 \in \mathbb{R}^{r \times d}$ and $W^2 \in \mathbb{R}^{1 \times r}$ are the weight matrices for the layer and $b \in \mathbb{R}^r$ is the bias in the hidden layer. Then the main desnity result for such networks is as follows.

Theorem 3.1 (Pinkus)**.** Let

$$\mathcal{M}(\sigma) = \operatorname{span}\{\sigma(w \cdot x + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

where $\sigma \in C(\mathbb{R})$. Then $\mathcal{M}(\sigma)$ is dense in $C(\mathbb{R}^n)$ with respect to the supremum norm on compact sets, if and only if σ is not a polynomial.

Essentially, Pinkus' theorem states that the network space $\mathcal{M}(\sigma)$ is dense in the space of continuous functions with respect to the supremum norm as long as the class of activation function chosen is non-polynomial. This means that given a target function $f \in C(\mathbb{R}^d)$ and a compact subset $X \subset \mathbb{R}^d$, for any $\epsilon > 0$, there exists $g \in \mathcal{M}(\sigma)$ so that

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

Examples of such σ which would satisfy the necessary conditions of Pinkus' theorem would be $\sin(\cdot), \cos(\cdot)$, etc. Let us consider the proof of the 1-dimensional case of the theorem; the extension to multiple dimensions can be found in [2].

Proof of the 1-dimensional case of Pinkus' theorem In this proof we consider the 1-dimensional case of $\mathcal{M}(\sigma)$, which we will denote

$$\mathcal{N}(\sigma) = \operatorname{span}\{(wx + b) : w, b \in \mathbb{R}\}$$

We will show that for any non-polynomial σ, $\mathcal{N}(\sigma)$ is dense in $C(\mathbb{R})$.

4 Week 4: 3/14/2024

5 Bibliography

[1] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice (Other Titles in Applied Mathematics)*. Society for Industrial and Applied Mathematics, USA, 2012. ISBN 1611972396.

[2] Mohammad Motamed. Approximation power of deep neural networks: an explanatory mathematical survey. preprint at https://arxiv.org/pdf/2207.09511.pdf, 2022.