Introduction

At its heart, approximation theory is concerned with understanding and quantifying the extent to which one function can be approximated by elements of another *given* collection of functions. As one would expect, it plays a central role in many fields, *inter alia* numerical analysis, differential equations, statistics, and data science. It is intimately related to the geometry of function spaces and compressibility. Indeed, a motif throughout these notes is the role that smoothness plays in the effective dimension of a function space, understood in a suitable sense.

It has a long history, dating at least back to Euler's 1777 work on minimizing errors in distances on maps of Russia; continuing through Laplace's work in 1843 on determining the best ellipsoidal approximation to the Earth; and Bernstein's work on constructive function theory in the 20th century, to today, where it underpins modern approaches to everything from PDE solvers, fast linear algebra, signal processing, and neural networks. In all of this, a key figure is Pafnuty Chebyshev, who in 1853 considered the following problem which arose in the study of locomotives:

Given an interval $[a, b]$, a function $f : [a, b] \to \mathbb{R}$, and a natural number n, solve

$$\min_{p \in \mathcal{P}_n} \max_{a \leq x \leq b} |f(x) - p(x)|,$$

where $\mathcal{P}_n$ denotes the set of polynomials of degree at most n.

Understanding this innocuous-looking optimization problem will be a central focus of these notes and, as we shall see, has far ranging implications, extensions, and applications. When considering an approximation problem, there are three fundamental mathematical questions to consider: what functions are we approximating? With what? In what sense? Additionally, from a computational perspective, when assessing a method for approximation, there are three fundamental computational traits that are desirable for a method to possess: is it accurate? Is it numerically stable? Can it be carried out in a reasonable amount of time?

A motivating example

As a first example that will highlight many of the themes of these notes, let us first consider the example of numerical integration. That is to say, we want to obtain information ($\int_0^1 f(x)\,\mathrm{d}x$) from samples. This leads to a problem of *quadrature*.

One of the simplest approaches to approximate such an integral would be through either the left-hand or right-hand Riemann rules:

$$L_n(f) := \frac{1}{n}\sum_{k=0}^{n-1} f(k/n), \quad R_n(f) := \frac{1}{n}\sum_{k=1}^{n} f(k/n),$$

respectively. Note that both of these approximations take the general form

$$\sum_{k=0}^{n} f(x_k)w_k,$$

where $x_k = k/n$ and $w_k = 1/n$ except at the right endpoint (for L_n) or the left endpoint (for R_n). The x_k are called *quadrature points* and the w_k are referred to as the corresponding *quadrature weights.* The error is clearly dependent on the smoothness of f. Indeed, it is easy to construct a function for which the integral, in a Lebesgue sense, exists and is equal to zero, but the left-hand and right-hand Riemann rules with equispaced points take on any desired value for all n.

Thus, we must buy convergence at the price of imposing more assumptions on our function f. One way to think about this is that the space of all functions, even all bounded functions, is too large to allow for numerical integration. Instead, we might consider restricting our attention to a subset of functions which are almost low dimensional. Indeed, approximation theory is fundamentally about function "compression" - the extent to which a given set of functions can be well-approximated by a finite dimensional subspace. One mechanism for a function space to be compressible is smoothness. An old game in numerical analysis (or analysis in general) is to buy decay at the cost of derivatives.

Returning to our example, we see that from Taylor's theorem, if f is smooth enough, then

$$|R_n(f) - I(f)| \leq C\|f'\|_\infty n^{-1}.$$

Here $\|\cdot\|_\infty$ denotes the L^∞ norm, C is a constant independent of f, and $I(f)$ is the true integral of f

$$I(f) := \int_0^1 f(x)\,\mathrm{d}x.$$

A similar result holds for the left-hand sum. These rules have a classical interpretation: if the integral denotes the (signed) area under the curve $y = f(x)$, then $R_n(f)$ is the area under the piecewise constant approximation to f which is equal to $f(k/n)$ for all $x \in ((k-1)/n, k/n]$, $k = 1, \dots, n$.

Using this intuition, generalizing is rather straightforward. Rather than piecewise constant, we could instead use a piecewise linear approximant, giving us the trapezoid rule:

$$T_n(f) := \frac{f(0)+f(1)}{2n} + \frac{1}{n}\sum_{k=1}^{n-1} f(k/n).$$

For smooth enough functions this looks much better. And it is! Indeed, one can easily show that

$$|T_n(f) - I(f)| \le C\|f''\|_\infty n^{-2}.$$

Exercise 1. *Prove this. Note that the increase in accuracy relies on smoothness.*

Exercise 2. *Can you find a counterexample when the function is differentiable, but not twice differentiable?*

Clearly, one could continue on in this way. But, there is a more subtle point that the geometric picture misses. Looking at the formulae for the three rules, we get

$$T_n(f) = \frac{L_n(f) + R_n(f)}{2},$$

and so the trapezoid rule is the average of the left and right. So far, so reasonable. Generically speaking, one "overshoots" the other "undershoots" and the two errors almost cancel. But, returning to the formula for T_n,

$$T_n(f) = \frac{f(0/n)}{2n} + \frac{1}{n}\sum_{k=1}^{n-1} f(k/n) + \frac{f(n/n)}{2n}.$$

In particular, only the endpoints are affected. From this, we conclude that almost all of the error comes only from the endpoints. How should we interpret this?

A first approach: Fourier series

In the following, we will assume that f is smooth, and it, along with all of its derivatives, is periodic (i.e. $f(0) = f(1)$, $f'(0) = f'(1), \cdots$). Then it can be shown that f can be expressed as a Fourier series

$$f(x) = \sum_{m \in \mathbb{Z}} e^{2\pi i m x} f_m,$$

where

$$f_m = \int_0^1 e^{-2\pi i m x} f(x)\,\mathrm{d}x.$$

In general, one needs much less regularity for the Fourier representation to exist. Later in these notes we will revisit decompositions of this type, exploring under what conditions they are defined and in what sense they converge to f. For now, we will not worry too much about these issues.

After substituting our expression for f into the formula for $T_n(f)$, we find that

$$T_n(f) = \sum_{j=0}^{n-1} \left(\sum_{m \in \mathbb{Z}} e^{2\pi i m x_j} f_m \right) \frac{1}{n} = \sum_{m \in \mathbb{Z}} \frac{f_m}{n} \sum_{j=0}^{n-1} \left(e^{2\pi i m/n} \right)^j.$$

Here $x_j = j/n$, and we have used the periodicity of f to equate the contributions from $j = 0$ and $j = n$. In particular, as a first sign that there is something interesting happening, for equispaced points the left-hand Riemann rule, the right-hand Riemann rule, and the trapezoid rule all give the same result.

The inner sum on the right-hand side of the previous expression is a truncated geometric series, and can be explicitly summed, which gives

$$\sum_{j=0}^{n-1} \left(e^{2\pi i m/n} \right)^j = \begin{cases} n, & e^{2\pi i m/n} = 1, \\ \frac{e^{2\pi i m} - 1}{e^{2\pi i m/n} - 1}, & e^{2\pi i m/n} \neq 1. \end{cases}$$

So,

$$T_n(f) = \sum_{m \in n\mathbb{Z}} f_m.$$

Now, $I(f) = f_0$ and hence

$$E_n(f) := |T_n(f) - I(f)| = \left| \sum_{m \in n\mathbb{Z},\, m \neq 0} f_m \right|.$$

Thus, the speed of the convergence is directly related to the rate of decay of the Fourier coefficients. How does this relate to smoothness?

We begin by observing that if we take the expression for f_m and integrate by parts, for $m \neq 0$,

$$f_m = -\frac{1}{2\pi i m} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}x} \left(e^{-2\pi i m x} \right) f(x)\,\mathrm{d}x = \frac{1}{2\pi i m} \int_0^1 e^{-2\pi i m x} f'(x)\,\mathrm{d}x.$$

The boundary conditions arising in the integration by parts disappear because of the periodicity of f and its derivatives. Indeed, now we see an explanation for the heuristic observation that the error is dominated by the endpoints. If f is not periodic then we must keep the boundary terms, which would in turn contribute to slower decay in the Fourier coefficients.

Naturally, if f is smooth enough, then we can iterate, to obtain

$$f_m = \frac{1}{(2\pi i m)^k} \int_0^1 e^{-2\pi i m x} f^{(k)}(x)\,\mathrm{d}x.$$

This formula yields the following bound for f_m

$$|f_m| \leq \frac{1}{2\pi |m|^k} \|f^{(k)}\|_1,$$

and hence

$$E_n \leq \frac{\|f^{(k)}\|_1}{(2\pi)^k} \sum_{j\in\mathbb{Z}, j\neq 0} \frac{1}{|jn|^k} \leq \frac{2\|f^{(k)}\|_1}{(2\pi n)^k} \sum_{j=1}^{\infty} \frac{1}{j^k}.$$

The last sum is independent of f and n, and converges for all $j > 1$. A crude bound can be obtained by comparing it to the integral of $1/x^k$ on $[1, \infty)$, which gives the upper bound $k/(k-1)$. So, for all $n \geq 2$,

$$E_n \leq C \frac{\|f^{(k)}\|_1}{(2\pi n)^k}.$$

Here we see the trade-off between smoothness and integration error very clearly.

A second approach: the Euler-Maclaurin formula

For our next approach, we will use the Euler-Maclaurin formula to bound the error in our quadrature rules. To do so, we need a little bit of notation. Given integers m, n and an integrable function $g : [m, n] \to \mathbb{R}$, we set

$$I_{m,n}(g) = \int_m^n g(x)\,\mathrm{d}x$$

and

$$S_{m,n}(g) = g(m+1) + \cdots + g(n).$$

Theorem 1. *If $k \in \mathbb{N}$ and $g \in C^k([m, n])$ then*

$$S_{m,n} - I_{m,n} = \sum_{j=1}^{k} \frac{B_j}{j!} \left[f^{(j-1)}(n) - f^{(j-1)}(m) \right] + R_k.$$

Here, B_k are the Bernoulli numbers (see Table 1). The remainder term, R_k, is bounded by

$$|R_k| \leq \frac{2\zeta(k)}{(2\pi)^k} \int_m^n |f^{(k)}(x)|\,\mathrm{d}x$$

if k is even. For odd k the Riemann zeta function, $\zeta(k)$, can be omitted.

n	B_n
0	1
1	1/2
2	1/6
4	-1/30
6	1/42
8	-1/30
10	5/66
12	-691/2730
14	7/6
16	-3617/510
18	43867/798
20	-174611/330
⋮	⋮

Table 1. *Bernoulli numbers. All odd Bernoulli numbers greater than one are zero.*

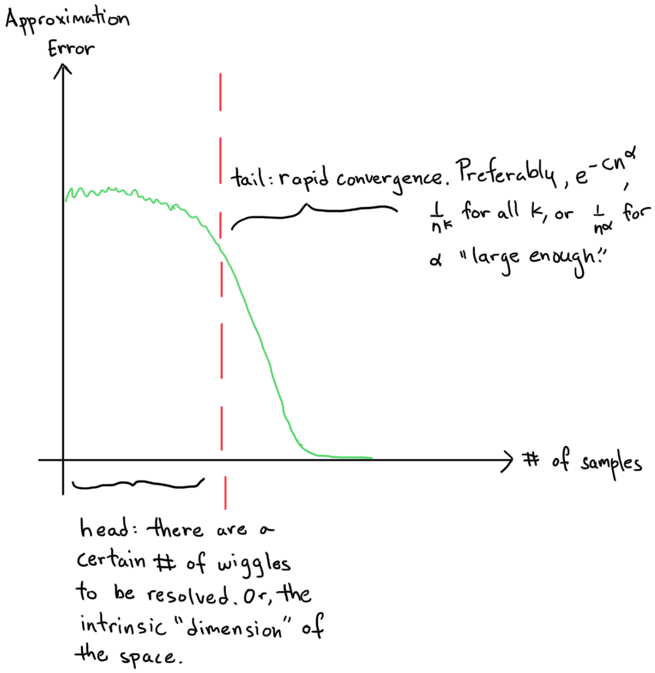
Rescaling to $[0,1]$ and setting $k = 2$, $m = 0$, we see that

$$T_n(f) - I(f) = \frac{n^{-2}}{12}|f'(1) - f'(0)| + R_2 \in O(n^{-2}).$$

Again, higher-order error terms come from the failure of derivatives of f to be periodic.

Summary

For periodic, smooth functions, the trapezoid, left-hand, and right-hand rules are the gold standard. They converge faster than any power of n (though one has to be a little bit careful with interpreting this). Schematically, the approximation error of an ideal scheme looks like the plot below.



For the trapezoid rule, we use differentiability to get compressibility. Finally, the error is not spread uniformly throughout the interval.

Additional Exercises

Exercise 3. *Assuming that $f \in C^4(0,1)$, construct an improved quadrature rule which computes the integral of f on $(0,1)$ with an error which is $O(h^4)$. Your rule should only involve a constant number of modifications to the standard left-hand or right-hand Riemann sums as $n \to \infty$. Implement your quadrature rule and plot the convergence for: i) $f(x) = e^x$*

and ii) $f(x) = \sqrt{x}$ *(of course this isn't differentiable up to the boundary!). Compare your numerical results with those obtained from Simpson's rule – using a quadratic approximation on each subinterval, where the quadratic is chosen to agree with* f *at the two endpoints and the midpoint of each subinterval.*

References and Further Reading

For an interesting account of the early days of approximation theory, and a glimpse into some of the methods used, see the book *The History of Approximation Theory: From Euler to Bernstein* by Anastassiou, George A.

The Basics of Bases

In this chapter, our goal will be to formulate means of describing, classifying, and quantifying spaces of functions and notions of distance therein. The properties of the functions we wish to approximate and those with which we approximate them, determine the rate of convergence, the algorithm, and the manner of the proofs. Generic theorems (or algorithms) work for broad classes of functions but then do not typically exploit the unique structure of a specific problem. The same considerations apply to measures of distance, our definition of "success".

More importantly, in applications there is typically a natural "budget" for accuracy of certain quantities, which induce natural objective functions for approximation problems. Do we care about outliers (i.e. catastrophic failure due to rare events, machining parts, etc.)? Or, do we care about the average, or a weighted average, being small (like antenna design)?

One of the most natural and convenient ways of specifying collections of functions is through *bases*. Every finite dimensional linear vector space, E, has a basis - a maximal linearly independent collection of vectors in E. This generalizes (rather unsatisfyingly) to infinite dimensions in the following way.

Definition 1. *A subset $\mathcal{X}$ of a vector space E is a Hamel basis if every vector $v \in E$ can be expressed uniquely as a finite linear combination of elements in $\mathcal{X}$. That is to say, for every $v \in E$, there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $v_1, \ldots, v_n \in \mathcal{X}$ such that*

$$v = \sum_{i=1}^{n} \alpha_i v_i.$$

Without more, we are limited to finite sums. In particular, we need a notion of convergence to use infinite sums. Even for relatively pedestrian spaces, $\mathcal{X}$ is typically uncountable.

Proposition 1. *Every linear vector space has a Hamel basis.*

Proof. A somewhat sterile application of Zorn's lemma which we leave to the reader. Note that linear independence involves only finite linear combinations. □

Still, this gives us our first approximation result. We now impose additional structure on our vector spaces, with the hope of obtaining more useful results. In particular, we now focus on the case in which E is a Banach space.

Moving to Banach spaces

Definition 2. *A sequence $\{v_j\}_1^\infty$ in a Banach space X is called a Schauder basis if for all $v \in X$ there exist unique coefficients $\alpha_1, \alpha_2, \ldots$ such that*

$$v = \sum_{j=1}^{\infty} \alpha_j v_j,$$

where implicitly we require the sum on the right-hand side in the last expression to converge.

It is easy to show that X must be separable to have a Schauder basis. Unfortunately, this is not sufficient (though finding counterexamples is non-trivial - see Enflo 1973).

The next result allows us to control the size of the partial sums in terms of the norm of v.

Theorem 2. *Suppose $\{v_j\}_1^\infty$ is a Schauder basis of the Banach space X. Then there exists a constant M such that for all $v \in X$,*

$$\left\| \sum_{j=1}^{N} \alpha_j v_j \right\| \leq M \|v\|$$

for all $N = 1, 2, \cdots$. Here, the α_j's are the coefficients such that

$$v = \sum_{j=1}^{\infty} \alpha_j v_j.$$

The minimal M for which the above inequality holds for all M and N is called the basis constant.

Proof. We define a new space E and show that it is isomorphic to X. In particular, set

$$E := \left\{ \{\alpha_j\}_1^\infty \mid \sum_{j=1}^{\infty} \alpha_j v_j \, \text{converges in} \, X \right\}.$$

Moreover, for $\alpha \in E$, set

$$\|\alpha\|_E := \sup_{N \in \mathbb{N}} \left\| \sum_{j=1}^{N} \alpha_j v_j \right\|.$$

It is easily verified that, equipped with this norm, E is a Banach space. Moreover, clearly there is a bounded bijection from E to X. Thus, by the inverse mapping theorem, this is an isomorphism. □

Now, for each positive integer N we can define the operator $S_N : X \to X$ by

$$S_N(v) := \sum_{j=1}^{N} \alpha_j v_j,$$

where again $\{v_j\}_1^\infty$ is a Schauder basis and the $\{\alpha_j\}_1^\infty$ are the corresponding coefficients of v. Note that the α_j are independent of N. Clearly S_N is a projection onto the span of $\{v_1, \ldots, v_N\}$ and, by the previous theorem, all the S_N's are uniformly bounded. It follows that that the coefficients are linear functionals on X, called *biorthogonal functionals*, and are denoted by v_j^*. Hence,

$$v = \sum_{j=1}^{\infty} v_j v_j^*(v).$$

Clearly,

$$\|v_j\| |v_j^*(v)| = \|v_j v_j^*(v)\| = |\|S_j(v) - S_{j-1}(v)\| \le 2M\|v\|$$

and so $\|v_j\| \, \|v_j^*\| \le 2M$ for all j.

This gives a slick characterization of compactness.

Theorem 3. *A subset $A \subset X$ is precompact if and only if A is bounded and the basis expansions converge uniformly, i.e.*

$$\|v - S_N(v)\| \le \epsilon_N \to 0,$$

for all $v \in A$.

Note here that the ϵ_N do not depend on v.

Proof. Necessity: from above, we know that S_N converges to the identity operator pointwise. Using a standard ϵ-net argument, since convergence is uniform on compact sets, it follows that S_N converges uniformly to the identity on A.
Sufficiency: For any ϵ, there exists an N such that $\|v - S_N(v)\| \le \epsilon$ for all $v \in A$. The image of S_N is a finite dimensional subspace and $A \subset \mathcal{N}_\epsilon(S_N)$ (A is contained within an ϵ-neighborhood of the image of S_N). □

We now impose even more mathematical structure to obtain stronger results. For Banach spaces we added a notion of length. Going to Hilbert spaces, we now also add a notion of angle.

Hilbert spaces

Definition 3. *A set $\{v_j\}_1^\infty$ in H (a Hilbert space) is an orthogonal system if $\langle v_j \, , \, v_k \rangle = 0$ for all $j \neq k$. It is orthonormal if additionally $\|v_j\| = 1$ for all j.*

Theorem 4. *Let $\{v_j\}_1^\infty$ be an orthogonal system. The following are equivalent:*

i $\sum_j \alpha_j v_j$ *converges in H*

ii $\sum_j |\alpha_j|^2 \|v_j\|^2 < \infty$

iii $\sum_j \alpha_j v_j$ *converges unconditionally*

If there is convergence then

$$\|\sum_j \alpha_j v_j\|^2 = \sum_j |\alpha_j|^2 \|v_j\|^2.$$

Proof. We leave the proof as an exercise. □

Definition 4. *Let $\{v_j\}_1^\infty$ be an orthonormal system in H. The* Fourier series *of v with respect to $\{v_j\}_1^\infty$ is*

$$\sum_{j=1}^{\infty} \langle v_j, v \rangle v_j.$$

The scalar quantities $\langle v_j, v \rangle$ are called the Fourier coefficients.

We end this section with several useful results about orthonormal systems on Hilbert spaces.

Lemma 1. *If $\{v_j\}_1^\infty$ is an orthonormal system then for each N, S_N is an orthogonal projection.*

Lemma 2 (Bessel's inequality)**.** *If $\{v_j\}_1^\infty$ is again an orthonormal system in a Hilbert space H,*

$$\sum_{j=1}^{\infty} |\langle v_j, v \rangle|^2 \leq \|v\|^2.$$

We also have the following optimality result.

Lemma 3. *Among all convergent sequences $s = \sum_j \alpha_j v_j$, $\|v - s\|$ is minimized by the Fourier series of v. The same is true of partial sums.*

Thus, finding a best approximation from a subspace is easy in a Hilbert space provided we have an orthonormal system. Here we should add that "best" is being measured with respect to the norm of the Hilbert space. A natural follow-up question is how to find orthonormal systems. If H is separable, it's easy, at least in practice. Just apply Gram-Schmidt to the countably dense set.

Projections

Projection operators arise naturally in approximation theory. Here we briefly review some general theory of projections which will be useful throughout these notes.

Definition 5. *Given a Banach space E, the linear operator $P : E \to E$ is called a* projection *if it is bounded and* idempotent, *i.e.* $P^2 = P$.

The following theorem gives a handy characterization of projections.

Theorem 5. *Suppose $P : E \to E$ is a projection and $\mathcal{R}(P) = V$. Then V is a closed subspace of E, and $V = \ker(I - P)$. Here I is the identity operator on E.*

Proof. To prove the first assertion, note that $I - P$ is continuous, so its kernel is closed. Thus, it suffices to show that $V = \ker(I - P)$. To see this, observe that if $v \in V$, then there exists a $u \in E$ such that $Pu = v$. Then $Pv = P^2u = Pu = v$. Thus $v \in \ker(I - P)$ and hence $V \subseteq \ker(I - P)$. Similarly, if $w \in \ker(I - P)$ then $Pw = w$ and so $w \in V$, implying the reverse inclusion. □

Given a projection P, there are a few other natural projections we can construct.

Proposition 2. *Let $P : E \to E$ be a projection. Then $(I - P) : E \to E$ is also a projection. So is the adjoint of P, $P^* : E^* \to E^*$.*

Proof. Both are clearly linear and bounded. For idempotency, for $I - P$, we observe that $(I - P)^2 = I - 2P + P^2 = I - P$. For the adjoint, recall that it is defined by: $(P^*\ell)(v) = \ell(Pv)$, for all $v \in E$. Hence, $((P^*)^2\ell)(v) = (P^*\ell)(Pv) = \ell(P^2v) = \ell(Pv) = (P^*\ell)(v)$. □

Given two projections, $P, Q : E \to E$ it is natural to ask whether it is possible to combine them. Unfortunately, in general, neither PQ nor QP are projections. The following result gives some conditions on which certain new projections can be created from two existing ones.

Proposition 3. *Let $P, Q : E \to E$ be two projections. Define the operator $P \oplus Q = P + Q - PQ$. Suppose $PQP = QP$. Then*

1. *$P \oplus Q$ is a projection onto $\mathcal{R}(P) + \mathcal{R}(Q)$.*

2. *QP is a projection onto $\mathcal{R}(P) \cap \mathcal{R}(Q)$.*

Proof. The proof follows by a direct calculation.

$$\begin{aligned}(P \oplus Q)^2 = (P + Q - PQ)^2 &= (P^2 + PQ - P^2Q) + (QP + Q^2 - QPQ) - (PQP + PQQ - PQPQ)\\ &= P + QP + Q - QPQ - PQP - PQ + (PQP)Q = P + Q - QPQ - PQ + QPQ\\ &= P + Q - PQ.\end{aligned}$$

Moreover, $(P \oplus Q)P = P$, and so $(P \oplus Q)u = u$ for all $u \in \mathcal{R}(P)$. Similarly, $(P \oplus Q)Q = Q$ and hence $(P \oplus Q)v = v$ for all $v \in \mathcal{R}(Q)$. If

$w \in \mathcal{R}(P) + \mathcal{R}(Q)$ then $w = u + v$ where $u \in \mathcal{R}(P)$ and $v \in \mathcal{R}(Q)$. From above, both u and v are fixed by $P \oplus Q$. The proof of the second statement follows similarly, and we leave it as an exercise. □

Remark 1. *The operator $P \oplus Q$ is called the* Boolean sum *of the linear operators.*

We finish our discussion with the following result which shows that if we are willing to consider only a finite-dimensional subspace U, a projection onto U always exists, and the norm is not too big.

Theorem 6. *Let U be an n-dimensional subspace of a Banach space E. There exists a projection $P : E \to U$ with norm at most n.*

Proof. Before proving the main result we need the following tool from functional analysis (it is a simplified version of Auerbach's theorem):

If U is an n-dimensional normed space, then there exist vectors $u_1, \cdots, u_n$ and functionals $\ell_1, \cdots, \ell_n$ such that $\|u_i\| = 1$, $\|\ell_i\| = 1, 1, \cdots, n$, and $\ell_j(u_i) = \delta_{i,j}$.

We sketch the proof briefly. Let $v_1, \cdots, v_n$ be a basis of U. Define $M : (U^*)^{\otimes n} \to \mathbb{R}$, by $M(\nu_1, \cdots, \nu_n) = |\det(\nu_j(v_i))|$. Since U^* is finite-dimensional, M attains its maximum m_* in the unit ball $B_{U^*} \times \cdots B_{U^*}$ at some point $(\ell_1, \cdots, \ell_n)$. Note that $m_* > 0$ since otherwise $A_{i,j} = \nu_j(v_i)$ would have a left nullvector for any choice of ν's. This in turn would imply that there is a v such that $\ell(v) = 0$ for all $\ell \in U^*$ (just choose $\nu_1, \cdots, \nu_n$ to be a basis of U^* for example).

Now set $A_{j,i} = \ell_i(v_j)$ and let $C = A^{-1}$. We set $u_j = \sum_{k=1}^n C_{j,k} v_k$. It is easy to see that $\ell_i(u_j) = \delta_{i,j}$, $1 \le i, j \le n$. To show the bound on $\|u_j\|$, note that for any $\psi \in E^*$,

$$\psi(u_i) = \sum_{j=1}^n C_{i,j} \psi(v_j).$$

Next, we observe that $C_{i,j}\psi(v_j)$ (as a vector) is a solution to the equation $Ax = \psi(\vec{v})$. By Cramer's rule,

$$\sum_{j=1}^n C_{i,j}\psi(u_j) = \frac{\det A_i}{\det(A)},$$

where A_i is the matrix formed by replacing the ith column with $\psi(u_j)$. In particular,

$$|\psi(u_i)| = \left| \frac{M(\ell_1, \cdots, \ell_{i-1}, \psi, \ell_{i+1}, \cdots, \ell_n)}{M(\ell_1, \cdots, \ell_n)} \right| \le \|\psi\|.$$

Thus, since ψ was arbitrary, $\|u_i\| \le 1$.

We now turn to the proof of the theorem. From Auerbach's theorem, we can find $u_1, \cdots, u_n \in U$ and $\ell_1, \cdots, \ell_n \in U^*$. By Hahn-Banach, the latter can be extended to act on all of E, while not increasing the norm. Then, set $P(v) = \sum_i u_i \ell_i(v)$. Clearly, $\mathcal{R}(P) = U$ and $P^2 = P$. Finally,

$$\|Pv\| = \|\sum_i u_i \ell_i(v)\| \leq \sum_i \|u_i\| \|\ell_i\| \|v\| \leq n\|v\|.$$

□

References and Further Reading

There are many good functional analysis textbooks which cover this. Here some of the presentation follows the notes by R. Vershynin. For projections, see *A Comprehensive Course in Analysis, Volume 5* by B. Simon. For a more applied approach, see *A Course in Approximation Theory* by W. Cheney and W. Light, which served as the basis for the results presented here.

Additional Exercises

Exercise 4. *Let $\{v_j\}_{j=1}^{\infty}$ be an orthonormal basis of a Hilbert space H. Let A be a bounded operator from H to itself with bounded inverse. Consider the sequence of vectors $w_j := Av_j$. Is $\{w_j\}_1^{\infty}$ an orthonormal basis? A Schauder basis? If so, can you bound its basis constant?*

Exercise 5. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be the function defined by*

$$\psi(x) = \begin{cases} 0, & x \notin (0,1), \\ 2x, & x \in [0, 1/2], \\ 2(1-x), & x \in [1/2, 1]. \end{cases}$$

For $i = 0, \cdots, \infty$ and $k = 0, \cdots, 2^i - 1$ define $\psi_{i,k}$ by

$$\psi_{i,k}(x) = \psi(2^i x - k).$$

Consider the collection of functions $\mathcal{X} = \{1, x, \psi_{0,0}, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \cdots\}$. Sketch the first five functions of $\mathcal{X}$ on $[0,1]$. Show that $\mathcal{X}$ is a Schauder basis of $C([0,1])$.

Exercise 6. *Let H be a Hilbert space and let $P, Q : H \to H$ be two projections. In the following, set $C = i[P,Q] = i(PQ - QP)$. We also assume that P and Q are* orthogonal, *i.e. $P^* = P$ and $Q^* = Q$.*

a) *Find explicit examples of P and Q which do not commute. Make sure to give clear definitions of both them and H. Sketch a geometric characterization of when this happens.*

b) *Show that* $0 \leq \langle v, Pv \rangle \leq \|v\|^2$ *for all* $v \in H$.

c) *For any* $v \in H$, *set* $\bar{P}(v) = \langle v, Pv \rangle / \|v\|^2$, $\bar{Q}(v) = \langle v, Qv \rangle / \|v\|^2$, $\sigma_P^2(v) = \langle v, (P - \bar{P}\, I)^2 v \rangle$ *and* $\sigma_Q^2(v) = \langle v, (Q - \bar{Q}\, I)^2 v \rangle$. *Show that* $\sigma_P, \sigma_Q \leq \|v\|/2$. *Here* I *is the identity operator on* H.

d) *If* C *is the (scaled) commutator of* P *and* Q, *show that* $\frac{1}{4} \left|\langle v, Cv \rangle\right|^2 \leq \sigma_P^2(v)\sigma_Q^2(v)$. *Argue that* $\|C\| \leq 1/2$. *Hint: use Cauchy-Schwarz. For the last part, you can use without proof that* $\|C\| = \sup_{\|v\| \leq 1} |\langle v, Cv \rangle|$ *(this follows from the fact that* C *is Hermitian).*

Some General Principles of Approximation Theory

In the following, we take X to be a Banach space. If A is a nonempty subset of X, we set

$$d(v, A) := \inf_{u \in A} \|v - u\|.$$

Notation 1. *Given $v \in X$, let $P_A(v)$ denote the set of nearest points to v in A, i.e. the set of best approximations to v in A.*

$$P_A(v) := \{u_0 \in A \mid \|v - u_0\| = d(v, A)\}.$$

Definition 6. *A set A is called an* existence set *if $P_A(v)$ is nonempty for all $v \in A$, and a* uniqueness set *if the cardinality of $P_A(v)$ is always at most one. A is called a* Chebyshev set *if it is an existence and uniqueness set.*

Clearly, existence sets are closed, but closed does not imply existence.

Example 1. *Set $X = \ell_2$, $A = \{(1 + 1/j)e_j\}_1^\infty$ where e_j denotes the jth canonical basis vector. This is closed, but not an existence set.*

Example 2. *Set $X = \ell_\infty$, $A = \{v \in \ell_\infty \mid \|v\|_\infty \leq 1\}$. This is an existence set, but not a uniqueness set.*

Example 3. *Set X to be a normed vector space and A any finite dimensional linear subspace. Then A is an existence set. This follows from a theorem by F. Riesz that says that if $v \in X$, then $B_{2\|v\|} \cap A$ is compact. B_r here denotes the ball of radius at r centered at the origin.*

Lemma 4. *If A is convex, then $P_A(v)$ is always convex.*

Proof. Given $v \in X$ suppose u_0 and u_1 in $P_A(v)$. For any $\lambda \in [0, 1]$ set $u_\lambda = \lambda u_0 + (1 - \lambda)u_1$. Then

$$\begin{aligned} \|v - u_\lambda\| &= \|\lambda(v - u_0) + (1 - \lambda)(v - u_1)\| \\ &\leq \lambda\|v - u_0\| + (1 - \lambda)\|v - u_1\| \\ &= d(v, A). \end{aligned}$$

□

In order to get more control on existence and uniqueness, we can introduce some additional geometric assumptions on the space X. Let $S = \partial B$ where B is the unit ball in X.

- X is *strictly convex* if for all $u, v \in S$ with $u \neq v$, $\|u+v\| < 2$.
- X is *uniformly convex* if, given $\epsilon > 0$, there is a $\delta = \delta(\epsilon)$ such that for any $u, v \in S$ with $\|u+v\| > 2-\delta$ then $\|u-v\| < \epsilon$.
- X is *smooth* if for each $v \in S$, there is a unique supporting functional (i.e. $\ell \in X^*$ with $\|\ell\| = 1$ and $\ell(v) = \|v\|$).

As a counterexample, consider the space of $\{\alpha_j\}$ equipped with the norm

$$\|\{\alpha_j\}\| := \left(\sum_{j=1}^{\infty} \frac{1}{j^2}|\alpha_j|^2\right)^{1/2} + \sum_{j=1}^{\infty}\left(1-\frac{1}{j}\right)|\alpha_j|$$

Lemma 5. *If M is convex and the space is strictly convex, then M is a uniqueness set.*

Remark 2. *Note that $C(X)$ is not strictly convex (exercise: prove this). To characterize uniqueness in this space, other tools are required.*

A geometric criterion for best approximations

Using geometric tools from functional analysis, we now formulate a general criterion for best approximations.

Theorem 7. *Suppose M is a convex set in a Banach space X, and $v \in X$, $v \notin \overline{M}$. Then u_0 is a best approximation to v in M if and only if there is a linear functional $\ell \in X^*$ satisfying the following conditions:*

i) $\|\ell\|_{X^*} = 1$,

ii) $\ell(v-u_0) = \|v-u_0\|$

iii) $\Re\,\ell(u-u_0) \leq 0$ *for all* $u \in M$.

Proof. The main ingredient of the proof in the forward direction will be the geometric Hahn-Banach theorem, a simplified version of which states:

> *Suppose A and B are disjoint non-empty convex subsets of a Banach space X. Moreover, suppose A is open. There exists a bounded linear functional $\ell \in X^*$, and a real number α such that*
>
> $$\Re\,\ell(u) < \alpha \leq \Re\ell(v)$$
>
> *for all $u \in A$ and $v \in B$.*

Now, we suppose u_0 is a best approximation to v from M. We set $r = \|u_0 - v\|$ and let $\mathring{B}_r(v)$ denote the interior of the ball of radius r in X which is centered at v. Then $\mathring{B}_r(v)$ and $\overline{M}$ are disjoint and so, by the geometric Hahn-Banach theorem, there exists an $\tilde{\ell} \in X^*$

and $\alpha \in \mathbb{R}$ with $\Re\,\tilde{\ell}(\mathring{B}_r(v)) > \alpha$ and $\Re\tilde{\ell}(\overline{M}) \leq \alpha$. In particular, $\Re\tilde{\ell}(B_r(v)) \geq \alpha$.

Now we set $\beta = \Re\tilde{\ell}(v - u_0)$ and $\ell = \frac{r}{\beta}\tilde{\ell}$. Clearly, by linearity

$$\Re\,\ell(v - u_0) = \|v - u_0\|.$$

This is a start, but we still need to: a) prove boundedness, b) remove the '$\Re$' from the previous equality. We argue by contradiction: suppose $\|\ell\| > 1$. Then, there exists $w \in B_r(0)$ with $\ell(w) > r$. Setting $z = v - w$, we observe that $z \in B_r(v)$, but

$$\Re\,\tilde{\ell}(z) = \Re\,\tilde{\ell}(v - u_0) + \Re\,\tilde{\ell}(u_0) - \Re\,\tilde{\ell}(w) \leq \beta + \alpha - \frac{\beta}{r}\ell(w) < \alpha.$$

This yields a contradiction, since $z \in B_r(v)$ implies that $\Re\,\tilde{\ell}(z) \geq \alpha$. Thus, $\|\ell\| = 1$ and hence $\ell(v - u_0) = \|v - u_0\|$.

In the other direction, assume such an ℓ exists. Then for any $u \in M$,

$$\|u - v\| \geq \Re\,\ell(v - u) = \Re\,\ell(v - u_0) - \Re\,\ell(u_0 - u) \geq \|v - u_0\|.$$

□

Remark 3. *This can be modified to allow for the choice of a different functional for each element of M. In this form it is called the* generalized Kolmogorov criterion.

Example 4. *$X = H$ a Hilbert space. A point u_0 in a convex set M is a best approximation to $v \in H$ if and only if*

$$\Re\langle v - u_0, u - u_0\rangle \leq 0,$$

for any $u \in M$.

Example 5. *$X = L_p(X, \mu)$, $1 < p < \infty$. If M is a* subspace *of X, then $u_0 \in M$ is a best approximation to $v \in L_p(X, \mu)$ if and only if*

$$\int_X |v(x) - u_0(x)|^{p-1}(\overline{u}(x) - \overline{u_0}(x)) = 0$$

for all $u \in M$.

In the next section we dive into the theory for approximation of continuous functions in a little more detail.

Haar subspaces and continuous functions

For spaces of continuous functions equipped with the supremum norm, *Haar subspaces* give a convenient characterization of when a unique best approximation to a given function exists in a subspace.

Definition 7. *A Haar subspace M is an n-dimensional subspace of $C(\Omega)$ such that any $u \in M$ has at most $n-1$ zeros in Ω.*

Lemma 6. *M is an n-dimensional Haar subspace if and only if for any n points $x_1, \ldots, x_n \in \Omega$ and $\beta_1, \ldots, \beta_n \in \mathbb{C}$ the interpolation problem:*

$$\text{Find } u \in M \text{ such that } u(x_i) = \beta_i,\ i = 1, \ldots, n$$

always has a solution.

Remark 4. *This is equivalent to saying that given any basis $u_1, \ldots, u_n$, the $n \times n$ matrix Φ defined by $(\Phi)_{i,j} = u_j(x_i)$ is invertible for any distinct points $x_1, \ldots, x_n$.*

A natural question is: who cares? The answer is given in the following theorem (which we won't prove here).

Theorem 8 (Haar's uniqueness theorem). *If Ω is locally compact, then a finite dimensional subspace M of $C(\Omega)$ is a Haar subspace if and only if every $f \in C(\Omega)$ has a unique best approximation in M, i.e. M is a Chebyshev set.*

Unfortunately, as it turns out, having high-dimensional Haar subspaces is somewhat difficult.

Theorem 9 (Mairhuber-Curtis theorem). *Suppose $\Omega \subseteq \mathbb{R}^d$, $d \geq 2$ contains an interior point. Then there is no Haar subspace of $C(\Omega)$ of dimension $n \geq 2$.*

Proof. Given a basis $u_1, \ldots, u_n$ consider the function $d(x_1, \ldots, x_n) = \det(u_j(x_i))$. Since the u_j are continuous, so is d. Consider two points x_1 and x_2 lying in a ball $B_r(x_*) \subset \Omega$ for some $r > 0$ and $x_* \in \Omega$. Note that the existence of such a ball is guaranteed by the condition that Ω have non-empty interior. Let $x_3, \ldots, x_n$ be arbitrary in Ω. Consider two non-intersecting continuous paths $\gamma_{1,2} : [0,1] \to \Omega$ with $\gamma_1(0) = x_1, \gamma_1(1) = x_2, \gamma_2(0) = x_2, \gamma_2(1) = x_1$ chosen so that both γ_1 and γ_2 do not pass through $x_3, \ldots, x_n$. Then, since $d(x_1, x_2, \ldots, x_n) = -d(x_2, x_1, \ldots, x_n)$ and $d(\gamma_1(t), \gamma_2(t), x_3, \ldots, x_n)$ is continuous as a function of t, there exists a $t_* \in [0,1]$ for which d vanishes. □

In fact this has deeper implications. It shows that for interpolation in dimensions $d \geq 2$, one cannot in general used a fixed basis for arbitrary scattered data. We will see more about Haar spaces later in these notes when we discuss polynomial approximations of continuous functions.

Interpolation in finite dimensional subspaces

In the last section, we reached a somewhat depressing conclusion: colloquially, except in one dimension given a fixed basis there always exists a set of points for which the interpolation problem is not solvable. But, we could take the opposite perspective: given a finite dimensional subspace, find points $x_1, \ldots, x_n$ which are good for interpolation.

Let us be a bit more precise. Let S be an arbitrary set with $|S| \geq n$ and $u_1, \ldots, u_n : S \to \mathbb{C}$ be bounded and linearly independent. Moreover, suppose $u : S \to \mathbb{C}$ is defined by

$$u(x) = \sum_{j=1}^{n} \alpha_j u_j(x)$$

for some (unknown) coefficients α_j. For interpolation our goal is the following:

Given points $x_1, \ldots, x_n \subseteq S$ and samples $u^{(1)} = u(x_1), \ldots, u^{(n)} = u(x_n)$, find functions $v_j : S \to \mathbb{C}$ such that

$$u(x) = \sum_{j=1}^{n} v_j(x) u^{(j)}.$$

Clearly, a necessary and sufficient condition is

$$d(x_1, \ldots, x_n) := \det(u_j(x_i)) \neq 0.$$

Also, clearly for generic problems many choices of $x_1, \ldots, x_n$ are possible. Are some better than others computationally?

Idea 1: We could try to choose sample points $x_1^*, \ldots, x_n^*$ such that

$$(x_1^*, \ldots, x_n^*) = \text{argmin}_{(x_1,\ldots,x_n)\in S^n} \kappa(u_j(x_i))$$

if possible. Here κ denotes the condition number of the matrix with (i,j)th entry $u_j(x_i)$. Of course many related conditions are possible. This approach tries to make the *coefficient recovery problem* as stable as possible. Note that even with optimal points, this might still be very poorly conditioned! For example, choosing $u_j(x) = 1 + \epsilon \sin(j\pi x)$. Though this seems rather pathological, situations like this arise frequently in applications. This is also somewhat roundabout since we only need v_j, though of course we could find it by setting

$$v_j(x) := \sum_{i=1}^{n} A_{i,j}^{-1} u_i(x),$$

where $A_{i,j} = u_j(x_i)$.

Idea 2: Using the previous definition of A it is easy to show that for any $x \in S$,

$$\sum_{j=1}^{n} A_{i,j} v_j(x) = u_i(x), \quad i = 1, \ldots, n.$$

Cramer's rule then gives

$$v_j(x) = \frac{d(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_n)}{d(x_1, \ldots, x_n)}.$$

The interpolation will be numerically unstable if $\|v_j\|_\infty \gg 1$. In particular, oscillations in signs of the v_j will lead to numerical catastrophic cancellation. A natural goal then is to try to make the norms of the v_j small.

We first claim that

$$B := \sup_{(x_1,\ldots,x_n) \in S^n} d(x_1, \cdots, x_n) < \infty$$

and $B > 0$.

Next, we claim that for any $0 < \epsilon < 1$ there exist points $x_1^*, \ldots, x_n^*$ with

$$d(x_1^*, \ldots, x_n^*) \geq B \max(1/2, 1 - \epsilon/2).$$

Choosing these points, and using our formula for v_j, together with the identity

$$\frac{B}{d} - 1 \leq \epsilon,$$

we find that

$$|v_j(x)| \leq \frac{B}{d} \leq 1 + \epsilon.$$

Remark 5. *For continuous functions we can choose $\epsilon = 0$.*

Two fundamental approximation problems

In the last two sections we have seen two fundamental problems in approximation theory:

- The "measurement" or "sampling" problem: given the value of a function at a collection of points, approximate its value at another location. Colloquially,

  given points, choose functions

- The "experimental" or "design" problem: given a collection of functions, choose points which allow stable interpolation. Colloquially

  given functions, choose points

References and Further Reading

General definitions and Kolmogorov criterion: *Nonlinear Approximation Theory* by D. Braess
Haar subspaces and Mairhuber-Curtis: *Scattered Data Approximation* by H. Wendland
"Idea 2": *On interpolation and integration in finite-dimensional spaces of bounded functions* by P-G. Martinsson, M.Tygert, and V. Rokhlin.

Exercises

Exercise 7. *Assume that u_1 and u_2 are two best L_1 approximations to f from a convex set M. Moreover, assume that all three functions are continuous. Show that*

$$\operatorname{sgn}(f-u_1)(x) = \operatorname{sgn}(f-u_2)(x),$$

for all $x \in X$. Here we assume that f and all the functions in M are real-valued.

Exercise 8. *Prove or give a counter-example: every subspace of a Haar subspace is a Haar subspace.*

Exercise 9. *Show that if Ω contains a subset homeomorphic to the letter 'Y' then $C(\Omega)$ cannot contain a Haar subspace of dimension greater than one.*

Exercise 10. *Given $x_1, \ldots, x_n \in \mathbb{R}$, find an expression for the determinant of the $n \times n$ matrix V with (i,j)th entry x_i^{j-1}. Use this to show that the monomials form a Haar subspace. Hint: argue that the determinant is a polynomial in each x_i and the total degree is at most $n(n-1)/2$. Show that $(x_1 - x_2)$ must be a factor of the determinant and use this to guess a general form for the determinant.*

Exercise 11. *Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ and consider the set of functions $u_j : \mathbb{R}_+ \to \mathbb{R}$ with $u_j(x) = e^{-\lambda_j x}$. Show that the u_j are linearly independent and form a Haar subspace of $C(\mathbb{R}_+)$. Hint: use induction and Rolle's theorem.*

Continuous Functions

So far we have discussed a variety of results related to the existence and uniqueness of best approximations. We have even discussed a little bit about how to represent them and how to use them. One glaring hole in all this is that we so far haven't actually said much about how good these "best approximations" are! Of course, this is intimately related to the existence of "good" bases and the decay of coefficients in those bases.

In this part, we focus on continuous functions and give several results on when a given collection of functions can approximate an arbitrary one.

Our first goal is as follows: prove (and define the terms in) the following result.

Theorem 10 (Stone-Weierstrass[1]). *Let X be a compact metric space and $\mathcal{A}$ be a subalgebra of $C(X)$. If $\mathcal{A}$ separates points in X and vanishes at no point in X then $\mathcal{A}$ is dense in $C(X)$. Here $\mathcal{A}$ and $C(X)$ consist of real-valued functions.*

Before proving the theorem we recall some definitions. Firstly, a vector space $\mathcal{A}$ is an algebra if there is a multiplication operation such that

i) $(fg)h = f(gh)$

ii) $f(g+h) = fg + fh$ where "+" denotes the vector space addition operation

iii) $\alpha(fg) = (\alpha f)g = f(\alpha g)$ for any scalar α.

We say: $\mathcal{A}$ is commutative if $fg = gf$; has an identity if there exists $e \in \mathcal{A}$ with $ef = fe = f$ for all $f \in \mathcal{A}$; and is normed if $\|\cdot\|$ satisfies the usual norm properties *and* $\|fg\| \leq \|f\|\,\|g\|$. If $\mathcal{A}$ is normed and complete then it is a Banach algebra.

We say that $\mathcal{A}$ *separates points* if for any $x, y \in X$, with $x \neq y$ there is an $f \in \mathcal{A}$ with $f(x) \neq f(y)$.

We now turn to the proof.

Proof. The proof consists of three parts:

[1] Marshall H. Stone was chair of the mathematics department of the University of Chicago from 1946-1952. He was partly responsible for bringing André Weil, Antoni Zygmund, Saunders MacLane, Shiing-Shen Chern, Paul Halmos, Irving Segal and Edwin Spanier to Chicago. This period is sometimes called the "Stone age".

1. First, we show that for all $f \in \mathcal{A}$, $|f| \in \overline{\mathcal{A}}$. From this we deduce that for any $f_1, \dots, f_n \in \mathcal{A}$, the function $\max\{f_1, \dots, f_n\} \in \overline{A}$.
2. Next, we show that for any $\epsilon > 0$, for any $f \in C(X)$, and for any $x \in X$, there exists a function $g_x \in \overline{\mathcal{A}}$ such that $g_x(x) = f(x)$, and
$$g_x(y) > f(y) - \epsilon, \quad \forall y \in X.$$
3. Finally, we show that using a finite number of the g_x functions constructed in the previous step, we can construct a function $g \in \overline{\mathcal{A}}$ with $|f(y) - g(y)| < \epsilon$ for all $y \in X$.

We sketch some details of these steps below.

1: Given $w \in \mathcal{A}$ we wish to show that $|w|$ can be approximated from within $\mathcal{A}$. Here we use a neat idea. Given any polynomial p, since $\mathcal{A}$ is an algebra, $wp(w) \in \mathcal{A}$. Thus, it suffices to find a polynomial $x\,p(x)$ which approximates $|\cdot|$ on $[-\|w\|_\infty, \|w\|_\infty]$. Without loss of generality we set $\|w\| = 1$. In particular, the problem has been reduced from finding an approximation from a continuous function defined on an *arbitrary* compact metric space from a general algebra, to approximating a *specific* function (absolute value) from a *specific* algebra (namely polynomials restricted to the interval $[-1, 1]$).

We begin by observing that for any $s \in [-1, 1]$,
$$|s| = \sqrt{1 - (1 - s^2)} = \sum_{n=0}^{\infty} (1 - s^2)^n (-1)^n \binom{\frac{1}{2}}{n},$$
which converges uniformly on $[-1, 1]$.

Letting q_N denote the N^{th} partial sum, then we have that for any $\epsilon > 0$ there exists an N such that $|q_N(s) - |s|| < \epsilon$ for any $s \in [-1, 1]$. Now, it follows that $|q_N(0)| < \epsilon$ and so $|q_N(s) - q_N(0) - |s|| < 2\epsilon$. Finally, since $q_N(s) - q_N(0) = sp_N(s)$ for some polynomial p_N, we have that $q_N(w) - q_N(0) \in \mathcal{A}$.

2: For any x, y with $x \neq y$ there exists an $h_{y,x} \in \mathcal{A}$ with $h_{y,x}(x) = f(x)$, $h_{y,x}(y) = f(y)$. To see this, we note that $\tilde{A} := \{(g(x), g(y)), g \in \mathcal{A}\}$ is a subalgebra of $\mathbb{R}^2$. It is easy to argue that it must in fact be all of $\mathbb{R}^2$, since $\mathcal{A}$ separates points and vanishes at no point in X.

Now, we argue by compactness that there are a finite number of y's such that
$$g_x(t) := \max\{h_{y_1,x}(t), \dots, h_{y_n,x}(t)\} \geq f(t) - \epsilon, \quad \forall t \in X.$$

3: Finally, we argue by compactness of X that there are a finite number of x's such that
$$\min\{g_{x_1}(t), \dots, g_{x_n}(t)\} \leq f(t) + \epsilon, \quad \forall t \in X.$$

□

Remark 6. *For the complex case, $\mathcal{A}$ should be self-conjugate (i.e. if $f \in \mathcal{A}$ then $\overline{f} \in \mathcal{A}$).*

Corollary 1. *The set of all functions $f(x,y) = f(x)g(y)$ with $f \in C(X), g \in C(Y)$ is dense in $C(X \times Y)$.*

Corollary 2. *If K is a compact subset of $\mathbb{R}^n$, then the set of all n-variate polynomials is dense in $C(K)$.*

This is great in that it gives us a (sort of) constructive way of obtaining an approximation of arbitrary fidelity. For polynomials (particularly in one dimension) it is overkill and the construction is a bit clunky. For polynomials on $[0,1]$ we will now give a few alternate proofs. Apart from being more constructive, they highlight interesting techniques that are frequently used in other contexts.

Theorem 11 (Weierstrass approximation theorem)**.** *For any $f \in C[0,1]$ and $\epsilon > 0$ there exists a polynomial p such that $\|f - p\| < \epsilon$.*

proof (Bernstein): For any bounded f on $[0,1]$ define the *Bernstein polynomial of f* by

$$B_n(f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k}, \quad 0 \le x \le 1.$$

In particular,

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

Also,

$$\begin{aligned} B_n(x)(x) &= \sum_{k=0}^{n} \binom{n}{k} \frac{k}{n} x^k (1-x)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1} x^k (1-x)^{n-k} \\ &= \sum_{k=0}^{n-1} \binom{n-1}{k-1} x^{k+1} (1-x)^{n-1-k} \\ &= x. \end{aligned}$$

Finally,

$$\begin{aligned} B_n(x^2)(x) &= \sum_{k=0}^{n} \binom{n}{k} \frac{k^2}{n^2} x^k (1-x)^{n-k} \\ &= \frac{s}{n}\frac{\partial}{\partial s}\frac{s}{n}\frac{\partial}{\partial s}(s+t)^n\Big|_{s=x,t=1-x} \quad = s(s+t)^{n-1}\Big|_{s=x,t=1-x} + \frac{n-1}{n}s^2(s+t)^{n-2}\Big|_{s=x,t=1-x} \\ &= x + x^2\left(1 - \frac{1}{n}\right) \\ &= x^2\left(1 - \frac{1}{n}\right) + \frac{x}{n}. \end{aligned}$$

So, B_n is exact on 1 and x and converges like $1/n$ for x^2.

Continuing, given a $\delta > 0$ we let F denote the set of $k \in \{0, \ldots, n\}$ for which

$$\left| \frac{k}{n} - x \right| \geq \delta.$$

Then

$$\begin{aligned}
\sum_{k \in F} \binom{n}{k} x^k (1-x)^{n-k} &\leq \frac{1}{\delta^2} \sum_{k \in F} \binom{n}{k} \left(\frac{k}{n} - x \right)^2 x^k (1-x)^{n-k} \\
&\leq \frac{1}{\delta^2} \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n} - x \right)^2 x^k (1-x)^{n-k} \\
&\leq \frac{1}{\delta^2} \left[x^2 B_n(1)(x) - 2x B_n(x) + B_n(x^2)(x) \right] \\
&= \frac{1}{\delta^2} \left[x^2 - 2x^2 + x^2 \left(1 - \frac{1}{n} \right) + \frac{x}{n} \right] \\
&= \frac{x - x^2}{n\delta^2} \\
&\leq \frac{1}{4n\delta^2}.
\end{aligned}$$

Then,

$$\begin{aligned}
|f(x) - B_n(f)(x)| &= \left| f(x) - \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n} \right) x^k (1-x)^{n-k} \right| \\
&= \left| \sum_{k=0}^{n} \binom{n}{k} \left(f(x) - f\left(\frac{k}{n} \right) \right) x^k (1-x)^{n-k} \right|.
\end{aligned}$$

Now, for any $\epsilon > 0$ there exists a δ such that $|f(x) - f(y)| < \epsilon$ whenever $|x - y| < \delta$. Thus, with that choice of δ,

$$\begin{aligned}
|f(x) - B_n(f)(x)| &\leq \sum_{k \in F} 2\|f\| \binom{n}{k} x^k (1-x)^{n-k} + \epsilon \sum_{k \notin F} \binom{n}{k} x^k (1-x)^k \\
&\leq \frac{2\|f\|}{4n\delta^2} + \epsilon.
\end{aligned}$$

Choosing $n > \|f\|/(2\delta^2)$, we see that

$$\|f - B_n(f)\|_\infty \leq 2\epsilon.$$

□

This has the following interpretation. Suppose that one has a biased coin with probability of heads equal to x. At each turn, you take a step to the right if heads and stay put if tails. Then the probability you are at "k" after n turns is given by

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

Thus,

$$B_n(f)(x) = \mathbb{E}\left[f\left(\frac{k(n,x)}{n}\right)\right] = \mathbb{E}\left[f\left(\frac{\sum_{i=1}^n X_i}{n}\right)\right]$$

where the X_i are independent Bernoulli random variables with $P(X_i = 1) = x$ and $P(X_i = 0) = 1 - x$. Convergence implies that

$$f(x) = \lim_{n\to\infty} \mathbb{E}[f(\bar{X})].$$

We now turn to another proof, due to Weierstrass.

proof (Weierstrass). Extend f to $\tilde{f}$, continuous on all of $\mathbb{R}$ with compact support. Solve the heat equation

$$\frac{\partial}{\partial t} u(x,t) = \Delta u(x,t)$$

with initial data $\tilde{f}$. It can be shown that $\lim_{t\to 0} u(x,t) = f(x)$ for all $x \in [0,1]$. Now,

$$u(x,t) = \int_{\mathbb{R}} \frac{e^{-(x-y)^2/4t}}{\sqrt{4\pi t}} \tilde{f}(y)\, \mathrm{d}y.$$

For $t > 0$, this is entire in x and hence has a uniformly convergent Taylor series on $[0,1]$. □

We conclude our proofs with one attributed to Landau.

proof (Landau). Let $f \in C[0,1]$ and consider $\tilde{f} = f - [f(0) + x(f(1) - f(0)]$. We note that $\tilde{f}(0) = \tilde{f}(1) = 0$ and $\tilde{f}$ differs from f by a polynomial. We now extend $\tilde{f}$ to all of $\mathbb{R}$, setting $\tilde{f} \equiv 0$ outside of $[0,1]$. Set

$$L_n(x) = c_n \int_{-1}^{1} \tilde{f}(x-t)(1-t^2)^n \, \mathrm{d}t$$

where c_n is a normalization constant chosen so that

$$c_n \int_{-1}^{1} (1-t^2)^n \, \mathrm{d}t = 1.$$

Making the change of variables $s = x - t$ in the definition of L_n, we see that

$$L_n(x) = c_n \int_{x-1}^{x+1} \tilde{f}(s)(1-(x-s)^2)^n \, \mathrm{d}s.$$

Now, since $\tilde{f}$ vanishes outside of $[0,1]$, for $x \in [0,1]$,

$$L_n(x) = c_n \int_{0}^{1} \tilde{f}(s)(1-(x-s)^2)^n \, \mathrm{d}s$$

which is a polynomial in x. Moreover,

$$\psi_n(t) := c_n(1-t^2)^n \chi_{[-1,1]}$$

is an approximate identity, and hence $\tilde{f} \star \psi_n \to \tilde{f}$ as $n \to \infty$. More concretely, for $|t| < 1/\sqrt{n}$,

$$(1-t^2)^n \geq 1 - nt^2$$

and hence $c_n < \sqrt{n}$. For $0 < \delta < 1$,

$$c_n \int_{[-1,1]\setminus(-\delta,\delta)} (1-t^2)^n \,\mathrm{d}t \leq 2c_n(1-\delta^2)^n \to 0$$

as $n \to \infty$. So,

$$\begin{aligned} \left|L_n(x) - \tilde{f}(x)\right| &= c_n \left| \int_{-1}^{1} [\tilde{f}(x-t) - \tilde{f}(x)]\,(1-t^2)^n \,\mathrm{d}t \right| \\ &\leq c_n \int_{-\delta}^{\delta} (1-t^2)^n |\tilde{f}(x-t) - \tilde{f}(x)| \,\mathrm{d}t \\ &\quad + c_n \int_{[-1,1]\setminus(-\delta,\delta)} (1-t^2)^n 2\,\|\tilde{f}\|_\infty \,\mathrm{d}t. \end{aligned}$$

If we choose δ small enough so that $|\tilde{f}(x-y) - \tilde{f}(x)| < \epsilon$ for all $|y| < \delta$ and $x \in [0,1]$ and choose n large enough so that $2\|f\|_\infty \sqrt{n}(1-\delta^2)^n < \epsilon/2$, then we obtain

$$|L_n(x) - \tilde{f}(x)| \leq 2\epsilon, \quad \forall x \in [0,1].$$

□

There is a common theme in these last two proofs. Convolve f with a "nice" function. Prove that the new function satisfies the required property. Pass to the limit to obtain the required result for the original function.

The Weierstrass (or Stone-Weierstrass) approximation theorem guarantees that given a continuous function, we can find an arbitrarily close polynomial approximation. In principle one could back out precise bounds on convergence depending on the regularity of f (say for Hölder spaces) and even back out an algorithm of sorts for constructing these approximations. In practice, they can be far from optimal. In the next chapter we begin our quest to rectify this deficiency by first trying to establish useful properties of best polynomial approximations.

References and Further Reading

Stone-Weierstrass: *A Short Course on Approximation Theory* by N.L. Carothers
Weierstrass proofs: *ASCAT* by Carothers and *ATAP* by L.N. Trefethen

Additional exercises

Exercise 12. *In this problem we will explore weighted spaces. In the following, assume f is a continuous function on an interval $[a, b]$ and w is a positive continuous weight function on $[a, b]$. For $p \in [1, \infty]$, let $\|\cdot\|_p$ be defined by*

$$\|f\|_p = \left(\int_a^b |f(x)|^p \, w(x) \, \mathrm{d}x\right)^{1/p}.$$

For $p = 2$ we also define $\langle , \rangle_w$ by

$$\langle f, g\rangle_w = \int_a^b f(x) \, g(x) \, w(x) \, \mathrm{d}x.$$

1. *Show that for any $f \in C[a, b]$, $\|f\|_1 \leq c\|f\|_2$ and $\|f\|_2 \leq c\|f\|$ where*

   $$c = \left(\int_a^b w(t) \, \mathrm{d}t\right)^{1/2}.$$

   Here $\|\cdot\|$ denotes the usual sup *norm for $C[a, b]$.*

2. *Show that polynomials are dense in $C[a, b]$ under all three norms $\|\cdot\|$, $\|\cdot\|_1$, $\|\cdot\|_2$. Show that $C[a, b]$ is not complete under $\|\cdot\|_1$ or $\|\cdot\|_2$.*

Where have all the errors gone?

In this chapter we pick up our story of polynomial approximations, and try to answer the simple question: what are the properties of the best polynomial approximation, and how do we find them? The following remarkable result is attributed to Chebyshev, though special cases were known to Laplace and, before him, to Euler.

Equioscillations

Theorem 12 (Equioscillation)**.** *The space of polynomials of degree at most n on $[-1,1]$, which we denote by P_n, is a Chebyshev set in $C[-1,1]$, i.e. a best approximation always exists and is always unique. Moreover, if f is real, the best approximation p_* is real, and $f - p_*$ equioscillates in at least $n+2$ extreme points. That is to say, there exist points $-1 \leq x_0 < x_1 < \cdots < x_n < x_{n+1} \leq 1$ such that*

$$p_*(x_i) - f(x_i) = -[p_*(x_{i+1}) - f(x_{i+1})], \quad i = 0, \dots, n,$$

and $|p_*(x_i) - f(x_i)| = \|p_* - f\|_\infty$ for $i = 0, \dots, n+1$.

Aside 1. *A set of such points is also sometimes referred to as an* alternant.

Proof. Existence: $0 \in P_n$ so $d(f, P_n) \leq \|f\|$. So, it suffices to consider $S = \{p \in P_N \mid \|p - f\| \leq \|f\|\}$. This is a compact set and $\|p - f\|$ is continuous. So it attains its minimum.

Equioscillation implies optimality: We argue by contradiction. Suppose $f - p$ equioscillates at $x_0 < x_1 < \cdots < x_{n+1}$ and suppose there exists a $q \in P_n$ with $\|f - q\| < \|f - p\|$. In particular, we mean that $|f(x_i) - p(x_i)| = \|f - p\|_\infty$ for all $i = 0, \dots, n+1$. It follows that $q - p = f - p - (f - q)$ alternates in sign at $x_0, \dots, x_{n+1}$ and hence changes sign between each pair of consecutive equioscillation points. Thus $q - p$ has $n+1$ roots and (since they are polynomials of degree at most n) we have $q = p$ which is a contradiction.

Optimality implies equioscillation: We once again argue by contradiction. Suppose p_* is a best approximation but $f - p_*$ equioscillates at

$-1 \leq x_0 < \cdots < x_{k+1} \leq 1, k < n$, and has no alternating sequence which is longer. We are done if we can find a polynomial which is positive near the points x_i at which $f - p_* > 0$ and negative near the remaining x_i for which $f - p_* < 0$. Conceptually, this is easy - simply take a point between each x_i, x_{i+1}, call it s_i and form the polynomial $p(x) = (-1)^k (x - s_0) \dots (x - s_k)$. Here we have assume that $f(x_0) - p_*(x_0) > 0$. For the opposite sign, we multiply p by -1. Then for δ small enough,

$$\|f - (p_* + \delta p)\|_\infty \leq \|f - p_*\|_\infty$$

but $p_* + \delta p \in P_n$, since $p \in P_{k+1}$ and $k < n$. This gives our contradiction.

The only care comes in choosing the s_i and δ. Assuming without loss of generality that $f(x_0) - p_*(x_0) > 0$, the s_i should be chosen so that

$$(-1)^i (f(x) - p_*(x)) > -(1 - \epsilon)\|f - p_*\|_\infty$$

for all $x \in [s_{i-1}, s_i]$, $i = 0, \dots, k+1$ with $s_{-1} = -1$ and $s_{k+1} = 1$. Let $w_i = \max_{x \in [s_{i-1}, s_i]} (-1)^{i-1}(f(x) - p_*(x))$, $i = 0, \dots, k$ and $w = \max\{w_i\}$. Then, setting $\delta \leq 1/(2w)$, $p(x) = (-1)^k (x - s_0) \dots (x - s_k)$, and $\tilde{p} = p_* + \delta p$ gives the required contradiction.

Uniqueness: Suppose p and q are both optimal. Set $r = (p + q)/2$. By the equioscillation characterization, there exist $-1 \leq x_0 < \cdots < x_{n+1} \leq 1$ with $(f - r)(x_i) = (-1)^i \|f - r\|$ for all $i = 0, \dots, n+1$ or $(f - r)(x_i) = (-1)^{i-1}\|f - r\|$ for all $i = 0, \dots, n+1$.

But

$$\begin{aligned} |f(x_i) - r(x_i)| &= \left| \frac{1}{2}(f(x_i) - p(x_i)) + \frac{1}{2}(f(x_i) - q_(x_)) \right| \\ &\leq \frac{1}{2}|f(x_i) - p(x_i)| + \frac{1}{2}|f(x_i) - q(x_i)| \\ &\leq \frac{1}{2}\|f - p\| + \frac{1}{2}\|f - q\|. \end{aligned}$$

Where equality holds if and only if the signs are the same, or either are zero.

Also, $|f(x_i) - r(x_i)| = \|f - p\| = \|f - q\|$ so it follows that

$$|f(x_i) - p(x_i)| = \|f - p\| = \|f - q\| = |f(x_i) - q(x_i)|$$

and $\mathrm{sgn}(f(x_i) - p(x_i)) = \mathrm{sgn}(f(x_i) - q(x_i))$. Thus, $p(x_i) = q(x_i)$ for $i = 0, \dots, n+1$, and hence $p \equiv q$. □

The next theorem gives a lower bound on the error of the best approximation.

Theorem 13 (de la Vallée Poussin)**.** *Given $f \in C[a, b]$, suppose $q \in P_n$ and $f(x_i) - q(x_i)$ alternates in sign at $a \leq x_0 < x_1 < x_{n+1} \leq b$. If $E_n(f)$ denotes the error of the best approximation in P_n then*

$$E_n(f) \geq \min_{i=0,\dots,n+1} |f(x_i) - q(x_i)|.$$

Proof. We argue by contradiction. Setting p_* to be the best polynomial approximation to f in P_n, and suppose that

$$E_n(f) < \min_{i=0,\dots,n+1} |f(x_i) - q(x_i)|.$$

Then, $q - p_*$ has $n+1$ roots which is a contradiction. □

Remez exchange algorithm

The results of the previous section suggest a simple algorithm for finding the best polynomial approximation. We only sketch the details here. As input it takes a function f and an interval $[a, b]$ together with an initial guess for the equioscillation points, call them $x_0^{(0)}, \dots, x_{n+1}^{(0)}$. The idea is to find the equioscillation points of p_* iteratively using the de la Vallée Poisson theorem as a criterion or "monitor function". The output will be the candidate equioscillation points and coefficients in a monomial expansion.

The algorithm proceeds as follows. For each iteration $i = 1, 2, \dots$ we solve the linear system

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n & 1 \\ 1 & x_1 & x_1^2 & \dots & x_1^n & -1 \\ 1 & x_2 & x_2^2 & \dots & x_2^n & 1 \\ 1 & x_3 & x_3^2 & \dots & x_3^n & -1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{n+1} & x_{n+1}^2 & \dots & x_{n+1}^n & (-1)^{n+1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \\ \vdots \\ \\ \alpha_n \\ E \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \\ \vdots \\ \\ f_n \\ f_{n+1} \end{pmatrix}.$$

Here $x_0 = x_0^{(i-1)}, \dots, x_{n+1} = x_{n+1}^{(i-1)}$ are the current guesses for the equioscillation points, $\alpha_0 = \alpha_0^{(i-1)}, \dots, \alpha_n = \alpha_n^{(i-1)}$ are the current guesses for the coefficients, $E = E^{(i)}$ is the current guess for $\|f - p_*\|$ and $f_0 = f(x_0), \dots, f(x_{n+1})$.
Next, we compute

$$e^{(i)}(x) = f(x) - \sum_{j=0}^{n} \alpha_j^{(i)} x^j$$

and find its maximum (in absolute value). We move the adjacent $x_j^{(i-1)}$ for which $e^{(i)}(x_j^{(i-1)})$ has the same sign, to that point. Set $x_k^{(i)} = x_k^{(i-1)}$ for all other points.

We exit when $\max_j |e^{(i)}(x_j^{(i)})| - \min_j |e^{(i)}(x_j^{(i-1)})| < \epsilon$, a prespecified tolerance.

Remark 7. *It is easy to update this algorithm to move multiple points at the same time. Also, rather than solve the coefficients in a monomial bases, one can use other representations / bases.*

This is nice, and very useful, if there is one (or a small number) of functions you wish to approximate, at a fixed polynomial order. With many functions and multiple orders it can quickly become computationally infeasible. In the next chapter we relax the optimality criteria and see if we can get something almost optimal but more flexible.

General equioscillation theorems

Looking at the proof of the equioscillation theorem and the de la Vallée Poussin theorem, we can see that both proofs hinged on the repeated use of the property that a polynomial $p \in P_n$ can have at most n roots. This, however, is a property shared by any n-dimensional Haar subspace. So, a natural question is whether or not we can extend our equioscillation and de la Vallée Poussin theorems to this more general setting. The answer is an emphatic yes.

Theorem 14. *Suppose V is an $(n+1)$-dimensional Haar subspace of $C[-1,1]$. Given $f \in C[-1,1]$, let u_* denote the best approximation to f from V. Then u_* is unique and $f - u_*$ equioscillates in at least $n+2$ extreme points.*

Proof. Equioscillation implies optimality: This is exactly identical to the case of polynomials.

Optimality implies equioscillation: The intuition is the same as for polynomials, though we have to work a bit harder. Again we argue by contradiction. Suppose u_* is a best approximation, but $f - u_*$ has an alternant of length $k+1$, $-1 \le x_0 < \cdots < x_{k+1} \le 1, \quad k < n$, and no longer alternant exists. Set $E = \|f - u_*\|_\infty$.

Without loss of generality, suppose that $f(x_0) - u_*(x_0) > 0$. Fix $\epsilon > 0$. Since f and u_* are continuous, and no longer alternant exists, there exists an $s_0 \in [x_0, x_1]$ such that $f(x) - u_*(x) > -(1-\epsilon)E$ for all $x \in [-1, s_0]$ and $f(x) - u_*(x) < (1-\epsilon)E$ for all $x \in [s_0, x_1]$. Colloquially, s_0 is chosen between the last time $f - u_* = E$ and the first time $f - u_* = -E$ on the interval $[x_0, x_1]$. Proceeding to the next interval, we select $s_1 \in (x_1, x_2)$ so that $f - u_* < (1-\epsilon)E$ for all $x \in [s_0, s_1]$ and $f - u^* > -(1-\epsilon)E$ on $[s_1, x_2]$. We can continue in this way to obtain $s_0, s_1, \ldots, s_k$. Note that by construction $s_0 \neq -1$ and $s_k \neq 1$.

Now, for polynomials the proof involved constructing a polynomial which changed sign only at $s_0, \ldots, s_k$ and nowhere else. For polynomials this is trivial. For Haar subspaces this is a bit more subtle.

Suppose $k = n-1$. Since V is a Haar subspace, we know that there is a unique function $v \in V$ with $v(s_i) = 0, i = 0, \ldots, n-1$, $v(-1) =$

$\operatorname{sgn}(f(-1) - u_*(-1))$. One remaining concern is as to whether or not the function v changes sign at each of the s_i. By adjusting $v(s_i)$ one can argue that if there is no sign change then it would be possible to create a function in V with more than n roots - a contradiction. The function v so constructed has n roots and hence changes sign only at the s_i.

Suppose $k = n - 2$. We choose v to be the unique function in V such that $v(s_i) = 0, i = 1, \ldots, n-2$, $v(-1) = 1$, and $v(1) = -1$. Once again, the function only changes signs at the s_i.

Suppose $k < n - 2$. For any of the i we can redo the construction of s_i to obtain two new distinct points $s_{i,a}$ and $s_{i,b}$ which satisfy the same conditions required by s_i, and add them to our set. Proceeding in this way, we arrive either at the $k = n - 2$ or $k = n - 1$ case. Letting $\{\tilde{s}_i\}$ denote this new set. As before we construct a $v \in V$ which changes sign at each $\tilde{s}_i$, and nowhere else.

Thus, for any $k < n$ we can construct a $v \in V$ which alternates in sign at each x_i. We can assume that $\operatorname{sgn}(v(x_i)) = \operatorname{sgn}(f(x_i) - u_*(x_i))$ for all $i = 0, \ldots, k$, otherwise we replace v by $-v$. Setting $\hat{u} = u_* - \delta v$ with $\delta < \epsilon/(2\|v\|_\infty)$, we see that $\|f - \hat{u}\|_\infty < \|f - u\|_\infty$, which gives us our contradiction.

Uniqueness: The proof is identical to the polynomial case. □

We also have an analog of the de la Vallé Poussin theorem. The proof is once again identical to the polynomial case.

Theorem 15 (Generalized de la Vallée Poussin). *Given $f \in C[a,b]$, suppose $q \in P_n$ and $f(x_i) - q(x_i)$ alternates in sign at $a \leq x_0 < x_1 < x_{n+1} \leq b$. If $E_n(f)$ denotes the error of the best approximation in P_n then*

$$E_n(f) \geq \min_{i=0,\ldots,n+1} |f(x_i) - q(x_i)|.$$

Best polynomial approximations of 'x^n'

We conclude our discussion of best approximations by returning to polynomials and considering a simple example, due to Chebyshev. The question is:

> *What is the best approximation to x^n in P_{n-1} on the interval $[-1, 1]$?*

Colloquially, the question then is asking to what extent x^n may be considered a polynomial of degree at most $n - 1$ if we measure distance in the supremum norm.

We first note that this is equivalent to finding the monic polynomial in P_n which is closest to 0. Indeed, this is one of the simplest cases of a general class of problems related to finding functions 'least deviating from zero'. It was a particularly popular line of research in the Russian school of approximation theory associated with Chebyshev.

Let $p_* \in P_{n-1}$ denote the best polynomial approximation to x^n and $\hat{p} = x^n - p_*$. Then, by the equioscillation thoerem, there exist $n+1$ points $-1 \leq x_0 < x_1 < \cdots < x_n \leq 1$ with

$$|x_i^n - p_*(x_i)| = \|x^n - p_*(x)\|_\infty = \|\hat{p}\|_\infty$$

and the sequence alternates. It follows that $\hat{p}'(x_i) = 0$ for all $x_i \in (-1,1)$. Since $\hat{p}' \in P_{n-1}$ only has $n-1$ zeros, and there are $n+1$ points x_i, we see that $x_0 = -1$ and $x_n = 1$ (if either were inside the interval, then $\hat{p}'$ would have at least n roots). Thus, for some constant c,

$$\hat{p}'(x) = c(x - x_1)\ldots(x - x_{n-1}).$$

On the other hand, if we set $E = \|\hat{p}\|$, then $q := E^2 - \hat{p}^2$ has double roots at $x_1, \ldots, x_n$ and simple roots at $x = \pm 1$. This last follows from the fact that q has $2n$ roots counting multiplicity and the fact that $E^2 - p^2$ is non-negative on $[-1,1]$. Thus, for some constant $\tilde{c}$,

$$E^2 - \hat{p}^2 = \tilde{c}(1 - x^2)\prod_{i=1}^{n-1}(x - x_i)^2$$

and hence

$$\beta(1 - x^2)(\hat{p}')^2 = E^2 - \hat{p}^2$$

for some constant β. Comparing leading coefficients it is clear that $\beta = 1/n^2$. So,

$$\frac{\hat{p}'}{\sqrt{E^2 - \hat{p}^2}} = \pm\frac{n}{\sqrt{1 - x^2}}.$$

This is a separable ordinary differential equation and can be integrated to obtain

$$\arccos\frac{\hat{p}}{E} = \pm n \arccos x + C,$$

and hence $\hat{p} = E\cos(n \arccos x + C)$. At $x = -1$, $\hat{p} = \pm E$ and so $\pm 1 = \cos(n\pi + C)$ from which it follows that $C = k\pi$ for some $k \in \mathbb{Z}$. Thus,

$$\hat{p} = \pm E \cos(n \arccos x).$$

Now,

$$z^n + \frac{1}{z^n} = \left(z^{n-1} + \frac{1}{z^{n-1}}\right)(z + 1/z) - \left(z^{n-2} + \frac{1}{z^{n-2}}\right)$$

and hence

$$\cos(nt) = 2\cos(t)\cos((n-1)t) - \cos((n-2)t)$$

from which it follows that

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$$

where $T_n(x) = \cos(n \arccos x)$. These are called *Chebyshev polynomials*. From the above identity, it is easy to see that the leading order coefficient of T_n is 2^{n-1}. Thus

$$\hat{p} = \frac{1}{2^{n-1}} T_n$$

and so the error of approximating x^n by a polynomial of degree at most $(n-1)$ is $1/2^{n-1}$.

We can see immediately from the formula that if x_i are the zeros of T_n then

$$x_i = \cos\left(\frac{m\pi}{n} + \frac{\pi}{2n}\right).$$

The previous result has the following nice interpretation. The roots of T_n minimize

$$\max_{x\in[-1,1]} |(x-x_1)\dots(x-x_n)|$$

which is equivalent to minimizing

$$\max_{x\in[-1,1]} \sum_{j=1}^{n} \log|x-x_j|.$$

Now, up to a scaling, log is the Green's function for the Laplace equation in two dimensions. Roughly speaking $-\log|\mathbf{x}-\mathbf{x}'|/(2\pi)$ is the (2D) electrostatic potential due to a unit charged placed at $\mathbf{x}'$. So the Chebyshev roots are the places you should put n unit charges to minimize the maximum electrostatic potential on $[-1,1]$. Obviously one can generalize this problem to more general subsets of $\mathbb{C}$, or even more generally to arbitrary manifolds, as well as adding weight functions.

Finally, we are left with the following additional amusing interpretation: on $[-1,1]$, for large enough n, x^n "is" a polynomial of degree at most $n-1$. Here we mean that for any ϵ there exists an n (logarithmic in ϵ!) for which

$$\inf_{p\in P_{n-1}} \|x^n - p\| < \epsilon.$$

One can extend this to more general orders. The following is due to Rivlin and Newman (1976):

Theorem 16. *Let $k < n$ and set*

$$E_k(x^n) := \max_{p\in P_k} \|x^n - p\|_{L^\infty([-1,1])}.$$

Then

$$E_k(x^n) \le 2e^{-\frac{k^2}{2n}}.$$

We thus obtain a rather shocking result: measured in the supremum norm, x^n is approximately a polynomial of order $\sim c\sqrt{n}$.

References and Further Reading

Equioscillations and de la Vallé Poussin: *ATAP* by L.N. Trefethen
x^n approximation: *ASCAT* by N.L. Carothers.
Remez: *IAF* by T. Rivlin

Additional Exercises

Exercise 13. *For any $n \geq 0$ show that the mapping which takes the function $f \in C[-1,1]$ to its best polynomial approximation $p_* \in P_n$ is continuous with respect to the ∞-norm on $C[-1,1]$. Hint: uniqueness follows from above. Combine with compactness.*

Exercise 14. *In this problem we will explore best approximations on L^2.*

1. *For $n \geq 1$ let p_* denote the best approximation to f from P_{n-1} in the least-squares sense, i.e.*

$$p_* = \text{argmin}_{p \in P_{n-1}} \|f - p\|_2.$$

Show that $\langle f - p_*, p \rangle_w = 0$ for all $p \in P_{n-1}$. Moreover, show that if $q \in P_{n-1}$ and $\langle f - q, p \rangle_w = 0$ for all $p \in P_{n-1}$ then q is the unique best approximation to f from P_{n-1} in the least-squares sense. Hint: for uniqueness, use the parallelogram law applied to $f - q$ and $f - p$.

2. *Show that if p_* is the best least-squares approximation to f then $f - p_*$ has at least n zeros on $[a,b]$. Hint: use the previous part.*

3. *Let $p_{*,2}^{(n)}$ denote the best least-squares approximation to f from P_{n-1} and $p_{*,\infty}^{(n)}$ denote the best uniform approximation. Show that*

$$\|f - p_{*,2}^{(n)}\|_2 \leq \|f - p_{*,\infty}^{(n)}\|_2$$

and hence $\|f - p_{*,2}^{(n)}\|_2 \to 0$ as $n \to \infty$.

4. *If p_n denotes the nth Legendre polynomial, and α_n is its leading coefficient (i.e. the coefficient of x^n) then show that p_n/α_n is the smallest monic polynomial in P_n (measured in the $\|\cdot\|_2$ norm).*

Exercise 15. *A (simplified) Remez algorithm works by doing the following: Choose "control" points $x_0, \ldots, x_{n+1}$ in the interval $[0,1]$ to initialize. For each iteration, form the system*

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n & -1 \\ 1 & x_1 & x_1^2 & \cdots & x_1^n & 1 \\ \vdots & \vdots & \cdots & \ddots & \vdots & \\ 1 & x_{n+1} & x_{n+1}^2 & \cdots & x_{n+1}^n & (-1)^{n+2} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_n \\ E \end{bmatrix} = \begin{bmatrix} f(x_0) \\ \vdots f(x_{n+1}) \end{bmatrix}$$

and solve it. Find the location of the maximum of $e(x) = f(x) - \sum_{j=0}^{n} \alpha_j x^j$ on the interval $[0,1]$. Call it x_*. Move the closest control point x_i (for which the signs of $e(x_*)$ and $e(x_i)$ agree) to x_* and move to the next iteration.

1. *Perform one iteration of the Remez algorithm to compute the best polynomial approximation to x^5 by polynomials of degree at most 3 on the interval $[0,1]$. Start with equispaced nodes $x_0 = 0, \dots, x_4 = 1$.*

2. *Write a code to use the Remez algorithm to compute the best polynomial approximation to x^5 by polynomials of degree at most 3 on the interval $[0,1]$. Plot the equioscillation points and the error as a function of iteration number. Show plots for initialization both with equispaced points and random initial points. Hint: in Matlab, the commands vander, polyval, and roots might be helpful.*

3. *Extend your code in some way (multiple points at once, more general function, more stable formulae for solving the linear system, some different experiments, etc.).*

Exercise 16. *Recall the formula for Chebyshev polynomials:*

$$T_n(x) = \cos(n \operatorname{acos} x).$$

1. *Let $p \in P_n$ be a polynomial given by a finite Chebyshev series*

$$p(x) = \sum_{k=0}^{n} \alpha_k T_k(x)$$

   and let $s \in [-1,1]$. Show that $p(s)$ can be evaluated using the following algorithm. Set $u_{n+1} = 0$, $u_n = \alpha_n$ and

$$u_k = 2su_{k+1} - u_{k+2} + \alpha_k, \quad k = n-1, m-2, \dots, 0.$$

   Then $p(s) = \frac{1}{2}(\alpha_0 + u_0 - u_2)$.

2. *Show that T_n satisfies the ODE $(1-x^2)y'' - xy' + n^2 y = 0$.*

Exercise 17. *In this problem, we will give a brief proof of Theorem 16 (see Sachdeva and Vishnoi 2013).*

a) *For $i = 1, 2, \cdots$, let X_i be the Bernoulli random variable which takes values ± 1 with equal probability. For $n < 0$, define $T_n = T_{-n}$. Prove that*

$$x^n = \mathbb{E}\left[T_{X_1 + \cdots + X_n}(x)\right].$$

   Hint: use induction and the three-term recurrence for Chebyshev polynomials.

b) *One simple version of Hoeffding's inequality states that for $Z_1, \cdots, Z_n$ i.i.d. bounded random variables, $a \le Z_i \le a + L$ almost surely, if S_n denotes their sum, then for all $t > 0$,*

$$\mathbb{P}\left[S_n - \mathbb{E}(S_n) \ge t\right] \le e^{-t^2/(nL^2)}.$$

Using this, bound the probability that $|X_1 + \cdots + X_n| > c\sqrt{n}$ *for any* $c > 0$.

c) *Using (a), (b), and the bound* $\|T_m\|_{L^\infty([-1,1])} \leq 1$ *for all* m, *argue that* x^n *can be written as a linear combination of Chebyshev polynomials up to order* $c\sqrt{n}$ plus *an error term. Give an explicit bound for the error.*

d) *Check this numerically. Take* x^{100} *and compute the Chebyshev coefficients for* $m = 10, 20, 30, 40, 50$. *What is the error for each* m?

Convergence of Chebyshev Polynomial Approximations

Previously, we saw how Chebyshev polynomials arose naturally in the contexts of best approximations to monomials and electrostatics. Here we explore in more detail their utility for approximating more general functions, and take our first steps toward practical and provable algorithms for polynomial approximation and interpolation.

Convergence of Chebyshev series

Our first result establishes the representability of Lipschitz functions on $[-1,1]$ by Chebyshev series.

Theorem 17. *Suppose we are given a function f which is Lipschitz continuous on $[-1,1]$. Then f has a unique representation as a Chebyshev series,*

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

which is absolutely continuous and uniformly convergent. Moreover, the coefficients are given by the following formula,

$$\alpha_k = \frac{2}{\pi} \int_{-1}^{1} \frac{f(x) T_k(x)}{\sqrt{1-x^2}}\, \mathrm{d}x, \quad k > 0,$$
$$\alpha_0 = \frac{1}{\pi} \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, \mathrm{d}x.$$

Proof. The intuition of the proof echoes the construction of the Chebyshev polynomials themselves, and their roots.
Let's map f to the unit circle. Given $z \in \mathbb{C}$, $|z| = 1$, set $x = \frac{1}{2}(z + \frac{1}{z})$. Then $z = e^{i\theta}$ implies $x = \cos\theta$. Then

$$\mathrm{d}x = \frac{1}{2}\left(1 - \frac{1}{z^2}\right) \mathrm{d}z = \frac{i/z}{2i}\left(z - \frac{1}{z}\right) \mathrm{d}z = \pm \frac{i}{z}\sqrt{1-x^2},$$

where we take the $+$ if $\mathrm{Im}(z) \geq 0$ and the $-$ if $\mathrm{Im}(z) \leq 0$.

Next we define the function F by $F(z) = f(x(z))$. If f is Lipschitz, so is F. In fact, the regularity is better near ± 1. In this case, standard results in Fourier analysis (see later chapters) say that F is uniquely expressible as a Laurent series which is absolutely and uniformly convergent on the unit circle.

$$F(x) = \frac{1}{2}\sum_{k=0}^{\infty} a_k \left(z^k + \frac{1}{z^k}\right) = \frac{1}{2}\sum_{k=0}^{\infty} a_k T_k(x),$$

where the first equality follows from the symmetry of F ($z \mapsto 1/z$) and the second follows from making the substitution $= e^{i\theta}$ and the definition of T_k.

The coefficient of z^k ($a_k/2$) can be easily computed using the Cauchy integral formula, to obtain

$$a_k = \frac{1}{i\pi}\int_{|z|=1} z^{-1-k} F(z)\,\mathrm{d}z, \quad k > 0.$$

Similarly, for z^{-k}, for which the coefficient is also $a_k/2$, one finds that

$$a_k = \frac{1}{i\pi}\int_{|z|=1} z^{-1+k} F(z)\,\mathrm{d}z.$$

So, averaging the two expressions,

$$a_k = \frac{1}{2\pi i}\int_{|z|=1} \frac{1}{z}\left(z^k + z^{-k}\right) F(z)\,\mathrm{d}z = \frac{1}{\pi i}\int_{|z|=1} \frac{1}{z} T_k(x(z))\, F(z)\,\mathrm{d}z.$$

Changing variables to x, and noting that the part involving $\mathrm{Im}(z) > 0$ comes in with an extra minus sign, we obtain

$$\begin{aligned} a_k &= \frac{1}{i\pi}\int_{|z|=1} \frac{1}{z} T_n(x(z))\, f(x(z))\,\mathrm{d}z \\ &= \frac{-1}{i\pi}\int_{-1}^{1} T_k(x) f(x) \frac{1}{\sqrt{1-x^2}}\frac{\mathrm{d}x}{i} + \frac{1}{i\pi}\int_{-1}^{1} T_k(x) f(x) \frac{1}{-\sqrt{1-x^2}}\frac{\mathrm{d}x}{i}. \end{aligned}$$

So,

$$a_k = \frac{2}{\pi}\int_{-1}^{1} \frac{T_k(x) f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x, \quad k > 0.$$

The case where $k = 0$ is left as an exercise. □

Chebyshev and Best Approximations

The next theorem gives some further justification for all of this bother about Chebyshev polynomials. It follows by a 1967 result of MJD Powell (thought related results appear elsewhere).

Theorem 18. *Suppose f is bounded and continuous on $[-1, 1]$ and let p_* denote its best polynomial approximation from P_n. Let p denote its truncated Chebyshev expansion (truncated after $n+1$ terms). Then*

$$\frac{\|f - p\|_\infty}{\|f - p_*\|_\infty} \sim \frac{4}{\pi^2}\log n, \quad \text{as } n \to \infty.$$

Proof. We actually start with proving a much more general result. Suppose we are given a finite dimensional subspace M of E, and moreover that this space is equipped with an inner product $\langle\,,\,\rangle_E$.

Given $f \notin M$, let p denote the orthonormal projection of f onto M and p_* denote the best approximation to f in M. We emphasize here that $\langle u, u\rangle_E^{1/2} \neq \|u\|_E$ in general, i.e. that the norm on E is not necessarily the norm induced by the inner product. R need not be complete with respect to $\langle\,,\,\rangle_E$.

The trick boils down to this: $(f-p)-(f-p_*) \in M$. Then

$$f - p = f - p_* - P_M(f-p), \quad \text{with} \quad P_M(u) = \sum_{i=1}^{m} \phi_i \langle \phi_i, u\rangle_E,$$

and hence

$$\|f-p\|_E \leq \|f-p_*\|_E + \|P_M\|_{E\to E}\|f-p_*\|_E,$$

and hence

$$\frac{\|f-p\|_E}{\|f-p_*\|_E} \leq 1 + \|P_M\|_{E\to E}.$$

Remark 8. *Note that all we needed for this was actually the existence of a projection, not necessarily an orthonormal projection. From chapter 2, this can then be applied to any finite dimensional subspace of a Banach space.*

For Chebyshev polynomials, we do have a convenient inner product, defined by

$$\langle u, v\rangle := \int_{-1}^{1} \frac{uv}{\sqrt{1-x^2}}\,\mathrm{d}x,$$

and the orthonormal basis is $\phi_0 = 1/\sqrt{\pi}$, $\phi_k = \sqrt{2/\pi}T_k$, $k > 0$. Thus, in order to use the above result with $E = L^\infty([-1,1])$, $M = P_n$, and the inner product defined above, we need to estimate the operator norm (as a map from $L^\infty \to L^\infty$) of the projection P defined by

$$P(u)(x) = \int_{-1}^{1} \sum_{i=0}^{n} \frac{\phi_i(x)\phi_i(y)}{\sqrt{1-y^2}} u(y)\,\mathrm{d}y.$$

Our bound on $\|P\|$ will mostly consist of truncated geometric series combined with trigonometric identities. And so we begin,

$$\|P\| \leq \sup_x \int_{-1}^{1} \left| \sum_{i=0}^{n} \frac{\phi_i(x)\phi_i(y)}{\sqrt{1-y^2}} \right| \mathrm{d}y.$$

The sum is given explicitly by

$$\sum_{i=0}^{n} \phi_i(x)\phi_i(y) = \frac{2}{\pi}\sum_{k=0}^{n} \cos(k \arccos x)\cos(k \arccos y) - \frac{1}{\pi}.$$

We observe that $\cos(ks)\cos(kt) = \frac{1}{2}[\cos(k(s-t)) + \cos(k(s+t))]$ and

$$\sum_{k=0}^{n} \cos(k\alpha) - \frac{1}{2} = \frac{\sin((n+\frac{1}{2})\alpha)}{2\sin(\alpha/2)}.$$

So,

$$I(s) := \int_{-1}^{1} \left| \sum_{i=0}^{n} \frac{\phi_i(x)\phi_i(y)}{\sqrt{1-y^2}} \right| \mathrm{d}y = \frac{1}{2\pi} \int_{-1}^{1} \left| \frac{\sin((n+\frac{1}{2})(s-t))}{\sin(\frac{s-t}{2})} + \frac{\sin((n+\frac{1}{2})(s+t))}{\sin(\frac{s+t}{2})} \right| \frac{1}{1-y^2} \mathrm{d}y$$

where we have set $s = \arccos x$, and $t = t(y) = \arccos y$. Changing variables to from y to t, we find

$$\begin{aligned} I(s) &= \frac{1}{2\pi} \int_{0}^{\pi} \left| \frac{\sin((n+\frac{1}{2})(s-t))}{\sin(\frac{s-t}{2})} + \frac{\sin((n+\frac{1}{2})(s+t))}{\sin(\frac{s+t}{2})} \right| \mathrm{d}t \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \left| \frac{\sin((n+\frac{1}{2})(s-t))}{\sin(\frac{s-t}{2})} + \frac{\sin((n+\frac{1}{2})(s+t))}{\sin(\frac{s+t}{2})} \right| \mathrm{d}t \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\sin((n+\frac{1}{2})(s-t))}{\sin(\frac{s-t}{2})} \right| \mathrm{d}t \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\sin((n+\frac{1}{2})t}{\sin(\frac{t}{2})} \right| \mathrm{d}t \end{aligned}$$

We break the region of integration up into two pieces, $R_1 = [-\alpha, \alpha]$ and $R_2 = [-\pi, \pi] \setminus R_1$. Here $\alpha \in (0, \pi)$ is left free and will be chosen later.

On R_1, we observe that $|\sin((n+1/2)t)| < (n+1/2)|t|$ and

$$\frac{1}{|\sin(t/2)|} \leq \frac{1}{|t|/2 - |t|^3/(3!8)} = \frac{2}{|t|} + O(\alpha^2).$$

Then,

$$\frac{1}{2\pi} \int_{R_1} \left| \frac{\sin((n+\frac{1}{2})t}{\sin(\frac{t}{2})} \right| \mathrm{d}t \leq \frac{2n+1}{2\pi} \int_{-\alpha}^{\alpha} 1 \, \mathrm{d}t + O(n\alpha^3) = \frac{2n+1}{\pi}\alpha + O(n\alpha^3).$$

Away from R_1, we use the bound $|\sin((n+1/2)t)| < 1$ and hence

$$\frac{1}{2\pi} \int_{R_2} \left| \frac{\sin((n+\frac{1}{2})t}{\sin(\frac{t}{2})} \right| \mathrm{d}t \leq \frac{1}{\pi} \int_{\alpha}^{\pi} \frac{1}{\sin(t/2)} \mathrm{d}t = \frac{2}{\pi} \log \frac{\tan(\pi/4)}{\tan(\alpha/4)} = -\frac{2}{\pi} \log(\alpha/4) + O(\alpha^2 \log(\alpha)).$$

Setting $\alpha = 1/(n+1/2)$ and putting together our bounds on R_1 and R_2, we find

$$\|P\|_{\infty\to\infty} \leq \frac{2}{\pi} \left(1 + \log \left(\frac{n+1/2}{4} \right) \right) + O(n^{-2} \log(n)).$$

□

To recap, we know that for Lipschitz functions, truncated Chebyshev series provide uniformly convergent polynomial approximations. Moreover, they are within a log factor of optimal! It is remarkable, given the nonlinear nature of best approximations, that a non-adaptive linear method should get so close. Now that we know how good they are, let's try to get a more quantitative estimate on the rate of convergence. To do this we will impose more regularity and see how that regularity yields compressibility.

Regularity and decay

We begin our discussion about the interplay between regularity of a function and the decay of its Chebyshev coefficients with the following theorem.

Theorem 19. *For an integer $p \geq 0$, let $f, f', \cdots, f^{(p-1)}$ be absolutely continuous and suppose $f^{(p)}$ is of bounded variation with total variation V. Then, for $k \geq p+1$, the Chebyshev coefficients α_k of f satisfy*

$$|\alpha_k| \leq \frac{2V}{\pi k(k-1)\dots(k-p)} \leq \frac{2V}{\pi(k-p)^{p+1}}.$$

Proof. We begin by recalling that

$$\alpha_k = \frac{2}{\pi}\int_0^\pi f(\cos(\theta))\,\cos(k\theta)\,\mathrm{d}\theta.$$

Set

$$F_{j,n} = \int_0^\pi f^{(j)}(\cos(\theta))\,\cos(n\theta)\,\mathrm{d}\theta.$$

Integration by parts, together with the identity $\sin(x)\,\sin(ax) = \frac{1}{2}\left[\cos((a-1)x) - \cos((a+1)x)\right]$, yields

$$F_{j,n} = \frac{1}{2n}\left[F_{j+1,n-1} - F_{j+1,n+1}\right].$$

In particular,

$$F_{0,k} = \frac{1}{2^{p+1}}\sum_{i_1,\dots,i_{p+1}=1}^{2}(-1)^{p+1+i_1+\cdots i_{p+1}}\frac{F_{p+1,k+(-1)^{i_1}+\cdots+(-1)^{i_{p+1}}}}{k(k+(-1)^{i_1})\times\cdots\times(k+(-1)^{i_1}+\cdots+(-1)^{i_p})}.$$

Let Z_j be the Bernoulli random variable which is ±1 with probability $1/2$ in each case. Let $U_j = \sum_{i=1}^{j} Z_j$. Then,

$$F_{0,k} = \frac{1}{k}\mathbb{E}\left[\frac{(-1)^{p+1+U_{p+1}}F_{p+1,k+U_{p+1}}}{e^{\sum_1^p \log(k+U_i)}}\right].$$

We observe that the denominator inside the expectation is always bounded below by

$$(k-1)\cdots(k-p).$$

Moreover, an elementary estimate of the numerator gives an upper bound of

$$\|f^{(p)}\|_{TV} =: V.$$

Hence,

$$|\alpha_k| = \frac{2}{\pi}|F_{0,k}| \leq \frac{2V}{\pi k(k-1)\cdots(k-p)}.$$

□

Corollary 3. *If f satisfies the conditions of the previous theorem, then*

$$\|f - f_n\| \leq \frac{2V}{\pi(n-p)^p},$$

where f_n is the Chebyshev projection of f obtained by keeping the first $n+1$ terms in the Chebyshev series.

When f is analytic in a neighborhood of $[-1,1]$ this result can be dramatically improved with an elegant proof.

Definition 8. *Consider the "Joukowsky" map $x = \frac{1}{2}(z + 1/z)$. For $\rho > 1$, the image of the circle of radius ρ, under this map, is an ellipse with foci at ± 1. These ellipses are called* Bernstein ellipses *and we will denote their interiors by E_ρ.*

Theorem 20. *Suppose f defined on $[-1,1]$ is analytically continuable to the open Bernstein ellipse E_ρ and $|f(x)| \leq M$ on that ellipse. If α_k denotes its Chebyshev coefficients then*

$$\begin{aligned}|\alpha_0| &\leq M \\ |\alpha_k| &\leq 2M\rho^{-k}.\end{aligned}$$

Proof. The case $k = 0$ is easy. For $k > 0$,

$$\begin{aligned}\alpha_k &= \frac{1}{\pi i}\int_{|z|=1} z^{-1-k}F(z)\,\mathrm{d}z \\ &= \frac{1}{\pi i}\int_{|z|=\rho} z^{-1-k}F(z)\,\mathrm{d}z,\end{aligned}$$

where the last equality follows from Cauchy's integral formula. Hence

$$|\alpha_k| \leq 2\pi\rho M\rho^{-k-1}.$$

□

Corollary 4. *If f satisfies the conditions of the previous theorem,*

$$\|f - f_m\| \leq \frac{2M\rho^{-n}}{\rho - 1}.$$

References and Further Reading

This chapter draws heavily from *Approximation Theory and Approximation Practice* by L.N. Trefethen.

Additional Exercises

Exercise 18.

In this problem we will explore Chebyshev interpolation. Consider the function $f(x) = e^x$ *on the interval* $[-1, 1]$.

1. *Interpolate* f *using* $2, 5, 10, 15$ *and* 20 *Chebyshev polynomials. Do this in two ways: firstly, by computing projections using the equation derived in class for coefficients in a Chebyshev expansion; and secondly, by forming the matrix* $V_{i,j} = T_j(x_i)$, $i, j = 0, \ldots, n$ *and using it to find the coefficients in the Chebyshev expansion. Include plots of the functions (all in the same plot), and errors.*

2. *Write down an explicit formula for the Bernstein ellipse with parameter* ρ *in terms of the real and imaginary parts of* x.

3. *Find a bound on the rate of convergence as a function of* n. *Use the bound based on Bernstein ellipses and choose the* ρ *depending on* n.

Exercise 19. *The Chebyshev polynomials can be expressed in the monomial basis as*

$$T_n(x) = \frac{n}{2} \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^k \frac{(n-k-1)!}{k!(n-2k)!} (2x)^{n-2k}, n > 0.$$

Verify that this gives the correct polynomials for $n = 1$ *and* $n = 2$.

1. *Prove that this expression satisfies the Chebyshev recurrence relation*

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

2. *Implement this formula numerically for* $n = 2, 5, 10, 15, 20, 40$ *and* 80 *and plot the solutions and the errors compared with computing them via the formula* $T_n(x) = cos(n \operatorname{acos}(x))$. *Is the result surprising?*

Exercise 20. *In this problem, we will examine some of the ramifications of approximation theory for linear algebra.*

1. *Write a code to find the best polynomial approximation to* $f(x) = 1/x$ *on the interval* $[1, 5]$ *for* $n = 4$.

2. *Construct a* 1000×1000 *symmetric random matrix* A *with eigenvalues all lying between* $[1, 5]$ *and a random vector* $b \in \mathbb{R}^{1000}$. *Use your answer from part (a) to approximate the solution of the linear system* $Ax = b$. *Give guarantees on the norm of the error (be careful with which norm you are using, though you can use any norm you want, to measure the error). Compute the actual error and compare. Note: the distribution of the random variables you use is up to you, as long as* A *has the right properties.*

3. *Find the coefficients in the (shifted and scaled) Chebyshev expansion for* $1/x$ *on the interval* $[1, 5]$ *for* $n = 5, 10, 20$. *Be careful about the mapping from* $[-1, 1]$ *to* $[1, 5]$ *and back.*

4. *Use the same A and b from part (b). Use your answer from part (c) to approximate the solution of the linear system $Ax = b$. Note: given a matrix M and a vector v you can compute $T_n(M)v$ from $T_{n-1}(M)v$ and $T_{n-2}(M)v$ using the recurrence relation! You do not need to calculate it each time, nor should you compute it using the cosine representation of T_n. What guarantees can you give on the accuracy of your solution?*

Interpolation and its Interpretations

So far we have considered the question of finding a "good" polynomial approximation to a given function f. We have not been overly concerned with computational considerations (finding maxima, integration, and the number of evaluations of f, for example). In this section we (partially) fill this gap. Our setup is the following:

> *Suppose we are given points $x_0, \ldots, x_n \in [-1, 1]$ and samples $f_1, \ldots, f_n$ taken from an unknown function f with certain known properties (Lipschitz,p-times differentiable, continuous, etc.).*

Then, we ask the following questions:

- Given $x \in [-1, 1]$ can we find $\tilde{f} \approx f(x)$?
- How do we construct $\tilde{f}$ from the given data?
- What is the accuracy of $\tilde{f}$ and how does that depend on the locations of the points and the properties of f?
- When can we do this in practice (on a computer with finite precision arithmetic)?
- Can we characterize good points? More broadly, what does it mean to be *good*?

Let's start with polynomials. Why? Well, previously we have found error bounds for polynomial approximation. If we can understand how to (stably!) interpolate polynomials, then we can try to use the fact that for smooth enough functions f, there is a nearby polynomial.

Idea 1: Compute the *Vandermonde* matrix and solve the following linear system

$$\underbrace{\begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix}}_{V} \underbrace{\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}}_{\vec{\alpha}} = \underbrace{\begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}}_{\vec{f}}$$

For each x, set $\delta_x = (1, x, x^2, \dots, x^n)^T$. Then, if $f \in P_n$,

$$f(x) = \delta_x^T \vec{\alpha}.$$

We have seen this idea (in a much more general context) before. There are some immediately visible drawbacks and natural questions. In general, V will be poorly-conditioned, so V^{-1} will magnify errors (rounding, experimental, etc.) by a large factor, leading to large errors in $\vec{\alpha}$. Does this affect the final result? The answer is somewhat surprising and depends on the points x_i. See Shen & Serkh 2023 for a good analysis of this approach.
Cosmetically, it requires computing the coefficients $\vec{\alpha}$ (which are auxiliary) and takes $O(n^2) - O(n^3)$ operations.

Remark 9. *Nothing stops us from using a different basis, besides monomials that is, and re-running the above arguments. In practice this works well, provided that we can find points for which the associated "Vandermonde matrix" $V_{i,j} = \phi_j(x_i)$ is well-conditioned.*

Idea 2: Lagrange interpolation.
We will first go through the naïve approach and then fix it. Construct

$$\ell_i(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}, \quad i = 0, \dots, n.$$

Clearly, ℓ_i is a polynomial of degree n, vanishing at all x_j except for x_i at which it is one. Using this basis, if $f \in P_n$,

$$f(x) = f_0 \ell_0(x) + \cdots + f_n \ell_n(x).$$

To see this, note that both sides are in P_n and agree at $n+1$ points.
A natural question is how stable this is.

Definition 9. *Given $x_1, \dots, x_n$ the* Lebesgue constant Λ *is defined via*

$$\Lambda = \sup_{f \in C([a,b])} \frac{\|p_f\|_{L^\infty}}{\|\vec{f}\|_{\ell^\infty}} = \sup_{f \in C([a,b])} \frac{\|p_f\|_{L^\infty}}{\|f\|_{L^\infty}}$$

where $\vec{f} = (f_0, \dots, f_n)^T$ and

$$p_f = \sum_{j=0}^{n} f_j \ell_j(x).$$

The function

$$\lambda(x) = \sum_{j=0}^{n} |\ell_j(x)|,$$

is called the Lebesgue function. *Note that here we assume the x_i's are in an interval $[a,b]$ and the ∞-norms are taken over that interval. In the first equality, the supremum should be taken over all $\vec{f}$ which do not vanish at the x_j's, and in the second, over all continuous functions $f \neq 0$.*

Proposition 4. *If the x_i are all distinct then $1 \leq \Lambda < \infty$.*

Proof. It is easy to argue that $\|p_f\|$ is a continuous function of $f_0, \ldots, f_n$. By scaling, it suffices to consider $\|\vec{f}\|_\infty \leq 1$. By compactness then, Λ is finite. □

Our next result relates the Lebesgue constant to the best polynomial approximation.

Theorem 21. *Let $x_0, \ldots, x_n \in [a,b]$ and Λ be the Lebesgue constant. Given $f \in C([a,b])$, let p be its Lagrange approximation and p_* be the best polynomial approximation to f in P_n. Then*

$$\|f - p\| \leq (1 + \Lambda)\|f - p_*\|.$$

Remark 10. *We have seen a very similar result before.*

Proof. We could argue by noting that $P : f \mapsto p_f$ is a projection of $C([a,b])$ to P_n and $\|P\|_\infty = \Lambda$. The result then follows the first part of the proof for Chebyshev errors versus best approximation errors.

More concretely, we note that $p - p_*$ is a polynomial. So

$$\begin{aligned} f - p - (f - p_*) &= \sum_{j=0}^{n} ((f-p) - (f-p_*))|_{x_j}\, \ell_j(x) \\ &= -\sum_{j=0}^{n} (f - p_*)|_{x_j}\, \ell_j(x), \end{aligned}$$

and hence

$$\|f - p\| \leq \|f - p_*\| \left(1 + \max_x \sum_{j=0}^{n} |\ell_j(x)|\right) = (1 + \Lambda)\|f - p_*\|.$$

□

Thus the Lebesgue constant is closely tied to how well Lagrange interpolation works at approximating continuous functions, at least relative to their best polynomial approximation.

Theorem 22. *On the interval $[-1,1]$, if Λ_n is the Lebesgue constant for any set of $n+1$ distinct points in $[-1,1]$, then* *a)* $\Lambda_n \geq \frac{2}{\pi}\log(n+1) + \frac{2}{\pi}\left(\gamma + \log\left(\frac{4}{\pi}\right)\right)$.

b) For Chebyshev roots,

$$\Lambda_n \leq \frac{2}{\pi}\log(n+1) + 1.$$

Proof. We do not prove (a) here. See the paper of Erdös (1961) and Brutman (1978).
For (b) we proceed as in Powell. The proof, though it is certainly not optimal, involves combining trigonometric identities with truncated geometric expansions. We begin by recalling that the roots of the Chebyshev polynomial T_{n+1} are

$$x_m = \cos\left(\frac{(m+1/2)}{n+1}\pi\right).$$

Then

$$\sum_{k=0}^{n}{}' T_k(x_j)T_k(x_i) = \sum_{k=0}^{n}{}' \cos\left(k\frac{(j+1/2)}{n+1}\pi\right)\cos\left(k\frac{(i+1/2)}{n+1}\pi\right) = \frac{(n+1)}{2}\delta_{i,j}.$$

We leave the last equality as an exercise. Here the prime is used to denote that the first term in the sum should be halved.

Thus,

$$\sum_{k=0}^{n}{}' T_k(x_j)T_k(x) = \frac{2}{(n+1)}\ell_j(x).$$

Now, setting $\theta_j = \arccos(x_j)$ and $\theta = \arccos(x)$,

$$\begin{aligned}\sum_{k=0}^{n}{}' T_k(x_j)T_k(x) &= \sum_{k=0}^{n}{}' \cos(k\theta_j)\cos(k\theta)\\ &= \frac{1}{4}\frac{\sin((n+1/2)(\theta+\theta_j))}{\sin((\theta+\theta_j)/2)} + \underbrace{\frac{1}{4}\frac{\sin((n+1/2)(\theta-\theta_j))}{\sin((\theta-\theta_j)/2)}}_{S_n(\theta-\theta_j)}.\end{aligned}$$

Summing over j we obtain

$$\sum_{j=0}^{n}|\ell_j(x)| = \frac{1}{2(n+1)}\sum_{j=0}^{n}|S_n(\theta-\theta_j) + S_n(\theta+\theta_j)|.$$

Next we use the identity

$$\sin((n+1/2)(\theta\pm\theta_j)) = \pm(-1)^j\left[\cos((n+1)\theta)\cos\left(\frac{\theta\pm\theta_j}{2}\right) + \sin((n+1)\theta)\sin\left(\frac{\theta+\theta_j}{2}\right)\right].$$

Substituting this into our expression for $\lambda(x)$, we obtain the bound

$$\sum_{j=0}^{n}|\ell_j(x)| \leq \frac{|\cos((n+1)\theta)|}{2(n+1)}\sum_{j=0}^{n}\left|\cot\left(\frac{\theta+\theta_j}{2}\right) - \cot\left(\frac{\theta-\theta_j}{2}\right)\right|.$$

We claim that is suffices to optimize over $[0, \frac{\pi}{2(n+1)}]$. On this interval all $|\ell_j(x(\theta))|$ are decreasing, so the right hand side is bounded above by

$$\begin{aligned}\frac{2}{n+1}\sum_j\left|\cot\left(\frac{\theta_j}{2}\right)-\cot\left(\frac{-\theta_j}{2}\right)\right| &= \frac{1}{n+1}\sum_j\left|\cot\left(\frac{\theta_j}{2}\right)\right| \\ &\leq \frac{1}{n+1}\cot\left(\frac{\theta_0}{2}\right)+\frac{1}{\pi}\int_{\theta_0}^{\pi}\cot(x/2)\,\mathrm{d}x \\ &= \frac{1}{n+1}\cot\left(\frac{\theta_0}{2}\right)+\frac{2}{\pi}\log\frac{\sin(\pi/2)}{\sin(\theta_1/2)}.\end{aligned}$$

Setting $\theta_1 = \pi/(4(n+1))$, we arrive at the bound

$$\Lambda \leq \frac{4}{\pi}+\frac{2}{\pi}\log\left(\frac{4(n+1)}{\pi}\right)+O(n^{-2}).$$

□

This is somewhat encouraging. What about equispaced?

Theorem 23. *For equispaced points, Λ_n satisfies*

$$\Lambda_n > \frac{2^{n-2}}{n^2}, \quad \Lambda_n \sim \frac{2^{n+1}}{en\log n}.$$

Proof. Try to get a bound as close as you can. □

Runge Phenomena

The blow up of the Lebesgue constant for equispaced interpolation is associated with a behavior known as *Runge phenomena*. Polynomial interpolation from equispaced points is exponentially ill-conditioned. Typically, the interpolant is relatively fine in the middle of the interval but oscillates wildly near the endpoints.

Example 6. *The witch of Agnesi. Consider the function*

$$f(x) = \frac{1}{1+a^2x^2}.$$

To make this more precise, set $\ell(x) = \prod_{j=0}^{n}(x-x_j)$ and let $p_f(x)$ be the Lagrange interpolant of f, which is assumed to be sufficiently smooth. For $x \neq x_j$, $j = 0, 1, 2, \ldots, n$, we define

$$\phi_x(t) = f(t) - p_f(t) - \frac{f(x)-p_f(x)}{\ell(x)}\ell(t).$$

Note that $\phi_x(x_i) = 0$ and $\phi_x(x) = 0$. Since ϕ_x is a polynomial in t, the interlacing of roots then says that $\phi_x^{(n_1)}$ has a root on $[-1, 1]$ at ξ_x, say. Then

$$\phi_x^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - 0 - \frac{f(x)-p_f(x)}{\ell(x)}(n+1)! = 0.$$

The zero arises since $p_f \in P_n$. After rearranging, we find

$$|f(x) - p_f(x)| = \frac{1}{(n+1)!} |\ell(x)| \, |f^{(n+1)}(\xi_x)|.$$

Example 7. *The witch continued. It is clear that*

$$\|\ell\|_\infty \leq h^{n+1} n!.$$

Also,

$$\|f^{(n+1)}\|_\infty \leq a^n n!$$

and so the bound becomes

$$|f(x) - p_f(x)| \leq \frac{h^{n+1}(n!)^2 a^n}{(n+1)!} \sim \frac{1}{n^{3/2}} \sqrt{8\pi} \left(\frac{2a}{e}\right)^n.$$

So $\|f - p_f\|$ converges if $2a/e < 1$. Of course this is an upper bound only, but it gives some reasonable intuition.

How do we avoid Runge phenomena?

i) Change the points.

i) Change the problem. Sacrifice passing through every point using, for example, least squares.

i) Change the approximation space, i.e. use piecewise polynomials or other, more general, approaches (see "radial basis functions" for example).

Potential theory, the Hermite integral formula, and interpolation

In this section we recast Lagrange interpolation as a contour integral. Apart from its aesthetic appeal, the resulting formula gives a relatively slick way of bounding interpolation errors, and introduces deep connections to *potential theory*. Given $x_0, \ldots, x_n$ let ℓ_j denote the j^{th} Lagrange interpolant. Set

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

Then, it is easy to check that

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

If f is analytic in a suitable region, we can get a nice formula for the Lagrange interpolant. We begin by recalling that the interpolating polynomial to f passing through $x_0, \ldots, x_n$ is given by

$$p(x) = \sum_{j=0}^{n} f(x_j)\, \ell_j(x).$$

Now, by Cauchy's formula,

$$\ell_j(x) = \frac{1}{2\pi i} \int_{\Gamma_j} \frac{\ell(x)}{\ell(t)\,(x-t)}\,\mathrm{d}t$$

where Γ_j encloses x_j but no other x_i nor x. It is an exercise to show that the residue of the pole at $t = x_j$ is given by $1/(\ell'(x_j)\,(x-x_j))$. Then, if f is analytic in a neighborhood containing Γ_j,

$$\ell_j(x) f(x_j) = \frac{1}{2\pi i} \int_{\Gamma_j} \frac{\ell(x) f(t)}{\ell(t)\,(x-t)}\,\mathrm{d}t.$$

So, provided $x \notin \bigcup_j \Gamma_j$, and f is analytic inside all the Γ_j's,

$$p(x) = \frac{1}{2\pi i} \sum_{j=0}^{n} \int_{\Gamma_j} \frac{\ell(x)\, f(t)}{\ell(t)\,(x-t)}\,\mathrm{d}t.$$

Clearly, if Γ is a contour enclosing all x_j's but not x, then

$$p(x) = \frac{1}{2\pi i} \int_{\Gamma} \frac{\ell(x)\, f(t)}{\ell(t)\,(x-t)}\,\mathrm{d}t,$$

provided f is suitably analytic.

Now, near x, the integrand has a pole at $t = x$, with residue $-f(x)$. Thus, if we expand our contour to include x, calling this new contour $\tilde{\Gamma}$, then

$$p(x) - f(x) = \frac{1}{2\pi i} \int_{\tilde{\Gamma}} \frac{\ell(x)\, f(t)}{\ell(t)\,(x-t)}\,\mathrm{d}t.$$

Also, clearly

$$p(x) = \frac{1}{2\pi i} \int_{\tilde{\Gamma}} \frac{(\ell(x) - \ell(t))\, f(t)}{\ell(t)\,(x-t)}\,\mathrm{d}t.$$

The first equation is called the Hermite integral formula. While they are not necessarily helpful in practice for computing p, they are very useful for understanding the errors in approximation. To wit,

$$|p(x) - f(x)| \le \underbrace{\max_{t \in \tilde{\Gamma}} \left| \frac{\ell(x)}{\ell(t)} \right|}_{T_1} \cdot \underbrace{\frac{1}{2\pi} \int_{\tilde{\Gamma}} \frac{|f(t)|}{|t-x|}\,\mathrm{d}t}_{T_2}.$$

Typically one then argues that T_1 decays rapidly in n, while T_2, which is independent of n, remains bounded. Of course this depends on the distribution of points. Define

$$\gamma_n(x,t) = \left| \frac{\ell(t)}{\ell(x)} \right|^{1/(n+1)}$$

and

$$\alpha_n = \min_{x \in X,\, t \in \Gamma} \gamma_n(x,t).$$

If $\alpha_n \ge \alpha > 1$, then we get the bound

$$\|p - f\| \in \mathcal{O}(\alpha^{-n}).$$

The Numerical Implementation of Lagrange Interpolation

So far we have been focused primarily on the theoretical implications of Lagrange interpolation. We now turn our attention to practical matters: how do we do it quickly and stably on the computer. Let's start with speed.

For the naïve algorithm,

- each denominator can be formed in precomputation (once per set of points, not once per x). This requires $\mathcal{O}(n^2)$ floating point operations,
- evaluating ℓ_j requires $\mathcal{O}(n)$ operations, and there are $n+1$ of them, so evaluating all of them is $\mathcal{O}(n^2)$ operations. This must be done for *each new x*.

A first improvement

Given

$$w_j = \frac{1}{\prod_{i \neq j}(x_j - x_i)},$$

we observe that if $\ell(x) = (x-x_0)\cdots(x-x_n)$, then

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

For each new x, ℓ can be computed in $2n+2$ floating point operations (flops), and ℓ_j can be computed from ℓ in 3 flops. So, we have improved the evaluation time per x to $\mathcal{O}(n)$. Note also that the "barycentric weights" are independent of both *x and f*. It is also easy to modify this procedure to give a fast update if only a few of the x_i are changed.

We can make this look prettier in the following way. By interpolating the function 1, we see that

$$1 = \sum_{j=0}^{n} \ell_j(x) = \sum_{j=0}^{n} w_j \frac{\ell(x)}{(x-x_j)} = \ell(x) \sum_{j=0}^{n} \frac{w_j}{x - x_j}.$$

Similarly, if p_f is our Lagrange interpolant,

$$p_f(x) = \sum_{j=0}^{n} \ell_j(x) f_j = \ell(x) \sum_{j=0}^{n} \frac{w_j f_j}{x - x_j}.$$

Taking the ratio,

$$p_f(x) = \frac{\sum_{j=0}^{n} \frac{w_j f_j}{x - x_j}}{\sum_{j=0}^{n} \frac{w_j}{x - x_j}}.$$

This is called the barycentric interpolation formula.

Remark 11. *This can be evaluated in $\mathcal{O}(n)$ flops after $\mathcal{O}(n^2)$ flops in precomputation. The formula is quite striking - it represents a polynomial as the ratio of two rational approximations!*

Remark 12. *In principle one can compute the weights w_j approximately in $\mathcal{O}(n \log \epsilon)$, where ϵ is the error tolerance. In practice this is seldom done.*

What about accuracy?

There seems like an awful lot of subtraction going on, and one might get concerned, for example, about evaluations at x's near an x_j. Our next result gives us some reassurance on this front.

Theorem 24 (Higham (2004)). *Suppose $\hat{p}$ is the polynomial one actually computes and ϵ is machine precision. Then*

$$\frac{|p(x) - \hat{p}(x)|}{|p(x)|} \leq (3n+4)\epsilon C_{x,f} + (3n+2)\epsilon C_{x,1} + \mathcal{O}(\epsilon^2),$$

where

$$C_{x,f} = \frac{\sum_{j=0}^{n} \left| \frac{f_j w_j}{x - x_j} \right|}{\left| \sum_{j=0}^{n} \frac{f_j w_j}{x - x_j} \right|}.$$

Note, in particular, that $C_{x,1} = \lambda(x) \leq \Lambda$, where λ is the Lebesgue function and Λ the Lebesgue constant.

proof sketch. Let $\oplus, \ominus, \oslash$ and $\otimes$ denote the floating point operation of $+, -, /$, and $\times$. Then, for any floating point numbers a, b

$$(a \oplus b) = (a + b)(1 + \delta),$$

for some $|\delta| \leq \epsilon$. The same goes for $\ominus, \oplus$ and $\oslash$.

Let ϕ_j denote the floating point result of computing $1/w_j$. Then

$$\begin{aligned}
\phi_j &= (x_j \ominus x_0) \otimes \cdots \otimes (x_j \ominus x_n) \\
&= ((x_j - x_0)(1 + \delta_0)) \otimes \cdots \otimes ((x_j - x_n)(1 + \delta_n)) \\
&= (x_j - x_0)(x_j - x_1)(1 + \delta_0)(1 + \delta_1)(1 + \eta_0) \otimes \cdots ((x_j - x_n)(1 + \delta_n)) \\
&= \frac{1}{w_j} \prod_{i \neq j} (1 + \delta_i) \prod_{i \neq j, i < n-1} (1 + \eta_i).
\end{aligned}$$

Here $|\delta_i|, |\eta_i| \leq \epsilon$.

Thus, the computation of w_j in floating point, which we denote by $\hat{w}_j$ is given by

$$\hat{w}_j = w_j (1 + \xi) \prod_{i \neq j} (1 + \delta_i)^{-1} \prod_{i \neq j, i < n-1} (1 + \eta_i)^{-1},$$

for some $|\xi| \leq \epsilon$ and so

$$|\hat{w}_j - w_j| = |w_j|(1 + \mathcal{O}(n\epsilon) + \mathcal{O}(\epsilon^2).$$

□

For the first barycentric formula

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

one can prove the following theorem.

Theorem 25 (Higham (2004))**.**

$$\frac{|p_f(x) - \hat{p}(x)|}{|p_f(x)|} \leq \frac{(5n+5)\epsilon}{1-(5n+5)\epsilon} C_{x,f}.$$

The proof can be found in the paper *The numerical stability of barycentric Lagrange interpolation* by Nicholas Higham in the IMA Journal of Numerical Analysis (2004).

Hermite interpolation

Frequently we are also interested in the case in which we not only have the values of a function at a collection of points, but also its derivative, or derivatives. In the linear algebraic picture of interpolation this poses no great trouble. Indeed, if our interpolating functions $\{\phi_j\}$ are differentiable and we are given information about the derivative of f at a point x_k then we can simply add an extra row to $\phi_j(x_\ell)$ consisting of $\phi_j'(x_k)$, $j = 0, \ldots, n-1$. For monomials this would give a linear system which looks like

$$\begin{pmatrix} 1 & x_0 & \cdots & x_0^{n-1} \\ 1 & x_1 & \cdots & x_1^{n-1} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n-2} & \cdots & x_{n-2}^{n-1} \\ 0 & 1 & \cdots & (n-1)\, x_j^{n-2} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-2} \\ \alpha_{n-1} \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_{n-2}) \\ f'(x_j) \end{pmatrix},$$

where $\alpha_0, \ldots, \alpha_{n-1}$ are the coefficients of our approximant in the monomial basis.

This is fine, insofar as it is relatively easy to show that it is easy to show that the matrix on the right-hand side is invertible, and it is straightforward to see how to generalize the approach to add more derivative information (i.e. derivative information at more points and/or higher derivatives). Each new piece of information allows us to fit an extra function.

For notational convenience, we show that derivative information at x_j is included by repeating x_j twice when we list our points. Extending this, if we have information about $f(x_j), \ldots, f^{(k)}(x_j)$ then x_j would appear $k+1$ times in our list of nodes.

Example 8. *Fitting the n-term truncated Taylor series of f at a point x_j would correspond to the interpolation problem with nodes $x_j, x_j, \ldots, x_j$ where x_j appears n times in the list.*

This gives us a convenient way of keeping track of which derivatives to include. One can then create the interpolation matrix as above and solve for corresponding coefficients. This approach is a little unsatisfying for the following reason. When all n points are distinct and relatively far away, the 'coefficients to values' matrix $V(x_0, \dots, x_{n-1})$ is invertible. As two points, x_j and x_{j+1} approach each other, the two corresponding rows become more and more parallel, and the condition number of V goes to infinity. When the two points are exactly equal, then we replace one by the derivative, and the system is (relatively) well-conditioned again. Thus, our interpolation problem is discontinuous in the location of the nodes $(x_0, \dots, x_{n-1})$ as a vector in $\mathbb{R}^n$.

In order to fix this situation, we start with the following definition.

Definition 10. *Given $v \in C^k$, the* divided difference of order j, $j \leq k$, *is the symmetric function defined inductively by the relation*

$$v[x_0, \dots, x_j] = \begin{cases} \frac{v[x_1,\dots,x_j] - v[x_0,\dots,x_{j-1}]}{x_j - x_0} & x_j \neq x_0, \\ \frac{1}{j!} v^{(j)}(x_0) & x_0 = x_1 = \cdots = x_j. \end{cases}$$

The following theorem then gives a solution to the general interpolation problem.

Theorem 26. *If $p \in P_n$ and $p[x_0, \dots, x_n]$ denotes the divided difference, then*

$$p(x) = p[x_0] + p[x_0, x_1](x - x_0) + \cdots + p[x_0, \dots, x_n](x - x_0) \cdots (x - x_{n-1}).$$

In particular, we have the following obvious corollary.

Corollary 5. *For any $x_0, \dots, x_n$, the Hermite interpolation problem associated with the monomials $1, x, \cdots, x^{n-1}$ is always solvable. In particular, for any set of real numbers $q_0, \dots, q_{n-1}$, there exists a polynomial p such that $p(x_j) = q_j$, where once again repetition of a point x_j means that p should be replaced by an appropriate derivative evaluated at x_j.*

This property can be generalized, and extends the notion of a Haar space.

Definition 11. *A space is called an* extended Chebyshev space *if any Hermite interpolation problem has a unique solution.*

As for Haar subspaces, one can formulate many different characterizations. For example, we have the following.

Lemma 7. *An n-dimensional space H is an extended Chebyshev space if and only if any non-zero element of H vanishes at most $n - 1$ times, including multiplicities.*

The following proposition also follows straightforwardly from the definition.

Proposition 5. *Let H be an extended Chebyshev system with basis $v_0, \dots, v_n$. For any collection of points $x_0, \dots, x_n$, (allowing for repetitions) the matrix*

$$V_{i,j} = \sum_{\ell=0}^{j} v_i[x_0, \dots, x_\ell]$$

is invertible. Here $v_i[\cdot]$ denote the divided differences.

References and Further Reading

Lagrange interpolation, Lebesgue constants, Lebesgue constant bounds, and Hermite interpolation: *ATAP* by L.N. Trefethen.
For further reading on barycentric interpolation, see the paper *Barycentric Lagrange Interpolation* by J-P. Berrut and L.N. Trefethen.

Additional exercises

Exercise 21. *In this problem we will play around with interpolation in theory and practice.*

1. *Determine what happens to the Hermite integral formula when two points coincide, i.e. suppose one uses the points $x_0, x_0, x_2, x_3, \dots, x_n$ in the Hermite integral formula. What expression does this correspond to for the interpolant p? Does it still make sense? Hint: start with*

$$p(x) = \frac{1}{2\pi i} \int_\Gamma \frac{f(t)\,(\ell(t) - \ell(x))}{\ell(t)\,(t - x)}\,\mathrm{d}t$$

   and go backwards through the derivation of the Hermite integral formula.

2. *Confirm your answer to part (a) numerically for $f(x) = e^{-0.3x}$. That is, use the Lagrange-type formula you derived to compute $p(x)$ and then compute it by directly integrating*

$$p(x) = \frac{1}{2\pi i} \int_\Gamma \frac{f(t)\,(\ell(t) - \ell(x))}{\ell(t)\,(t - x)}\,\mathrm{d}t.$$

   For Γ choose a circle of radius 1.5 and perform the integration using left-hand Riemann sum (why?). For points, use a 2 point Chebyshev rule with each node repeated twice (to give a total of 4 points - two sets of two identical points). How does the error at $x = 1/3$ change as you add more points. Compare with the unrepeated Chebyshev rule.
   Note: *you can use either Chebyshev points of the first or second kind - whichever you prefer (either $\cos(1/(2n) + j/(n+1)$, $j = 0, \dots, n$ or $\cos(j\pi/n)$, $j = 0, \dots, n$, the second type are more common).*

3. *More generally, given a set of distinct points* $x_1, \ldots, x_n$*, let* α_n *be defined by*

$$\alpha_n = \min_{x \in [-1,1],\, t \in \Gamma} \frac{\left(\prod_{j=1}^{n} |t - x_j|\right)^{1/n}}{\left(\prod_{j=1}^{n} |x - x_j|\right)^{1/n}}.$$

   Now, consider the doubled-version with each node repeated twice $[x_1, x_1, \ldots, x_n, x_n]$*. Assuming that* $\alpha_n > 0$ *find a bound for the error in terms of* α_n*,* f*, and* Γ.

4. *Implement the barycentric interpolation formula for* 5 *Chebyshev nodes on* $[-1, 1]$ *and* $f(x) = e^{-0.3x}$*. Investigate what happens as the error approaches one of the nodes.*