APPROXIMATION THEORY

ABSTRACT. Approximation theory is in other words the study of function compression. In many times the compressibility is expressed in forms of smoothness.

In this class, we ask and try to answer the following questions:

Theoretical	Application
1. What function to approximate?	1.How fast? (Algorithmic)
2. With what?	2. Accuracy?
3. In what sense?	2. Numerically stable?

CONTENTS

1. 3/20: Quadratures; Basis 2
1.1. Quadratures 2
1.2. Basics of Basis 5
2. 3/22: Best approximations 8
2.1. Existence of best approximation 8
2.2. Uniqueness of best approximation 11
3. 3/27: Stone Weierstrass Theorem and plus 13
3.1. Different proofs of Weierstrass approximation 14
3.2. Equi-Oscillation Theorem 17
4. 3/29: Remez Exchange Algorithm; Monomial approximations 18
4.1. Remez Exchange Algorithm 18
4.2. Monomials are not so good. 19
5. 4/3: Convergence of Chebyshev polynomials 22
6. 4/5: Interpolations: Vandermonde matrix; Lagrange interpolation 26
6.1. Idea 1: Construct Vandermonde Matrix 26
6.2. Idea 2: Lagrange Interpolation 27
7. 4/10: Some potential theory; Barycentric polynomials 30
7.1. Potential theory, Hermite Intergral, Formulea and Interpolation 30
7.2. Numerical considerations: 31
8. 4/12: Integration 33

8.1. Gauss-Legendre quadrature and orthogonal polynomials: 36
9. 4/19: splines; Adaptive Discretization 39
10. 4/26: Conjugate Gradient 42
11. 5/1: Conjugate gradient continue; Trigonometric approximation 46
11.1. More on conjugate gradient 46
11.2. Trigonometric approximation 48
12. 5/3: Approximation with Trigonometric Polynomials; Gibb's Phenomenon 51
12.1. Best approximation with Trigonometric Polynomials 51
12.2. Gibb's Phenomenon 53
13. 5/8: Fourier Series; Convergence theory 56
13.1. Fourier Series 56
14. 5/10: Fejer Kernel; Fourier Transform 60
14.1. Fourier Transform 60
15. 5/15: Bandlimited function; Uncertainty Principle 62
15.1. Bandlimited function 62
16. 5/17: Reproducing Kernel Spaces 65
Appendix A. a 69
Appendix B. b 69
Appendix C. c 69

1. 3/20: Quadratures; Basis

Let's say that we have the goal of computing integrals numerically. That is, we face the problem of recovering information from sample points.

1.1. **Quadratures.**

Say that we want to get the integral of a function, then the very first thing we study is to use Riemann sums, that is, approximate with endpoint valued step functions. This approach depends on the smoothness of f, as we can see from the curves of the function.

Proposition 1.1. *For the Riemann sums and for f that is differentable, we have*

$$\left|R_n(f) - I(f)\right| \leq C \cdot h \cdot ||f'||_\infty$$

where $I(f)$ is the exact integral on the domain.

The proof is just since the error on each step is smaller than the triangle.

We might think that Trapezoid rule works better, judging from the picture of the approximation. And indeed we have the result below.

Proposition 1.2. *For the Trapezoid rule and for f that is twice differentable, we have*

$$\left|T_n(f) - I(f)\right| \leq C \cdot h^2 \cdot ||f''||_\infty.$$

The proof, again, is to note that the maximal difference has leading term to be the second order error, and for that the maximal difference occurs when the step function is exactly on top of the lowest point of the parabola.

A counter example that says smoothness is needed can be provided: Just take the abosolute value function at around 0 then we see that the decay is $O(h)$ rather than $O(h^2)$.

Yet, even though this approximation with trapezoids looks better, it really is just silly, as we can see from the formulas:

$$L_n(f) = \sum_{i=0}^{n-1} \frac{f(x_i)}{n}; \qquad R_n(f) = \sum_{i=1}^{n} \frac{f(x_i)}{n}$$

$$T_n(f) = \frac{L_n(f) + R_n(f)}{2} = \frac{f(x_0)}{2n} + \sum_{i=1}^{n-1} \frac{f(x_i)}{n} + \frac{f(x_n)}{2n}$$

so the only difference between these methods are that occurs at the endpoints. Silly!

Let's see via 2 different method how this approximation behaves.

Method 1:

Let's say that f is smooth enough and periodic (WLOG on $[0, 1]$). Now since f is periodic we have that the trapezoid rule and the endpoint rules are the same. Now we use the Fourier series and get

$$f(x) = \sum_{m\in\mathbb{Z}} e^{2\pi i m x} f_m; \quad f_m = \int_0^1 e^{-2\pi i m x} f(x) dx$$

and we apply the Trapezoid formula here to get (note that $x_j = jh$ where $h = \frac{1}{n}$ is the step size):

$$\begin{aligned} T_n(f) &= \sum_{j=0}^{n-1} \frac{1}{n} \left(\sum_{m\in\mathbb{Z}} e^{2\pi i m x_j} f_m \right) \overset{finite\ \ sum}{=} \sum_{m\in\mathbb{Z}} \frac{f_m}{n} \sum_{j=0}^{n-1} \left(e^{2\pi i m h}\right)^j \\ &= \sum_{m\in\mathbb{Z}} \frac{f_m}{n} \cdot \begin{cases} n & mh \in \mathbb{Z} \\ \frac{1-e^{2\pi i h m n}}{1-e^{2\pi i h m}} = \frac{1-e^{2\pi i m}}{1-e^{2\pi i h m}} = 0 & \text{otherwise} \end{cases} \end{aligned}$$

which yields that $T_n(f) = \sum_{m \in n\mathbb{Z}} f_m$. But note that $I(f) = f_0$ just by definition of Fourier coefficients, so we have

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

which means that the speed of convergence of the Trapezoid rule is related to the decay of Fourier coefficients. Now to compute the above we have

$$f_m = \int_0^1 -\frac{1}{2\pi i m} f(x) d e^{-2\pi i m x} = \frac{1}{2\pi i m} \int_0^1 e^{-2\pi i m x} f'(x) dx$$

and doing ibp k times we get that

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

which means

$$\begin{aligned} E_n := |T_n - I| &\leq \frac{||f^{(k)}||_1}{(2\pi)^k} \sum_{j \in \mathbb{Z}; j \neq 0} \frac{1}{(jn)^k} \leq \frac{2||f^{(k)}||_1}{(2\pi n)^k} \sum_{j=1}^{\infty} \frac{1}{j^k} \\ &\leq \frac{2||f^{(k)}||_1}{(2\pi n)^k} \zeta(k) \leq \frac{2||f^{(k)}||_1}{(2\pi n)^k} \frac{k}{k-1} \end{aligned}$$

where the last step we assumed $k \geq 2$. Now we see that E_n is decaying with n:

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

since we've assumed that f is smooth, so the nominator does not matter.

Method 2:

We use Euler Maclaurin formula to compute the error. See https://en.wikipedia.org/wiki/Euler-Maclaurin_formula.

Letting

$$I = \int_m^n f(x) dx; \qquad S = f(m+1) + \cdots + f(n)$$

and assuming for some $k \in \mathbb{N}$ we know $f \in C^k[m, n]$, then

$$S_I = \sum_{j=1}^{k} \frac{B_j}{j!} \left(f^{(j-1)}(n) - f^{(j-1)}(m) \right) + R_k$$

where B_j are the Bernoulli numbers and the remainder term has order

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

where if k is odd we can even take ζ off.

But now if we set $k = 1$ and rescale $[m, n]$ to $[0, 1]$ we will have

$$T_n(f) - I(f) = R_1$$

where

$$|R_1| \leq \frac{2h^2}{2\pi} \int_0^1 |f'(x)| dx$$

since

$$\int_m^n |f'(x)| dx \overset{x=m+hy}{=} h \int_0^1 |f'(m+hy)| dy \overset{?}{=} h^2 \int_0^1 |f'(z)| dz.$$

This also implies that the order is roughly h^k since for each derivative we take out more hs. Note that here $h = \dfrac{1}{n-m}$ so we get the same bound as before.

We can use this method for better quardratic approximation, where $B_2 = \dfrac{f'(1) - f'(0)}{2}$.

For more information, check out endpoint corrected quardratures.

To summarize, we know that for periodic smooth functions, Trap/LHS/RHS rules are the gold standard since they converge faster than any power of n. (That is, $\forall k \in \mathbb{N}$ it's fine.)

If we plot the error against how many interpolation points, then we should get a period of chaotic behavior then rapid convergence, thanks to the smoothness of f.

Note that in most cases errors are not spread uniformly, so there's almost always some parts of the function where most error occur.

1.2. **Basics of Basis.**

We know for finite dimensional vector spaces, there exists a set of basis. But this generates poorly to infinite dimensional spaces.

Def 1.3. *$\chi \subset E$ is a Hamel Basis if $\forall v \in E$ can be expressed uniquely as a finite linear combinations of elements in χ, i.e.*

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

The problem though is that χ is almost always uncountable, making things hard to deal with.

Proposition 1.4. *Every linear vector space has a Hamel basis.*

Now we move to Banach spaces.

Def 1.5. *$\{x_j\}_{j=1}^{\infty}$, a countable set in a Banach space X is a Schauder basis if $\forall v \in X$, there exists unique coefficients $\alpha_1, \alpha_2, \ldots$ such that*

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

In particular, note that the sum converges.

Remark 1.6. *Not all Banach space has a Schauder Basis. A necessary condition is that the space must be separable. But even this is not sufficient (the counterexample, however, is not trivial).*

Theorem 1.7. *If x_j is a Schauder basis, then there exists M such that $\forall v \in X$*

$$\left|\left|\sum_{j=1}^{n} \alpha_j x_j\right|\right| \leq M||v||, n = 1, 2, \dots$$

where we call M the basis constant.

Proof. Let

$$E := \left\{ \{\alpha_j\}_{j=1}^{\infty} : \sum_{j=1}^{\infty} \alpha_j x_j \text{ converges in } X \right\}$$

and

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

We claim:

(1) E is Banach;
- Find Cauchy sequence of sequences $\left\{\{\alpha_i^m\}_{i=1}^{\infty}\right\}_{m=1}^{\infty}$;
- Find limit of each i for $\{\alpha_i^m\}_{m=1}^{\infty} \in \mathbb{R}$, that is, the limit in $\mathbb{R}$ of the i-th term of the each sequence (Cauchy because we can bound each term by $\frac{2\varepsilon}{||x_j||}$);
- Show that this is indeed the sequence (the choice we find for each ith term guarantees this).

(2) There exists bounded bijection from E to X.
- This follows from the uniqueness in the definition of Schauder basis.

Then by inverse mapping theorem (open mapping, i.e. bounded inverse), we are done since M is the norm of the bounded bijection. □

Now let's consider more about the partial sums. Let

$$S_n(v) := \sum_{j=1}^{n} \alpha_j x_j$$

where α uniquely depends on v by definition of Schauder basis. Now, since S_n is a projection onto $\text{span}\{x_1, \dots, x_n\}$ and all S_n are uniformly bounded by above theorem, we can view the

coefficients α_j as linear functionals on X and call them biorthonormal functionals denoted by x_j^*, where

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

and remember we can view the identity as

$$I(\cdot) = \sum_{j=1}^{\infty} x_j x_j^*(\cdot).$$

So we can get a bound for the norms biorthonormal functionals usingthe following argument:

$$|x_j^*(v)| \cdot ||x_j|| = ||x_j^*(v)x_j|| = ||S_j(v) - S_{j-1}(v)|| \leq 2M \cdot ||v||$$

and by dividing $||v||$ on both sides we have

$$||x_j^*|| \cdot ||x_j|| \leq 2M.$$

Now, for Hilbert spaces, we have the regular definition of orthonormal system.

Def 1.8. *If $\{x_j\}$ is an orthonormal system in $\mathcal{H}$, then the Fourier Series of v with respect to x_j is*

$$v = \sum_j \langle v, x_j \rangle x_j$$

where the Fourier coefficients are

$$v_j = \langle v, x_j \rangle.$$

Theorem 1.9. *(Bessel's inequality)*

$$\sum |\langle v, x_j \rangle|^2 = ||v||^2.$$

This means that among all convergent sequences

$$s = \sum \alpha_j x_j$$

$|v - s|$ is minimized by the Fourier series of v.

And lastly, to get an orthonormal system we can use Gram-Schimidt.

2. 3/22: Best approximations

2.1. **Existence of best approximation.**

Def 2.1. *Let E be Banach, and M be a non-empty subset of E, define*

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

Def 2.2. *Given $f \in E$, we define the set of nearest points to f in M as*

$$P_M f := \left\{ u_0 \in M \middle| ||f - u_0|| = d(f, M) \right\}.$$

We note that $P_M f$ can well be empty, if say we choose M to be open.

Def 2.3.

- *M is an existence set if $P_M f$ is non-empty for all $f \in E$.*
- *M is a uniqueness set if $|P_M f|$ is at most 1.*
- *M is a Chebyshev set if M is both existence and uniqueness.*

We note that if M is existence then M is closed, since otherwise we can choose a point in the boundary. So we want to ask if the reverse direction is true, that is, whether all closed sets are existence. And the result is no, due to the following example.

Example 2.4. *Closed but not existence.*

Let $E = l^2$ and $M = \left\{ \left(1 + \frac{1}{j}\right) e_j \right\}_{j=1}^{\infty}$ and let $f = 0$, then M is closed because each point is away from others, but there's no closest point in M to f.

Example 2.5. *Existence not uniqueness.*

Let $E = l^\infty$ and $M = \left\{ v \in l^\infty \middle| ||v||_\infty \leq 1 \right\}$ and let $f = (2, 0, 0, \dots)$ then any $v \in M$ with $v_1 = 1$ is a best approximation.

Proposition 2.6. *Let E be Banach and M be a finite dimensional subspace, then M is an existence set.*

Proof. Since $0 \in M$ we know that $B_{||f||}(0) \cap M$ is compact (bounded closed finite dim), so $d(f, u)$ attains minimum in $B_{||f||}(0) \cap M$ as the distance map is continuous by triangle inequality. □

Proposition 2.7. *If M is convex, then $P_M f$ is convex.*

Proof. Just take two points in $P_M f$ so we know that their average of their distance to f is still the infimum, then we are done. □

Def 2.8. *Let $S = \partial B$, the unit sphere in the Banach space E. Then we define the following:*

- *E is strictly convex if $x, y \in S$, $x \neq y$ implies $||x + y|| < 2$.*
- *E is uniformly convex if given $\varepsilon > 0$, $\exists \delta = \delta(\varepsilon)$ such that $x, y \in S$ if $||x+y|| > 2-\delta$ implies $||x - y|| \leq \varepsilon$.*
- *A supporting functional at point u is a functional l such that $l \in E^*, ||l|| = 1, l(u) = ||u||$. The existence of such a functional is due to Hahn-Banach theorem.*
- *E is smooth if for each $u \in S$, $\exists!$ supporting functional.*

With this, we can give a generic criterion for best approximation. The geometric picture is that there's a l-level set such that the value of l at u_0 is σ and the value at f is nothing but $||u_0 - f|| + \sigma$.

Theorem 2.9. *Suppose M is convex in Banach space E and for $f \in E, f \notin \overline{M}$, then $u_0 \in M$ is a best approximation to f iff $\exists$ a linear functional $l \in E^*$ with the following properties:*

(i) $||l|| = 1$*;*
(ii) $l(f - u_0) = ||f - u_0||$*;*
(iii) $\mathrm{Re}(l(u - u_0)) \leq 0$ *for all* $u \in M$.

Proof.

$\Leftarrow$:

For this direction, we assume that such an l exists, then for any $u \in M$ we have

$$||u - f|| \overset{(i)}{\geq} \mathrm{Re}(l(f - u)) = \mathrm{Re}\left\{l(f - u_0) - l(u - u_0)\right\} \overset{(ii)}{\geq} \mathrm{Re}\left\{l(f - u_0)\right\} \overset{(iii)}{\geq} ||f - u_0||.$$

$\Rightarrow$:

Let $r = ||f - u_0||$ and let $B_r(f)$ be the open ball around f. Then $B_r(f) \cap \overline{M} = \emptyset$ since u_0 minimizes $d(u, f)$. Now using the Hahn-Banach separating theorem on complex linear functionals, we know that there exists $\tilde{l}$ with $\mathrm{Re}\{\tilde{l}(B_r(f))\} > \sigma$ and $\mathrm{Re}\{\tilde{l}(M)\} \leq \sigma$. Since u_0 is the closest point we know that $\mathrm{Re}\{\tilde{l}(u_0)\} = \sigma$ since otherwise there's a better approximation.

Now, since l is linear we let $\beta := \mathrm{Re}\{\tilde{l}(f - u_0)\} > 0$ by above definition of $\tilde{l}$. Now we pick

$$l = \frac{r}{\beta}\tilde{l}$$

and claim that this is the one we want.

By construction we have

$$\mathrm{Re}\{l(f - u_0)\} = ||f - u_0||$$

and so we need to take off the real value operator to prove (ii).

First we show (i), for which we assume that $||l|| > 1$, that is, since l linear, there exists $w \in \overline{B_r(0)}$ such that $||l(w)|| > r$. Now, let $z = f - w$ then

$$\mathrm{Re}\{\tilde{l}(z)\} = \mathrm{Re}\left\{\tilde{l}(u_0) + \tilde{l}(f - u_0) - \frac{\beta}{r}l(w)\right\} = \sigma + \beta - \frac{\beta}{r}\mathrm{Re}\{l(w)\} < \sigma$$

contradiction to the fact that $z \in B_r(0)$. Hence, $||l|| = 1$ and this implies $l(f-u_0) = ||f-u_0||$ directly. From which we conclude (i) and (ii). For (iii) we just use definition of l, since $\frac{r}{\beta}$ is just a constant. □

Remark 2.10. *One remark here is that we can modify the statement to allow a different functional for each element of M. This gives us the generalized Kolmogorov criterion.*

Now let's look at some examples.

Example 2.11.

Let $E := \mathcal{H}$ be Hilbert. Then for M convex and $u_0 \in M$, there is a nearest point to $f \in \mathcal{H}$ iff $\mathrm{Re}\langle f - u_0, u - u_0\rangle \leq 0$, $\forall u \in M$.

This is just because the inner product above describes the angle. Moreover, that linear functional is, by Riesz, described by the inner product with $f - u_0$ since it has to satisfy (ii).

Example 2.12.

Let $E = L^p(X, \mu)$ and $p \in (1, \infty)$. Now u_0 is a best approximation iff

$$\int |f(x) - u_0(x)|^{p-2}\overline{(f(x) - u_0(x))}u(x)d\mu = 0$$

for all $u \in M$.

To get a hint of what's going on, just take $u = e_0 := f(x) - u_0(x)$ and everything makes sense. This is just a version of dual product for L^p.

Example 2.13.

If $E = L^1(X, \mu)$ for X compact, then u_0 is best approximation iff

$$\mathrm{Re}\int_{X\setminus Z}\frac{\overline{e_0}(x)}{|e_0(x)|}u(x)d\mu \leq \int_Z |u(x)|d\mu$$

where $Z := \{x \in X | e_0(x) = 0\}$, that is, the set where f agrees with u_0. And if the converse is true the u is a better approximation.

Example 2.14.

Continuing on that route let $E = C(X)$ for X compact, then $u_0 \in M$ is a best approximation iff

$$\inf_{x\in P[e_0]} \mathrm{Re}\left\{\overline{e_0}(x)[u(x) - u_0(x)]\right\} \leq 0$$

for all u. Here

$$P[e_0] := \left\{ x \in X \middle| |e_0(x)| = ||e_0|| \right\}.$$

One way to see this is that if there is some $u \in M$ such that $\operatorname{sgn}(e_0) = \operatorname{sgn}(u - u_0)$, then some combination of u and u_0 is a better approximation since what's maintaining the largest $||e_0||$ is mitigated by that combination.

2.2. **Uniqueness of best approximation.**

Def 2.15. *A Haar subspace M is an n-dimensional subspace of $C(X)$ such that for all $u \in M$, u has at most $n-1$ zeros in X.*

Lemma 2.16. *M is a Haar subspace iff for all sets of n points $x_1, \dots, x_n \in X$ and any $\beta_1, \dots, \beta_n \in \mathbb{C}$, the interpolation problem $u(x_i) = \beta_i$ for each i has a solution for $u \in M$.*

Remark 2.17. *What this means is that for any basis $u_1, \dots, u_n$ of the Haar subspace M, we can define the n by n matrix Φ by*

$$\Phi_{ij} = u_j(x_i)$$

is invertible. That is, we can get a linear combination of u_js such that $\sum_j \alpha_j u_j(x_i) = y_i$, where

$$\Phi\alpha = y$$

is really the problem.

Theorem 2.18. *(Haar's Uniqueness theorem) If X is locally compact then a finite dimensional subspace of $C(X)$ is a Haar subspace iff it's a Chebyshev set, i.e. for each point there's a unique closest point.*

All seems good that if we have a Haar subspace then we basically have anything we want in terms of interpolating. But if you think about it, it really doesn't make sense for us to just be able to interpolate any set that is essentially high dimension because there's loops inside.

Theorem 2.19. *(Mairhuber-Curtis Theorem) Suppose $\Omega \subset \mathbb{R}^d$ for $d \geq 2$ and Ω has an interior point. Then there's no Haar subspace of $C(\Omega)$.*

Proof. Assume there is then we can find an open ball contained in Ω since it has interior. Then we pick two points in that open ball, say x_i and x_j. Then since $d(x_1, \dots, d_n) := \det(\Phi)$ is continuous in x, so we can continuously exchange x_i and x_j via a loop and get a new matrix that has two rows exchanged. Let's say $j = i + 1$ then the determinant is flipped sign, which means at some intermediate point the matrix is not invertible, hence the set not Haar. □

The question now is that, given a finite dimensional subspace, can I find points $x_1, \dots, x_n$ that are good for interpolation. That is, we find points for some base, rather than finding basis

for points. To this end we let S be arbitrary collection of points with $|S| \geq n$, and $f_1, \dots, f_n$ to be functions from $S \to \mathbb{C}$ that are bounded and linearly independent, i.e.

$$f(x) = \sum_{j=1}^{n} \alpha_j f_j(x).$$

Now the goal is to pick $x_1, \dots, x_n \in S$ and sample our function at the points, given $f_r = f(x_r)$. And we want to find coefficients such that

$$f(x) = \sum_{j=1}^{n} g_j(x) f_j$$

where $g_j : S \to \mathbb{C}$. This can be done since the dimension of f is n.

A necessary and sufficient condition for this is obviously $\det(f_j(x_i)) \neq 0$. But we might also improve for practical matters.

Idea 1: We can choose $(x_1^*, \dots, x_n^*) = \operatorname{argmin} K(f_j(x_i))$, i.e. the ones that minimizes the condition number. This is because if $g_j \gg 1$ then we lose digits due to the super large coefficients.

Some more claims are in the notes, and we've looked at two types of problems: scattered data and experimental design. More to come next time.

3. 3/27: Stone Weierstrass Theorem and plus

Theorem 3.1. *(Stone Weierstrass) Let X be a compact metric space and $\mathcal{A}$ be a sub-algebra of $C(X)$. If $\mathcal{A}$ separates points in X and nowhere vanishes in X, then $\mathcal{A}$ is dense in $C(X)$.*

Def 3.2. *$\mathcal{A}$ separates points if $\forall x, y \in X$, $x \neq y$, then there is an $f \in \mathcal{A}$ which has $f(x) = f(y)$.*

Proof.

Step 1: If $f \in \mathcal{A}$, then $|f| \in \overline{\mathcal{A}}$, which implies that $\max\{f_1, \dots, f_n\} \in \overline{\mathcal{A}}$.

Step 2: $\exists g \in \mathcal{A}$ such that $g(x) = f(x)$ and $g > f - \varepsilon$.

Step 3: Exists $||g - f||_\infty \leq \varepsilon$.

Now we fill in the details.

Step 1:

First, for step 1, we only need functions with norm in $[-||f||, ||f||]$, and the way we do it is to use the polynomial from binomial theorem:

$$|s| = \sqrt{1 - (1 - s^2)} = \sum_{n=0}^{\infty} t^n (-1)^n \binom{\frac{1}{2}}{n}$$

where $(1 - s^2) = t$ and we know the above sum converges uniformly on $[-1, 1]$. Thus we scale $||f||$ to 1 and let $q_n(s) = p_n(1 - s^2)$ where

$$p_n(t) = \sum_{n=0}^{n} t^n (-1)^n \binom{\frac{1}{2}}{n}.$$

Thus we know $\left|q_n(s) - |s|\right| \leq \varepsilon$ for $s \in [-1, 1]$. We don't have $q_n \in \mathcal{A}$ but we know $q_n(f) - q_n(0) \in \mathcal{A}$ since it is an algebra. But

$$|q_n(s) - q_n(0) - |s|| \leq 2\varepsilon$$

which is what we need.

To go from absolute value to maximal we just take the difference and do absolute and add to f.

Step 2:

For $x \neq y$ there exists $h_{x,y} \in \mathcal{A}$ with $h_{x,y}(x) = f(x)$ and $h_{x,y}(y) = f(y)$. Then by compactness there's a finite number of y such that

$$\max h_{x,y_1}(t), \dots, h_{x,y_n}(t) \geq f(t) - \varepsilon, t \in X.$$

Step 3:

We just for an ε finite cover of the compact set and find a point in each one and find g_n accordingly. Then we get

$$\min g_1, \dots, g_n(t) \leq f(t) + \varepsilon, t \in X.$$

This concludes our prove. □

Remark 3.3.

(1) Now, if we have $\mathcal{A}, \mathcal{B}$ dense in X and Y, then $f(x,y) = f(x)g(y)$ where $f \in \mathcal{A}$, $g \in \mathcal{B}$ is dense in $C(X,Y)$.
(2) If K is a compact subset of $\mathbb{R}^n$, then n variate polynomials are dense in $C(K)$,

3.1. **Different proofs of Weierstrass approximation.**

Theorem 3.4. *(Weierstrass Approximation Theorem) For $\forall f \in C[0,1]$ and $\varepsilon > 0$, there exist a polynomial p such that $||p - f|| < \varepsilon$.*

It's actually a weaker version of what is proven above, but the proof here gives us a direct construction, which we like.

We build up toward the proof.

Def 3.5. *Let the Bernstein polynomials be*

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

We can get the following by direct computation:

- $B(1)(x) = 1$.
- $B(x)(t) = t$ or we can confusingly write $B_n(x)(x) = x$ for fun.
- $B(x^2)(x) = x^2\left(1 - \frac{1}{n}\right) + \frac{x}{n}$.

Now we show that all points far away from the point we're approximating does not matter.

Lemma 3.6. *Define*

$$\mathcal{F}_n := \left\{ k \in \{0, 1, \dots, n\} \middle| \left| \frac{k}{n} - x \right| \geq \delta \right\}$$

then we have

$$\sum_{k \in \mathcal{F}_n} \binom{n}{k} x^k (1-x)^{n-k} \leq \frac{1}{4n\delta^2}$$

in particular, all points away from x is controlled by $\frac{1}{n}$.

Proof. We just plug in the inequality of the definition of $\mathcal{F}_n$ to get

$$\sum_{k\in\mathcal{F}_n}\binom{n}{k}x^k(1-x)^{n-k} \leq \frac{1}{\delta^2}\sum_{k\in\mathcal{F}_n}\left(\frac{k}{n}-x\right)^2\binom{n}{k}x^k(1-x)^{n-k}$$
$$= \frac{1}{\delta^2}\left(x^2B_n(1)-2xB_n(x)+B_n(x^2)\right) = O\left(\frac{1}{n\delta^2}\right)$$

where the middle step we used the expression for the Bernstein polynomials of degree 0 to 2 polynomials. □

Proof. (Proof 1 of Theorem 3.4)

Now we do the approximation proof. Note that

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

where it's nothing but using $B(1)(x)=1$.

Since f is uniformly continuous we have $\forall\varepsilon, \exists\delta$ with $|f(x)-f(y)|<\varepsilon$ for $|x-y|<\delta$.

Thus, we separate above sum into

$$\leq \left(\sum_{k\in\mathcal{F}}+\sum_{k\notin\mathcal{F}}\right)\left|\binom{n}{k}\left[f(x)-f\left(\frac{k}{n}\right)\right]x^k(1-x)^{n-k}\right| \leq \frac{||f||}{4n\delta^2}+c\cdot\delta$$

where the first is by lemma and the second is because x close to $\frac{k}{n}$, then we just pick n large enough such that $\frac{||f||}{4n\delta^2} = O(\varepsilon)$. □

But note that what is inside is really just the binomial distribution! So we can instead compress the above information into

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

where X_i are independent Bernoulli that has probability $\binom{n}{k}x^k(1-x)^{n-k}$ is the probability that you are at "k" after n turns of flip of coin, where $\mathbb{P}(X_i=1)=x$ and $\mathbb{P}(X_i=0)=1-x$.

So the above just tells us that

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

We now see some different proofs of the same thing.

Proof. (Weierstrass' proof)

We steal some conclusion from PDE and get that $u(x,t)$ is the solution of the heat equation. Then

$$\lim_{t\to 0} u(x,t) \overset{uni}{=} f(x)$$

where the uniform in $x \in [0,1]$ is the important result from PDE. In particular

$$u = \int_{\mathbb{R}} \frac{e^{-(x-y)^2/4t}}{\sqrt{4\pi t}} \hat{f}(y)dy$$

is entire in x so it has a uniformly convergent series. Thus we're done since the Taylor series converge to f by a diagonal argument. □

Proof. (Landau's proof)

For $f \in C[0,1]$, let $\tilde{f} = f - \Big(f(0) + x\big[f(1) - f(0)\big]\Big)$ on $[0,1]$ and 0 else where.

Then, let

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

where c_n is chosen such that

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

But with a change of variables to put the depend of x into $g(x) = (x-y)^2$ we get

$$L_n(x) = c_n \int_{-1}^{1} \tilde{f}(y)(1-g(x))^n dy$$

where the upper and lower bound of the integral uses the fact that $\tilde{f}$ is not supported outside of $[0,1]$. Thus we have that $L_n(x)$ is a polynomial in x.

Moreover, we note that it is the convolution with $\phi_n(t) = c_n(1-t^2)^n \chi_{-1,1}$ then this is an approximated identity, that is $\phi_n * f \to f$.

If $|t| \le \dfrac{1}{\sqrt{n}}$ this means $(1-t^2)^n \ge 1 - nt^2$ which by our definition of n means that $c_n \ge \sqrt{n}$. An exercise here is what is the order of c_n in this case.

Now for $0 < \delta < 1$ we have

$$c_n \int_{[-1,1]\setminus[-\delta,\delta]} (1-t^2)^n dt \le 2c_n(1-\delta^2)^n \to 0.$$

□

3.2. **Equi-Oscillation Theorem.**

Def 3.7. *Say p_* is the approximation of f, then the error function $E = f - p_*$ is equi-oscillating if there's an alternating sequence that reaches $|E|$, then $-|E|$, then $|E|$, alternatively continuing.*

Note that have n equi-oscillating points does not say there's only n points that takes maximum value. It just says that there are n alternating maximums.

Theorem 3.8. *Let P_N be a Chebyshev set in $C[-1,1]$. Now we focus on the case when f is real. We know that the best approximation P_* is real and $f - P_*$ is equi-oscillating in at least $n+2$ extreme points.*

Proof.

Existence:

We know $0 \in P_n$ but how good is it? The answer is of course as good as $||f||_\infty$. So we only consider polynomials of the subset

$$S := \left\{ p \in P_n \middle| ||p - f|| \leq ||f|| \right\}$$

which attains maximum since S is compact.

Equioscillation $\Rightarrow$ Optimality

Assume $f - p$ is equi-oscillates at $x_0 < x_1 < \cdots < x_{n+1}$ and $\exists q \in P_n$ with $||f - q|| \leq ||f - p||$, then $f - p > f - q$ at all positive oscillation points (i.e. $(f-p)(x_i) = |E|$) and reversed for negative oscillation points.

But then we know that $(p-q)(y_i) = 0$ for at least 1 y_i in between x_{i-1} and x_i. Thus $p - q$ has n zero points so it is 0. Contradiction as $p = q$.

Optimality $\Rightarrow$ Equioscillation

If the best approximation oscillates at fewer points, then we can find at most n $y_i \in (x_{i-1}, x_i)$ chosen such that these points separates the equi oscillations, and let $q = (-1)^{\pm}(x - y_1)\ldots(x - y_k)$ then let $h := p + \delta q$ for q small we will get a better approximation.

Uniqueness

If p, q both optimal, then $r := \dfrac{p+q}{2}$ is also optimal since the optimal set is convex. But then at the equi oscillation point x_i of r we have

$$|E| = |f(x_i) - r(x_i)| \leq \frac{1}{2}|f(x_i) - q(x_i)| + \frac{1}{2}|f(x_i) - p(x_i)| \leq |E|$$

hence they are also oscillating points of p and q. To show that they are not reversed oscillating, we use the result from last time such that

$$\operatorname{sgn}(f(x_i) - p(x_i)) = \operatorname{sgn}(f(x_i) - q(x_i))$$

and we are done.

□

4. 3/29: Remez Exchange Algorithm; Monomial approximations

4.1. **Remez Exchange Algorithm.**

One bound crutial to the algorithm is the following.

Proposition 4.1. *(de la Vallée Poussin) Given $f \in C[a,b]$, suppose $q \in P_n$ and $f(x_i) - q(x_i)$ has alternating sign at $a \le x_0 < x_1 < \cdots < x_{n+1} \le b$. Then*

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

where

$$E_n(f) := \min \left\{ ||p - f|| \middle| \forall p \in P_n \right\}.$$

Remark 4.2. *If we want to estimate $f = 0$, then the only n degree we can find that satisfies there's $n+2$ such points is the zero function, since otherwise there's $n+1$ zeros.*

Remez Exchange Algorithm:

Input: continuous f, interval $[a,b]$ and initial guessed equi-oscillation points $x_0^{(0)}, \ldots, x_{n+1}^{(0)}$.

Output: Approximation to f with ε error.

Procedure:

For any iterates, we have given data points $x_0^{(i)}, \ldots, x_{n+1}^{(i)}$ (later we skip i) we can generates the matrix A and value b via given definition:

$$A := \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n & 1 \\ 1 & x_1 & x_1^2 & \ldots & x_1^n & -1 \\ 1 & x_2 & x_2^2 & \ldots & x_2^n & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{n+1} & x_{n+1}^2 & \ldots & x_{n+1}^n & (-1)^{n+1} \end{pmatrix}; \quad \text{and} \quad b := \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_{n+1}) \end{pmatrix}$$

where we solve the linear system

$$Ax = b$$

and get

$$x := \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \\ E \end{pmatrix}.$$

Then just noting what $Ax = b$ means we know that the first n rows of x are the coefficients of the polynomial at the i-th iteration. Now we compute the error function

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

and find the point y of its absolute maximum value. Then we move the most adjacent x_j^{i-1} for which $e^{(i)}(x_j^{i-1})$ has the same sign to that point, and exit the loop when we have

$$\max_j |e^{(i)}(x_j^{i-1})| - \min_j |e^{(i)}(x_j^{i-1})| \leq \varepsilon$$

since the first part is obviously an upper bound, and the second is a lower bound by De La Vallee.

Remark 4.3.

- *We can update all points at each iteration by cutting into pieces.*
- *We can also play this game for other functions by* $\sum w_i e^{-\alpha_i x}$.

4.2. **Monomials are not so good.**

If we can get a really good approximation of x^n with P_{n-1} on $[-1, 1]$, then this means that "x^n" is really just in P_{n-1}.

So what we want to do is to approximate x^n with P_{n-1}. Thus, it's the same thing as minimizing $|x^n - p|$ for some $p \in P_{n-1}$. Now let's say that p_* is the best approximation, then we can find $0 \leq x_0 < x_1 < \cdots < x_n \leq 1$ that are equi-oscillation points such that

$$|x_i^n - p_*(x_i)| = ||x^n - p_*||_\infty$$

and the sign oscillates.

The trick here is to define

$$p := x^n - p_*$$

and in particular we have for $x_i \in (-1, 1)$ we know $p'(x_i) = 0$. But note that p' is a $n - 1$ degree polynomial so has at most $n - 1$ roots if non-zero, yet there are $n + 1$ such points, thus we have $x_0 = -1, x_n = 1$. So we can explicitly write

$$p' = c \cdot \prod_{i=1}^{n-1} (x - x_i)^2.$$

So if we let $E := ||x^n - p_*(x)||_\infty$ then $E^2 - p^2$ has double roots at $x_1, \ldots, x_{n-1}$ and since x_0 and x_n are roots, and there's only 2 left, they are single roots. Hence we have

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

for some β constant. Then $\beta \cdot (1-x^2)(p')^2 = E^2 - p^2$, where we align the highest order coefficient to get $\beta = \frac{1}{n^2}$. Now we note that this form is nothing but an ODE, even though we don't know the solution is a polynomial. But that's fine as we can just solve the ODE and check.

So we have the ODE

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

whose solution is

$$\arccos\left(\frac{p}{E}\right) = \pm n \arccos(x) + C$$

hence

$$p = E \cdot \cos\left(n \cdot \arccos(x) + C\right)$$

and we plug in $x = \pm 1$ to get the constant is $C = k\pi$. Thus

$$p = \pm E \cdot \cos\left(n \cdot \arccos(x)\right) =: T_n$$

which surprisingly is a polynomial!

The reason is this: We have offhand the expression

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

hence letting z to be the suitable form of e^{it} we have

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

so plug in $t = \arccos(x)$ we get the following relation:

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

and when $n = 0$ $T_0 = 1$ thus we get that T_n is a polynomial.

Def 4.4. *We define the above T_n to be the Chebyshev polynomials.*

But note that

$$P = E\cos(n \cdot \arccos(x))$$

so the error is exactly the reciprocal of leading coefficient of T_n, which by it's formula it's clear that $E = \frac{1}{2^{n-1}}$.

Now we consider the L^2 errors, then we get the following theorem:

Theorem 4.5. *Suppose $n \geq t \cdot \sqrt{m}$ where $t \geq 1$ and let $P_n(\phi_m)$ to be the orthogonal projection of*

$$\phi_m = x^m\sqrt{2m+1}$$

onto the subspace of $L^2(0,1)$ spanned by polynomials of degree $\leq n$. Here the coefficient is just to normalize the L^2 norm. Then

$$||P_n(\phi_m)||^2_{L^2[0,1]} \geq 1 - 16\sqrt{m}t^2e^{-2t^2}$$

thus x^m is roughly a polynomial of degree $\sqrt{m}$ in the L^2 sense!

We end the lecture with two remarks about T_n.

Remark 4.6.

- The roots of T_n maximizes the formula:

$$\max_{x\in[-1,1]} |(x-x_1)\cdot\cdots\cdot(x-x_n)| \iff \max_{x\in[-1,1]} \log \sum_{i=1}^{n} |x-x_i|.$$

- One place the above show up is the Green's functions.
- How to find the roots? A good way is to note that if $T_n(x) = 0$ it means

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

But this means if we draw equi-distance points on the semi-circle and projects onto the x-axis, then we've find them. In particular the endpoints has density the same as $O\left(\frac{1}{n^2}\right)$.

5. 4/3: Convergence of Chebyshev polynomials

We want to approximate continuous functions by Chebyshev polynomials today. Is it good?

Theorem 5.1. *Suppose f is Lipschitz 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 and uniformly convergent, here

$$\alpha_k = \frac{2}{\pi} \int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}} dx, k > 0$$

and if $k = 0$ everything's the same except that the constant in front changes to $\frac{1}{\pi}$.

Proof. For $x \in [-1,1]$, we imagine it to be the average of z and $\frac{1}{z}$, it's corresponding points on the unit circle, with $\mathrm{Re}(z) = x$. So the point on the unit circle directly above/below x. Now we define $F(z) := f(x(z))$ and with some care we can get $dx = \pm \frac{i}{z}\sqrt{1-x^2}dz$ where the sign is the same as $\mathrm{Im}(z)$.

Since f is Lipschitz, so is F. Thus expanding F with it's Laurent series (some slight modification at $k = 0$; Also, note $F(z) = F(1/z)$) we have

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

where just focusing on one of the terms and using Cauchy's integral formula we have

$$a_k = \frac{1}{\pi i} \int_{|z|=1} z^{-1-k} F(z) dz$$

whereas using the negative power term we have

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

thus

$$a_k = \frac{1}{2\pi i} \int_{|z|=1} \frac{1}{z}\left(z^k + \frac{1}{z^k}\right) f(x(z)) dz$$

and with a genuine skip of change of variables we get

$$\alpha_k = \frac{2}{\pi} \int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}} dx, k > 0$$

and if $k = 0$ everything's the same except that the constant in front changes to $\frac{1}{\pi}$ due to the little counting twice we did there. □

Theorem 5.2. *Suppose f is bounded continuous on $[-1, 1]$, let p_* denote it's best polynomial approximation from P_n and p denote it's truncated Chebyshev expansion at T_n. Then*

$$\frac{||f - p||}{||f - p_*||} \sim \frac{4}{\pi^2} \log n.$$

Remark 5.3. *If $||f - p_*||$ decays a lot faster than $O(\log n)$ then really $f - p$ decays pretty fast.*

Proof.

We can define an inner product on P_n such that $\phi_1, \dots, \phi_n$ is a basis. Note that this inner product is just chosen so $\langle \cdot, \cdot \rangle_E \neq ||\cdot||_E$ often.

Now the trick is to note $(f - p) - (f - p_*) \in M$ and thus

$$(f - p) - (f - p_*) = \sum_{i=1}^{m} \beta_i \phi_i$$

where we take inner product with ϕ_i on both sides to get

$$\beta_i = -\langle \phi_i, f - p_* \rangle_E$$

since p is the projection of f onto P_n. Denote $\tilde{P}_M = \sum_{i=1}^{n} \phi_i \langle \phi_i, \cdot \rangle$ as the projection map then we know by the coefficients that

$$f - p = f - p_* - \tilde{P}_M(f - p_*)$$

taking the norm on both sides we get

$$||f - p||_E \leq ||f - p_*||_E + ||\tilde{P}_M||_{E \to E} ||f - p_*||_E$$

and dividing we get

$$\frac{||f - p||}{||f - p_*||} \leq 1 + ||\tilde{P}_M||_{E \to E}.$$

Now if we plug in our Chebyshev polynomials where we define

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

and we know

$$\phi_0 = \frac{1}{\sqrt{\pi}}, \dots, \phi_k = \frac{2}{\sqrt{\pi}} T_k$$

forms an orthonormal basis. Hence

$$\tilde{P}(u)(x) = \int_{-1}^{1} \sum_{i=0}^{n} \frac{\phi_i(x)\phi_j(y)}{\sqrt{1 - y^2}} u(y) dy$$

so the norm is (taking $||u|| = 1$)

$$||\tilde{P}||_{\infty\to\infty} \leq \sup_x \int_{-1}^{1} \left| \sum_{i=0}^{n} \frac{\phi_i(x)\phi_j(y)}{\sqrt{1-y^2}} \right| dy$$

where after some trigonalmetrics we "clearly" get

$$||\tilde{P}|| \lesssim \frac{2}{\pi}\left(1 + \log\frac{n+1/2}{4}\right)$$

and hence our result. (we skip some toils here). □

We have shown that to the extend of approximating functions, using Chebyshev polynomials are good enough. So we might as well want to find the error to know how good we are. For some good results we need some kind of regularity.

Theorem 5.4. *For $\nu \geq 0$ an integer, $f, f', \ldots, f^{(\nu-1)}$ are absolutely continuous, and $f^{(\nu)}$ is of bounded variation V, then for $k \geq \nu + 1$ the Chebyshev coefficients α_k satisfy*

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

Remark 5.5. *This means that for one more controled degree of regularity we have one degree faster, assuming that $k \gg \nu$, that is, the denominator is of $O(k^{\nu+1})$.*

Moreover, this fits with what we've seen the first class, the approximations wanders around for some time then decays rapidly. This is exactly what happens when k get's larger than ν.

Proof. When $\nu = 0$, we have $V = ||f||_1$, not too surprising because that's really what it means to vary.

For $\nu = 1$ we have (again by Laurent series)

$$\alpha_k = \frac{1}{\pi i}\int_{|z|=1} f\left(\frac{1}{2}\left(z+\frac{1}{z}\right)\right) z^{k-1} dz \overset{ibp}{=} -\frac{1}{\pi i}\int_{|z|=1} f'\left(\frac{1}{2}\left(z+\frac{1}{z}\right)\right)\frac{z^k}{k}\frac{dx}{dz}dz$$

where the boundary vanishes using the absolutely continuity. Now again we do the same symmetry trick to get

$$\alpha_k = \frac{1}{\pi i}\int_{-1}^{1} f'(x)\frac{z^k - z^{1/k}}{k}dx = \frac{2}{\pi}\int_{-1}^{1} f'(x)\,\mathrm{Im}(z^k)\frac{1}{k}dx$$

and taking absolute value on both sides yields the bound

$$|\alpha_k| \leq \frac{2V}{\pi k}$$

which we see is the same for all ν. □

Corollary 5.6. *If f is as above, then for the chebyshev approximation f_n we have*

$$f - f_n \leq \frac{2V}{\pi\nu(n-\nu)^{\nu}}.$$

The idea of this is just we lose a power when taking an integral, that is, a sum.

We can even better the result if f is analytic.

Def 5.7. *We call the map* $x = \frac{1}{2}\left(z + \frac{1}{z}\right)$ *the Joukowsky map.*

For $\rho > 1$, the image of the circle of radius ρ under the Joukowsky map is an ellipse with foci at ± 1. We call then the Bernstein ellipses and their interiors are denoted E_ρ.

Note that for $\rho = 1$ the circle is just mapped to $[-1, 1]$.

Theorem 5.8. *Suppose* f *on* $[-1, 1]$ *is analytically continuable to* E_ρ *and is also bounded in that ellipse, i.e.* $|f(x)| \leq M$*, then we get*

$$|\alpha_0| \leq M; \qquad |\alpha_k| \leq 2M\rho^{-k}.$$

Proof. For $k = 0$ it is obvious.

For $k > 0$ we have

$$\alpha_k = \frac{1}{\pi i}\int_{|z|=1} F(z)z^{k-1}dz = \frac{1}{\pi i}\int_{|z|=\rho} F(z)z^{k-1}dz$$

and, just for simplicity, we assume that the bound M also works for the boundary (which we need not) then we know

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

as we claimed. □

Corollary 5.9.

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

What we see is that since the shape is an ellipse, we can deal with singularities out of the end points much more elegantly than those in the middle. One way to get around this is to cut the interval into two pieces and using two sets of ellipses to approximate. Then we're in better shape.

6. 4/5: Interpolations: Vandermonde matrix; Lagrange interpolation

Today we think about interpolations. The problem is this: say we have points $x_0, \dots, x_n$ and evaluations $f_0, \dots, f_n$ from an unknown function f. And of course for all the following to make sense we'll have to assume some reasonable structure of f.

We might face two problems:

- Experiments: We are given data points and are asked to find some good polynomials to interpolate.
- Experimental Design: We design which points to choose in order that for some basis the convergence is good.

More questions on this include the following:

- Given $x \in [-1, 1]$, can we find $\tilde{f}(x) \approx f(x)$?
- How do we find it?
- How do we measure the accuracy and how does that depend on the locations of points and properties of f.
- Can we characterize good points?
- Can we actually do this with floating point arithmetics in a reasonable amount of time?

So we will start with polynomials and do with 2 general ideas.

6.1. **Idea 1: Construct Vandermonde Matrix.**

We define

$$V := \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix}; \qquad \text{and} \qquad \vec{f} := \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

and define

$$\vec{\alpha} := \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}.$$

And to get the solution we just type $\vec{\alpha} = V \backslash \vec{f}$ and for each x we set $\delta_x = (1, x, x^2, \dots, x^n)^T$ then we have

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

The problem with this genuine method is that V is horribly conditioned so that small errors in f are magnified. Thus, the coefficients may be off. Weirdly, a 2023 research says that Yes and No: even though $||\vec{\alpha} - \alpha||$ might be large, $||V\vec{\alpha} - V\alpha||$ might still be small.

Cosmestically, it takes $\Theta(n^2)$ to $\Theta(n^3)$ operations to do this.

Remark 6.1. *Note that there's absolutely nothing that stops us from letting $V_{ij} = \phi_j(x_i)$ and let the corresponding $\vec{\beta} = V^{-1}\vec{f}$ be our solution as long as ϕ_i is a good basis.*

6.2. **Idea 2: Lagrange Interpolation.**

Let

$$l_i(x) = \prod_{j \neq i}(x - x_j) \Big/ \prod_{j \neq i}(x_i - x_j)$$

then l_i is a polynomial of degree n that vanishes at x_j for $j \neq i$, with $l_i(x_i) = 1$. So we can write $l_i(x_j) = \delta_{ij}$. Then for $f \in P_n$ we know

$$f = \sum_{i=0}^{n} f(x_i) l_i(x).$$

Def 6.2. *The Lebesgue constant Λ associated with $x_0, \dots, x_n$ is defined via*

$$\Lambda = \sup_{f \in C[a,b]} \frac{||Pf||_\infty}{||\vec{f}||_\infty}$$

where $Pf = \sum_{i=0}^{n} f(x_i) l_i(x)$. Thus, if we define $\lambda(x) = \sum_{i=0}^{n} |l_i(x)|$ then we have

$$\Lambda = \sup_x |\lambda(x)|.$$

Theorem 6.3. *Let $x_0, \dots, x_n \in [a, b]$ and Λ be the Lebesgue constant. Given f and let p be it's Lagrange approximation and $p_* \in P_n$ be it's best polynomial approximation. Then we have*

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

Proof. The trick is the same as before, note that $p - p_* \in P_n$ so we have

$$(f - p) - (f - p_*) = \sum_{j=0}^{n} \left[(f - p) - (f - p_*)\right] \Big|_{x = x_j} l_j(x)$$

and by definition of Lagrange interpolation the $(f - p)(x_j)$ term goes away and we have

$$(f - p) = (f - p_*) - \sum_{j=0}^{n} \left[(f - p_*)\right] \Big|_{x = x_j} l_j(x)$$

where takin absolute value we have

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

□

Theorem 6.4. *On* $[-1, 1]$*, let* Λ_n *be the Lebesgue constant for* $n+1$ *distinct points in* $[-1, 1]$*, then we have*

(a) $\Lambda_n \geq \frac{2}{\pi} \log(n+1) + \frac{2}{k}\left[\gamma + \log\left(\frac{4}{\pi}\right)\right]$.

(b) For Chebyshev points we have

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

(c) For equi-spaced points we have

$$\Lambda_n \geq \frac{2^{n-2}}{n^2} \sim \frac{2^{n+1}}{en \log n}$$

which is exponentially in n.

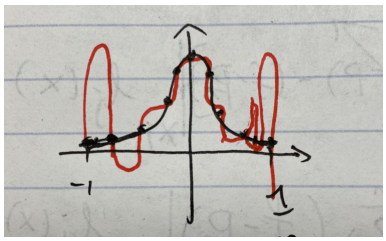
The proof for (b) is in the notes.

Remark 6.5. *Note that it is indeed reasonable that* Λ_n *should increase with n since the term* $||f - p_*||$ *is decreasing fast. So this really is fine.*

Note that the equispaced points has Λ_n acutually growing exponentially in n. This is corresponds to what is often denoted Runge's Phenomena: That interpolation from equispaced points are exponentially ill-conditioned.

Example 6.6. *The Witch of Agnesi*

Suppose we have $f(x) = \frac{1}{1 + a^2 x^2}$ ans say we choose $a = 5$ with 11 equispaced points. Then we get the interpolation like this:



Now we let $l(x) = \prod_{j=0}^{n}(x - x_j)$ and p be the Lagrange interpolation of a nice guy f. For $x \neq x_j$ define

$$\phi_x(t) = f(t) - p(t) - \frac{f(x) - p(x)}{l(x)} l(t)$$

then we know that $\phi_x(x_i) = 0 = \phi_x(x)$ has $n+2$ roots. Thus we take $n+1$ derivatives and get that $\phi_x^{(n+1)}$ has at least one root, call it ξ_x. Then we know (using definition of $l(x)$)

$$0 = \phi_x^{(n+1)}(\xi_n) = f^{(n+1)}(\xi_x) - \frac{f(x) - p(x)}{l(x)}(n+1)!$$

and hence

$$|f(x) - p(x)| \leq \frac{|f^{(n+1)}(\xi_x) l(x)|}{(n+1)!} \leq \frac{||l|| \cdot ||f^{(n+1)||}}{(n+1)!}.$$

Now we go back to the example of witch and see what this tells us. Note that roughly the maximum error occurs as the middle point in the end most intervals. So we can get the rough estimate (here h is the step size and $n = \frac{2}{h}$)

$$||l|| \sim \left(\frac{h}{2}\right) \cdot \left(\frac{h}{2}\right) \cdot \left(\frac{3h}{2}\right) \cdots \left(\left(n - \frac{1}{2}\right)\frac{h}{2}\right) \sim h^{n+1} n!.$$

Also, just knowing what f is we can compute

$$f^{(n+1)} = \frac{1}{2}\frac{d^n}{dx^n}\left(\frac{1}{1+iax} + \frac{1}{1-iax}\right) = \frac{(ia)^n}{2} n! \left(\frac{(-1)^n}{(1+iax)^{n+1}} - \frac{1}{(1+iax)^{n+1}}\right)$$

so

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

and putting things together we have

$$||f - p|| \sim \frac{1}{(n+1)!} h^{n+1} (n!)^2 a^n \overset{Sterling}{\sim} \frac{2}{n(n+1)} \left(\frac{2a}{n}\right)^n \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \sim \frac{\sqrt{8\pi}}{n^{3/2}} \left(\frac{2a}{e}\right)^n$$

and we see that if $2a < e$ we actually expect convergence!

The insight is that there no such saying as "equispaced points are just bad". We should investigate cases.

7. 4/10: Some potential theory; Barycentric polynomials

We know from last time that Runge phenomenon is bad when a large (say we are thinking of the witch), then there's pole at $\pm\frac{1}{a}i$, which if included inside the ellipse it will blow up. So today we try to kill Runge.

We can try the following:

- Change points: we've shown that Chebyshev points are very good. Roughly $O(1/n^2)$ is enough.
- Change problem: We can sacrifice the condition of passing through that point, but rather change it to close enough at the point.
- Change the approximation space. We can either do a larger polynomial space and minimize the error, or do with smaller space and find the best combo of points chosen.

7.1. **Potential theory, Hermite Intergral, Formulea and Interpolation.**

Let $x_0, \dots, x_n$ be the points and l_j be the lagrange interpolants. And define the polynomial

$$l(x) = \prod_{j=0}^{n}(x - x_j)$$

and we can check that

$$l_j(x) = \frac{l(x)}{l'(x_j)(x - x_j)}$$

and if f is analytic in a suitable region then by what we mean by interpolation

$$p(x) = \sum_{i=0}^{n} f(x_i) l_i(x).$$

Now by Cauchy's formula we have

$$l_j(x) = \frac{1}{2\pi i}\int_{\Gamma_j} \frac{l(x)}{l(t)(x-t)}dt$$

where Γ_j is a contour that encloses x_j, but non of the other x_k. This holds since the term is a residue.

Then we have

$$l_i f(x_i) = \frac{1}{2\pi i}\int_{\Gamma_i} \frac{l(x)f(t)}{l(t)(x-t)}dt$$

and thus if $x \notin \cap\Gamma_i$ and f analytic in side the union, we have

$$p(x) = \frac{1}{2\pi i}\sum_{i=0}^{n}\int_{\Gamma_i} \frac{l(x)f(t)}{l(t)(x-t)}dt = \frac{1}{2\pi i}\int_{\Gamma} \frac{l(x)f(t)}{l(t)(x-t)}dt$$

where we join the contours in a way that does not entail x.

Thus, the integrand has a simple pole at $x = t$ with residue $-f(x)$, so if we entend to $\tilde{\Gamma}$ that entails x. So we have the two Hermitian formula:

$$p(x) - f(x) = \frac{1}{2\pi i} \int_{\tilde{\Gamma}} \frac{l(x)f(t)}{l(t)(x-t)} dt$$

or alternatively

$$p(x) = \frac{1}{2\pi i} \int_{\tilde{\Gamma}} \frac{(l(x) - l(t))f(t)}{l(t)(x-t)} dt.$$

Thus we have

$$||p - f|| \leq \max_{t \in \tilde{\Gamma}} \left| \frac{l(x)}{l(t)} \right| \frac{1}{2\pi} \int_{\tilde{\Gamma}} \frac{f(t)}{x-t} dt.$$

Some connection to potential theory is if we let

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

and set

$$\alpha_n = \max_{x \in X, t \in T} \gamma_n(x,t)$$

then if $\alpha_n \geq \alpha > 1$, this means $|p - f| = \Theta(\alpha^{-n})$.

Another way to think of our points is to redefine a measure that's dense at the endpoints.

Some useful point of interpolation are:

- Fekete points: find $\{x_i\}_i$ that maximizes

$$\left(\prod_{j \neq k} |x_j - x_k| \right)^{\frac{2}{n(n+1)}}.$$

- Fejer points: obtained by a conformal mapping;
- Leja points: a greedy Fekete points, obtained by each time minimizing

$$\prod_{i=0}^{j-1} |x_j - x_i|$$

  while updating x_j.

7.2. **Numerical considerations:**

We start by thinking about the naive argument's run time. To form

$$l_i(x) = \prod_{j \neq i} (x - x_j) \Big/ \prod_{j \neq i} (x_i - x_j)$$

it takes $O(n^2)$ in the nominator, and $O(n^2)$ precomputations in the denominator. But if we let

$$w_j = \frac{1}{\prod_{j \neq i} (x_i - x_j)}; \quad l_j(x) = w_j \cdot \frac{l(x)}{(x - x_j)}$$

then we can compute l for each $x \sim n$ flop. And computing l_j needs only $O(1)$.

Now by interpolating

$$1 = \sum_{j=0}^{n} l_j(x) = \sum_{j=0}^{n} \frac{w_j l(x)}{x - x_j} = l(x) \sum_{j=0}^{n} \frac{w_j}{x - x_j}$$

we can actually cancel this $l(x)$! Since

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

Thus we have

$$Pf(x) = \frac{Pf(x)}{1} = \frac{\sum_{j=0}^{n} \frac{w_j f_j}{x - x_j}}{\sum_{j=0}^{n} \frac{w_j}{x - x_j}}$$

is what we call the Barycentric interpolation formula. Note that this canbe done with only $O(n)$ operations.

8. 4/12: Integration

Today we do integration. Suppose f is well-approximated by a set of basis $\phi_1, \dots, \phi_n$, and suppose that we have set of points $x_1, \dots, x_n$ such that the Vandermonde matrix is well-conditioned. Then we have

$$f \approx \sum_{i=1}^{n} \alpha_i \phi_i(x)$$

which means that

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} \alpha_i \phi_i(x)dx = \int_a^b \sum_{i=1}^{n} \alpha_n \int_a^b \phi_i(x)dx = \sum_{i=1}^{n} \alpha_i s_i = \alpha \cdot s$$

and from our discussion above (or homework) we know that we can compute the α vector via

$$\alpha_i \approx \sum_{j=1}^{n} (V^{-1})_{ij} f_j$$

and hence

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} \sum_{j=i}^{n} (V^{-1})_{ij} f_j s_i = \sum_{j=i}^{n} \left(\sum_{i=1}^{n} (V^{-1})_{ij} s_i \right) f_j = \sum_{j=1}^{n} w_j \cdot f_j$$

where we can precompute

$$w = (s^T \cdot V^{-1})^T$$

since s depends only on the basis functions, V depends only on the points.

Def 8.1. *We call the x_i quadrature nodes, and call w_i quadrature weights. Together, they give a quadrature rule.*

But in fact if we have $f = \sum_{i=1}^{n} \alpha_i \phi_i(x)$, then $f = \sum_{i=1}^{n} g_i(x) f_i$ with $f_i = f(x_i)$ What are the g_i? Are they dependent on f?

Thus if we have $f = \sum_{i=1}^{n} \alpha_i \phi_i(x)$ then

$$\int_a^b f(x)dx = \sum_{i=1}^{n} f_i \int_a^b g_i(x)dx.$$

Theorem 8.2. *Suppose S is a measure space, w is a non-negative real-valued integrable function on S, n is a positive integer, $f_1, \dots, f_n$ are bounded complex valued integrable functions on S, and $\varepsilon \leq 1$ is a positive real number. Then, there exists n complex numbers $w_1, \dots, w_n$ and n points $x_1, \dots, x_n \in S$ such that*

$$|w_k| \leq (1 + \varepsilon) \int w(x)dx, \quad \forall k = 1, \dots, n$$

and

$$\int f(x)w(x)dx = \sum_{k=1}^{n} w_k f(x_k)$$

for any f given by $f(x) = \sum_{j=1}^{n} c_j f_j(x)$ for some coefficients $c_1, \dots, c_n$.

Remark 8.3. *What could go wrong here is that even w_k are bounded, but if they are oscillating like $1000.1, -1000.1, 1000, \dots$, then because we are dealing with floating point arithmetic, then there's catastrophic cancellation around 5 digits.*

This is sufficient for most cases. But note that here we have n nodes x and n weights w. Thus we have in total 2n degrees of freedom to integrate n functions. Can we do better?

Preliminaries:

Recall that $H \subset C[a,b]$ is an n-dimensional Haar subspace if for any basis $\phi_1, \dots, \phi_n$ and any points $x_1, \dots, x_n$ the Vandermonde matrix is invertible.

We consider the general case in which x_i may be repeated. In this case we replace the repeated rows by derivatives of the ϕ_i (adding one order for each extra copy). i.e.

$$\tilde{V} = \begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) & \phi_3(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \phi_3(x_2) \\ \phi_1'(x_2) & \phi_2'(x_2) & \phi_3'(x_2) \end{pmatrix}$$

We call the above interpolation problem the "Hermite interpolation problem". It is solvable if the extended Vandermonde is always invertible for any choice of points.

Def 8.4. *We call the functions $\phi_1, \dots, \phi_n$ are called an extended Chebyshev system if they form a basis of an n-dimensional subspace of $C[a,b]$ and the Hermite interpolation problem is always solvable.*

Does this imply that ϕ has to be $n+$ smooth?

Theorem 8.5. *Let H be a Haar subspace of $C[a,b]$ spanned by an extended Chebyshev system $v_1, \dots, v_n$. Moreover, suppose $d\mu(x) = w(x)dx$ with a positive weight function $w \in C[a,b]$.*

Then, if $n = 2m$ is even, $\exists!$ set of points $a = x_0 < x_1 < \cdots < x_{m+1} = b$ such that

$$\int_a^b u d\mu(x) = \sum_{i=1}^{m} a_i u(x_i)$$

for all $u \in H$ where a_i are positive.

The proof uses the Borsuk antipodality theorem:

Lemma 8.6. *(Borsuk antipodality theorem) Let Ω be a bounded open symmetric neighborhood of 0 in $\mathbb{R}^{n+1}$ and suppose $T : \partial\Omega \to \mathbb{R}^n$ be an odd, continuous map. Then there exists some $x \in \partial\Omega$ for which $T(x) = 0$.*

What this really means is that if a circle is squeezed onto the center plane, some point must be mapped onto 0.

Proof. (Of Theorem 8.5)

Consider the set that contains the "vector of differences" between x_i. Define

$$S = \left\{ (y_0, \dots, y_m) : \sum_{i=0}^{m} |y_i| = b - a \right\}$$

so that the norm of dimensional of y are distances between x.

If $y \in S$ then set $x(y) = (a, a + |y_0|, a + |y_0| + |y_1|, \dots, b)$. So it's a map from S to $\mathbb{R}^n$.

So for $v \in C^k$ and $k \geq 1$ we define the symmetric function $v(x_0, x_1, \dots, x_k)$ iteratively by

$$v(x_0, x_1, \dots, x_k) = \begin{cases} \frac{v(x_1, x_2, \dots, x_k) - v(x_0, x_1, \dots, x_{k-1})}{x_k - x_0} & x_k \neq x_0 \\ \frac{1}{k!} u^{(k)}(x_0) & x_0 = x_1 = \cdots = x_k. \end{cases}$$

Now for $u \in H$ define

$$L_y(u) := \sum_{i=0}^{m} \operatorname{sgn}(y_i) \int_{x_i(y)}^{x_{i+1}(y)} u(x) d\mu - \sum_{j=1}^{n} b_j u(z_1, \dots, z_j)$$

where $z = (x_1, \dots, x_m, x_m, \dots, x_1)$, and b_j are chosen so that $L_y(v_j) = 0$ for all j.

Now it can be shown that b_j are continuous functions of y. Then we define

$$T : S \to \mathbb{R}^m; \quad T(y) = (b_n(y), b_{n-1}(y), \dots, b_{n+1-m}(y))$$

is continuous and odd. Thus, by Lemma we know $\exists y_*$ such that $T(y_*) = 0$.

Now if $s(x) = \operatorname{sgn}(y_*)$ then

$$\sum_{i=0}^{m} \operatorname{sgn}(y_i) \int_{x_i(y)}^{x_{i+1}(y)} u(x) d\mu = \int_a^b u(x) s(x) d\mu = \sum_{j=1}^{m} b_j(y_*) u(x_1^*, \dots, x_j^*)$$

where $x^* = x(y_*)$.

And just writing things according to multiplicity we have

$$\int_a^b u(x) s(x) d\mu = \sum_{j=1}^{l} \sum_{i=1}^{k_j} c_j^i u^{(i-1)}(x_j).$$

Now we want the integral inside the left part to be positive, so we choose u accordingly and such that $u^{(i)}(x_j) = 0$, $u(x)s(x) \geq 0$. If $l < m$ the this last condition can be guaranteed with fewer than m extra conditions. Thus we can always find such a function. Then the left hand side above is positive while the RHS vanishes. □

Conclusion of the above theorem: For a general class of subspaces, one an integrate 2n functions with n points and n positive quadrature weights. The positivity of weights helps us avoid catastrophic cancellation. Analogous results hold for spaces of odd dimension. Frequently, even when the system is not extended Chebyshev, one can exactly integrate n functions with $< n$ quadrature points. Finding them can be difficult.

8.1. **Gauss-Legendre quadrature and orthogonal polynomials:** We now focus on the case of polynomials P_n on $[-1, 1]$. Here we assume n is even. For odd n there's some floor signs involved so let's just assume we have them.

We first think about quadratures for integrating all polynomials in P_n, $n = 2m$. For all $p \in P_n$, we want to write it as a linear combination of our base polynomials in P_m. So we want to find points $x_1, \dots, x_{m+1}$ with weights $w_1, \dots, w_{m+1}$ such that

$$\int_{-1}^{1} q(x)\tilde{q}(x)dx = \sum_{j=1}^{m+1} q(x_j)\tilde{q}(x_j)w_j$$

where we can decompose ring of polynomial with $p = q\tilde{q}$. Thus, we define $T : P_m \to \mathbb{R}^{m+1}$ by

$$T(q) = \left(q(x_1)\sqrt{w_1}, \dots, q(x_{m+1})\sqrt{w_{m+1}}\right)^T$$

is an isometry from $P_m \in L^2$ to $l^2(\mathbb{R}^{m+1})$.

So we try to write out an orthogonal basis for P_m. The idea is that we suppose we have $p_0, \dots, p_k$ with $\langle p_i, p_j\rangle = 0$, $i \neq j$, $p_j \in P_j$. Then if $i < j - 1$

$$\langle p_j x, p_j\rangle = \langle p_j, xp_i\rangle = 0$$

where the right equality is because we are assuming that $p_j \perp P_{j-1}$, because $p_0, \dots, p_{j-1}$ span P_{j-1}.

So we set

$$a_j := \langle p_j x, p_{j+1}\rangle, \quad b_j := \langle p_j x, p_j\rangle, \quad c_j := \langle p_j x, p_{j-1}\rangle.$$

Note $a_j = c_{j+1}$ and by symmetry

$$\int_{-1}^{1} xp_j^2 dx = 0$$

and so $b_j = 0$. Thus, $(xp_j - c_j p_{j-1}) \perp \text{span}\{p_0, \dots, p_j\}$ and so P_{j+1}. we can see this because

- $\langle xp_j, p_j\rangle - \langle c_j p_{j-1}, p_j\rangle = 0 - 0$ since p_j is perpendicular to P_{j-1}.
- $\langle xp_j, p_{j-1}\rangle - \langle c_j p_{j-1}, p_{j-1}\rangle = \langle xp_j, p_{j-1}\rangle - \langle xp_j, p_{j-1}\rangle \cdot \langle p_{j-1}, p_{j-1}\rangle = 0$? Do we have normal?
- For $i < j - 1$ we can use the above argument that $\langle p_j x, p_j\rangle = \langle p_j, xp_i\rangle = 0$.

Thus, $p_{j+1} = A(xp_j - c_j p_{j-1})$ i.e.

$$xp_j = \alpha_j p_{j+1} + c_j p_{j-1}$$

now finding α_j and c_j is not strictly unique, so we follow the tradition of letting $p_j(1) = 1$ in which case $\alpha_j = \frac{j+1}{2j+1}, c_j = \frac{j}{2j+1}$ and thus

$$(j+1)p_{j+1} = (2j+1)xp_j - jp_{j-1} \quad p_0 = 1; p_1 = x$$

are our Legendre polynomials. We have that

$$\int_{-1}^{1} p_n p_m dx = \frac{2}{2n+1}\delta_{n,m}$$

so the norm is

$$||p_n|| = \sqrt{\frac{2}{2n+1}}.$$

So now we try to get the formula for Legendre quadratures. Let $n = 2m$ and for $q \in P_{2n}$ we can always write it as

$$q(x) = p(x)p_{m+1}(x) + r(x)$$

where $p \in P_{m-1}$ and $r \in P_m(x)$. Moreover, $p_{m+1} \in P_{m+1}$ has $m+1$ roots $x_1, \dots, x_{m+1}$ so

$$q(x_i) = p(x_i)p_{m+1}(x_i) + r(x_i) = r(x_i)$$

and since $p_{m+1} \perp P_m$ we have

$$\int_{-1}^{1} p(x)p_{m+1}dx = 0$$

and so $\int_{-1}^{1} q = \int_{-1}^{1} r$.

Now, since $r \in P_m$ there exist constants $c_0, \dots, c_m$ such that

$$r(x) = \sum_{j=0}^{m} c_j p_j(x)$$

so $q(x_i) = \sum_{j=0}^{m} c_j p_j(x_i)$ for all i. Setting $V_{ji} = p_j(x_i)$ we solve the matrix to obtain w, whose use we'll see later, by

$$Vw = (2, 0, \dots, 0)^T$$

then

$$\int_{-1}^{1} p_j(x)dx = \int_{-1}^{1} p_j(x) \cdot 1 dx = \langle p_j, p_0 \rangle = 2\delta_{j=0} = \sum_{i=1}^{m+1} p_j(x_i)w_i$$

and hence

$$\int_{-1}^{1} r(x)dx = \sum_{j=0}^{m} c_j \sum_{i=1}^{m+1} p_j(x_i)w_i = \sum_{i=0}^{m} r(x_i)w_i$$

This is just the "interpolation based quadrature" from before, applied to $p_0, \dots, p_m$. Now, if $q(x) \in P_m$, then

$$\int_{-1}^{1} q(x)dx = \int_{-1}^{1} p(x)p_{m+1}(x)dx + \int_{-1}^{1} r(x)dx = \int_{-1}^{1} r(x)dx = \sum_{i=1}^{m+1} r(x_i)w_i$$
$$= \sum_{i=1}^{m+1} [p(x_i)p_{m+1}(x_i) + r(x_i)]w_i = \sum_{i=1}^{m+1} q(x_i)w_i$$

And thus we have actually found a way to find the nodes and weights.

9. 4/19: SPLINES; ADAPTIVE DISCRETIZATION

Today we do approximation in a new way: we do it with polynomials in pieces. To do this we introduce splines, which is just to interpolate with fewer points each time and combine the result. One good thing about this is that the global approximation is easily polluted by outliers, but in this case since we're deviding by number of pieces, the pollution is minimal.

Let's say we want to interpolate many points and values with degree 2 polynomials. We are given n points, where the endpoints are not given, that is, on the interval $[x_0, x_{n+1}]$, we only know points x_1 through x_n. We want the interpolation to be smooth up till second order, so the derivatives match. So we have in total $n+1$ pieces, which gives us $3n+3$ degrees of freedom, but we only have $3n$ constraints (values agree (both side), derivatives agree). One easy way to count this is that we can freely choose 2 degrees of freedom at the left endpoint, but since we're to match derivatives every middle pieces is determined, and we're left with only 1 endpoint of freedom: so 3 real freedom.

This generates to polynomials with degree p. We have $p+1$ degrees of freedom, and to get a unique interpolant we impose an extra $p+1$ constraints. All possible polynomials are in the set:

$$S_p(x) := \left\{ s \in C^{p-1} \middle| s|_{[x_{i-1},x_i]} \in P_p, 0 \le i \le n+1 \right\}$$

where for convenience we let $x_0 = a$, $x_{n+1} = b$.

Now we focus on S_3, because choosing too much point just down right falls into global approximation. One of our choices is to require that the approximation be linear on the first and last interval, so we're cancelling 2 degree 2 terms and 2 degree 3 terms which left us with exactly no freedom. So we define this space to be

$$NS_3 := \left\{ s \in S_3 | s|_[a, x_1] \in P_1; s|_[x_n, b] \in P_1 \right\}$$

now if a spline passes through all points then it's called a interpolant.

To find natural interpolating cubic splines, we can do the same as for Lagrange interpolation. Let $\phi_j \in NS_3$, and $\phi_j = \delta_{ij}$ just by letting the nodes be $0, 0, 0, \dots, 1, \dots 0$, say. Then

$$s(x) = \sum_{i=1}^{n} f(x_i)\phi_i(x).$$

Ok we can do that, but why are they good? This is illustrated in the following proposition (geometry):

Proposition 9.1. *Given $f \in H^2$, $x_1, \dots, x_n \in [a, b]$ and $f_i = f(x_i)$. Then if s is the natural cubic interpolant, then $\langle f'' - s'', s'' \rangle = 0$.*

Before we see the proof we see why this is a pretty good property. Assuming the above is right, we have

$$\begin{aligned}\langle f'', f''\rangle &= \langle f'' - s'', f''\rangle + \langle s'', f''\rangle \\ &= \langle f'' - s'', f'' - s''\rangle + \langle s'', f'' - s'' + s''\rangle = ||f'' - s''||^2 + ||s''||^2\end{aligned}$$

where we've used the proposition in the last step. This means

$$||f''||^2 = ||f'' - s''||^2 + ||s''||^2$$

and thus for all function in H^2 that agrees with f at the interpolation points, s is the one that minimizes the second derivative norm (since we run the same argument for any "smaller g" to get they are not smaller). So:

Corollary 9.2. *s is the function in H^2 with $s(x_i) = f_i$ that minimizes $|\cdot|_2 := ||u''||$.*

Proof. (Property 9.1) since s'' is linear on each interval, we have

$$\begin{aligned}\int_{x_i}^{x_{i+1}} (f'' - s'')s''dx &= -\int_{x_i}^{x_{i+1}} (f' - s')s'''dx + (f' - s')s''\Big|_{x_i}^{x_{i+1}} \\ &= \int_{x_i}^{x_{i+1}} (f - s)s''''dx + (f' - s')s''\Big|_{x_i}^{x_{i+1}} - (f - s)s'''\Big|_{x_i}^{x_{i+1}} \\ &= (f' - s')s''\Big|_{x_i}^{x_{i+1}}\end{aligned}$$

where the first and last 0 are obvious. For the middle part, we note that this turns out to be an oscillating sum after summation, and since s'' vanishes on the endpoints, we get the desired result. □

Now clearly (since we can write polynomials as polynomials plus each other)

$$s(x) = \sum_{j=1}^{n} \alpha_j (x - x_j)_+^3 + \sum_{j=0}^{3} \beta_j x^j$$

and the choice of natural polynomials means that $\beta_2 = \beta_3 = 0$. On the last interval we have

$$s(x) = \sum_{j=1}^{n} \alpha_j (x - x_j)^3 + \beta_0 + x\beta_1$$

and our assumptions that quadratic and cubic terms vanish on this inteval means

$$\sum_{j=1}^{n} \alpha_j = 0; \quad \sum_{j=1}^{n} \alpha_j x_j = 0.$$

To be more clear we can write $x_+^3 = \dfrac{|x|^3 + x^3}{2}$ so on the whole interval we have

$$s(x) = \sum_{j=1}^{n} \frac{\tilde{\alpha}_j}{2} |x - x_j|^3 + \tilde{\beta}_0 + x\tilde{\beta}_1$$

because the higher order terms vanish by above condition. So interpolating spline are of the form

$$s(x) = \sum \alpha_j \phi(|x - x_j|) + p$$

where $p \in P_3$. And here we've chosen $\phi = r_+^3$, but of course there's other choices.

Think about more general ϕ and more general approximation spaces leads naturally to radial basis functions, but that we'll not go over here.

Proposition 9.3. *For*

$$s = \sum_{j=1}^{n} \alpha_j \phi(|x - x_j|) + p$$

and

$$\tilde{s} = \sum_{j=1}^{n} \tilde{\alpha}_j \phi(|x - y_j|) + \tilde{p}$$

are 2 natural splines on the interval $[a, b]$*. Then*

$$\langle s'', \tilde{s}'' \rangle = 12 \sum_{j=1}^{n} \sum_{l=1}^{m} \alpha_j \beta_l \phi(|x_j - y_l|) = \vec{\beta}^T K \vec{\alpha}.$$

Note that if $x = y$ *we have*

$$K_{ij} = \phi(|x_i - x_j|).$$

Proposition 9.4. *There exists a constant* $c > 0$ *such that for all* $f \in H^2$*,*

$$||f - s||_\infty \leq C h_x^{3/2} ||f''||_{L^2}$$

where s *is the natural spline, and* h *is the maximal separation.*

Remark 9.5. *Note that this method means that maybe equispace points are the best. This makes sense because there's no issue of Runge phenomenon.*

Now we move on to adaptive discretization.

We might face the following condition of a function for us to integrate: it behaves fine some where, bad some where, and really really bad or explodes somewhere else. We know that global approximation does work eventually, but too slow, so we use the following algorithm to do it:

(1) Use first 16, say, Chebyshev polynomials to approximate. If the approximation coefficient decays pretty well we are done;
(2) If the coefficients does not decay, then we cut the space into half, then do the above for each part.
(3) Recursively do this we'll be done. And integrate over each piece.

10. 4/26: Conjugate Gradient

We want to solve the system $Ax = b$, and start with an initial guess and construct a sequence of approximations. Then, at each step we choose a direction d_n and step in that direction.

Now, what we do below "a.s." contains some fence error.

So first we want the error in one direction each step to be orthogonal to the direction, i.e. we exhaust all differences in that direction, that is

$$\langle x - (x_n + \gamma_n d_n), d_n \rangle = 0.$$

If we can do that then we are just done. But we do not know where is x! So rather than using orthogonality as it is, we can try to work with another inner product. And since all norms are equivalent, more or less, we know that one convergence implies another. So our natural choice is the A-norm, and for this to really be a norm we'd require A to be symmetric, positive definite (we can of course try solve the system $A^*Ax = A^*D$ when the conditions are not satisfied, but then the condition number is doubled). So all things considered we

$$\langle u, v \rangle_A := u^T A v; \qquad ||u||_A = \langle u, u \rangle_A^{1/2}$$

is well-defined. So rather than focusing on the true error, we can try to instead focusing on the residual

$$r_n = A(x - x_0) = b - Ax_n$$

and require d to be A-orthogonal to $x - x_{n+1}$, that is

$$\langle d_n, x - x_{n+1} \rangle_A = \langle d_n, A(x - x_{n+1}) \rangle = \langle d_n, b - Ax_{n+1} \rangle$$

where now everything is known. So again writing out $x_{n+1} = x_n + \gamma_n d_n$ we have

$$\langle x - x_{n+1}, Ad_n \rangle = 0 \Rightarrow \langle r_n, d_n \rangle - \gamma_n \langle d_n, Ad_n \rangle = 0$$

$$\Rightarrow \gamma_n = \frac{\langle r_n, d_n \rangle}{\langle d_n, d_n \rangle_A}$$

and to get the direction d_{n+1} we follow our instinct, that is, since we are going to exhaust all possible direction of errors, then how about just letting d_n be the result of Gram Schimidt. So if we can find a suitable set such that the space is spanned by $\{u_0, \dots, u_n\}$ that are A-orthogonal, then we can let $d_0 = u_0$, and let

$$d_i = u_i + \sum_{k=0}^{i-1} \beta_{ik} dk$$

which is just Gram-Schimidt.

$$\langle d_i, d_j \rangle_A = \langle u_i, d_j \rangle_A + \sum_{k=0}^{i-1} \beta_{ik} \langle d_k, d_j \rangle$$

and the usual trick results in

$$\beta_{ij} = -\frac{\langle u_i, d_j \rangle_A}{\langle d_j, d_j \rangle_A}$$

where we note that by its construction we have to have $j < i$, so it's a purely upper diagonal matrix's terms. Now we define $x - x_0 = e_0$ and so for some α_i we have

$$e_i = e_0 + \sum_{j=0}^{i-1} \alpha_j d_j.$$

Again, the idea is that each time we exhaust a direction, so if our dimension is finite, then

$$e_0 = \sum_{j=0}^{N-1} \sigma_j d_j$$

and we know $\sigma_j = -\alpha_j$ and thus

$$e_i = \sum_{j=i}^{N-1} \alpha_j d_j.$$

If $v \in \text{span}\{d_0, \dots, d_{i-1}\}$, then

$$\begin{aligned} ||v - e_0||_A^2 = \langle v - e_0, v - e_0 \rangle_A &= \sum_{j=0}^{i-1} (v_j - \sigma_j)^2 \langle d_j, d_j \rangle_A + \sum_{j=i}^{N-1} \sigma_j^2 \langle d_j, d_j \rangle_A \\ &\geq \sum_{j=i}^{N-1} \sigma_j^2 \langle d_j, d_j \rangle_A = ||e_i||_A^2 \end{aligned}$$

and this tells us that

Proposition 10.1. $\sum_{j=0}^{i-1} \alpha_j d_j$ *is the best approximation to* e_0 *from the space.*

Now this all seems well and good, we're guaranteed to get to the point in N steps. But one thing is that when N large we're really doing nothing than Gauss Elimination, and the other is that due to machine error, a little error in the first iteration can lead to gigantic missing in the first few directions. Moreover, there's yet many choice of u that we can have. So all things considered, we want to let u each time be the $u_i = r_i$.

This is in fact very natural, since if we know where x is we'd rather just go in that direction. But we just know the A-error, so let's go that direction.

It's not hard to show that with this choice, $r_i \cdot r_j = 0$ and since $x_{i+1} = x_i + \gamma_i d_i$ we have

$$r_{i+1} = A(x - x_{i+1}) = A(x - x_i - \gamma_i d_i) = r_i - \gamma_n A d_i$$

from which we have a 3 term recurrence relation between γ, r, d.

Moreover, we have from above reasoning that

$$d_i = r_i + \sum_{j=0}^{i-1} \beta_{ij} d_j$$

where

$$\beta_{ij} = -\frac{\langle r_i, d_j \rangle_A}{\langle d_j, d_j \rangle_A}.$$

Now compute

$$\langle r_j, r_{i+1} \rangle = \langle r_j, r_i \rangle - \gamma_i \langle r_j, d_i \rangle_A \Rightarrow \langle r_j, d_i \rangle_A = \frac{1}{\gamma_i} \left[\langle r_j, r_i \rangle - \langle r_j, r_{i+1} \rangle \right]$$

but what do we know already? From that exercise above we know that r_i are orthogonal, so for most i, j we know $\langle r_j, d_i \rangle_A = 0$, except when $i = j - 1$, in which case the β is on the upper off-diagonal, for which we label

$$\beta_i := \beta_{i,i+1} = \frac{1}{\gamma_{i-1}} \frac{\langle r_i, r_i \rangle}{\langle d_{i-1}, d_{i-1} \rangle_A}$$

and remember from above that

$$\gamma_n = \frac{\langle r_n, d_n \rangle}{\langle d_n, d_n \rangle_A} \Rightarrow \beta_i = \frac{\langle d_{i-1}, d_{i-1} \rangle_A}{\langle r_{i-1}, d_{i-1} \rangle} \frac{\langle r_i, r_i \rangle}{\langle d_{i-1}, d_{i-1} \rangle_A} = \frac{\langle r_i, r_i \rangle}{\langle r_{i-1}, d_{i-1} \rangle} = \frac{\langle r_i, r_i \rangle}{\langle r_{i-1}, r_{i-1} \rangle}$$

where the last equality is because

$$d_i = r_i + \sum_{j=0}^{i-1} \beta_{ij} d_j = r_i + \sum_{j=0}^{i-1} a_j r_i$$

due to recursively plugging in, and r_i are orthogonal.

The Algorithm:

- $d_0 = r_0 = b - Ax_0$.
- $\gamma_i = \dfrac{\langle r_i, r_i \rangle}{\langle d_i, d_i \rangle_A}$.
- $x_{i+1} = x_i + \gamma_i d_i$.
- $r_{i+1} = r_i - \gamma_i A d_i$.
- $\beta_{i+1} = \dfrac{\langle r_{i+1}, r_{i+1} \rangle}{\langle r_i, r_i \rangle}$.
- $d_{i+1} = r_{i+1} + \beta_{i+1} d_i$.

and so we can write this in a polynomial way using recurrence: $e_i = P(A)e_0$, $P(0) = I$. And to analyze this we diagonlize into

$$e_0 = \sum_j \xi_j v_j$$

where v_j is the eigenvector of A. Then

$$e_i = \sum_j P_i(\lambda_j)\xi_j v_j$$

and

$$Ae_i = \sum_j P_i(\lambda_j)\lambda_j\xi_j v_j.$$

$$||e_i||_A^2 = \langle e_i, Ae_i\rangle = \sum_j \xi_j^2 (P_i(\lambda_j))^2\lambda_j$$

where we want to have the worst b, which means there's a worst b with $||\xi|| = 1$. Then the above has a bound (maximum λ because we let ξ be in that direction) since we can pick our starting point with respect to any polynomial we want to pick

$$\leq \min_{p\in P_i}\max_{\lambda\in\Lambda(A)} |P_i(\lambda)|^2||e_0||_A^2 = \min_{p_i\in P_i}\max_{\lambda\in[\lambda\,\max,\lambda\,\min]} \frac{|P_i(\lambda)|}{|P_i(0)|}.$$

Moreover, we have

Conjecture 1. *(Jeremy's weak conjecture) the scaled and shifted Chebyshev polynomials are optimal in the above sense.*

Proof. Let's WLOG assume that $[\lambda\,\min, \lambda\,\max] = [-1, 1]$,then since A is positive definite we know that 0 is mapped to some s such that $s < -1$. Thus we want to show that T_i is optimal, that is for all $q \in P_i$ we have

$$\frac{|T_i(x)|}{|T_i(s)|} \geq \frac{|q(x)|}{|q(s)|}$$

whereas we know that T_i is equioscillating in the interval.

First we define the difference

$$p(x) = \frac{q(x)}{|q(s)|} - \frac{T_i(x)}{|T_i(s)|}$$

then it is a polynomial of degree i, so it has i roots at most if non-zero. If it's zero then we're done.

Note that one of the roots is at $x = s < -1$ (let's just pick the sign of q). For the other ones, we know that T_i is equioscillating from 1 to -1 (just knowing what it is) between -1 and 1, then we know at the points where $T_i(x) = 1$, we have

$$p(x) = \frac{q(x)}{|q(s)|} - \frac{1}{|T_i(s)|} \leq 0$$

and at point that's -1 we get

$$p(x) = \frac{q(x)}{|q(s)|} + \frac{1}{|T_i(s)|} \geq 0$$

and thus there's $i + 1$ roots of p by continuity, thus $p = 0$ and we are done. □

11. 5/1: Conjugate gradient continue; Trigonometric approximation

11.1. More on conjugate gradient.

Recall from last time that the algorithm for computing the conjugate gradient is

The Algorithm:

- $d_0 = r_0 = b - Ax_0$.
- $\gamma_i = \dfrac{\langle r_i, r_i \rangle}{\langle d_i, d_i \rangle_A}$.
- $x_{i+1} = x_i + \gamma_i d_i$.
- $r_{i+1} = r_i - \gamma_i A d_i$.
- $\beta_{i+1} = \dfrac{\langle r_{i+1}, r_{i+1} \rangle}{\langle r_i, r_i \rangle}$.
- $d_{i+1} = r_{i+1} + \beta_{i+1} d_i$.

And an analysis on this is that we note

$$x_i = x_0 + \sum_{j=0}^{i-1} c_j r_j$$

where if you look really close enough at the algorithm, you will note that $\sum_{j=0}^{i-1} c_j r_j$ is a polynomial of A acting on r_0. Thus we can write it as $P(A)r_0 = P(A)Ae_0$. Thus

$$e_i = x - x_i = (x - x_0) - P(A)Ae_0 = e_0 - P(A)Ae_i = q(A)e_0$$

where $q(0) = I$ just by the above deduction.

Thus, diagonalizing A gives eigenvalues λ_j and eigenvectors v_j. So we have (for ξ being coefficients)

$$e_0 = \sum_j \xi_j v_j \Rightarrow e_i = q_i(A) \sum_j \xi_j v_j = \sum_j \xi_j q_i(\lambda_j) v_j$$

and hence

$$Ae_i = \sum_j \xi_j q_i(\lambda_j) \lambda_j v_j$$

and by definition of A-norm we have (by orthonormal)

$$||e_i||_A^2 = \sum_j \xi_j^2 (q_i(\lambda_j))^2 \lambda_j$$

and we want to choose the coefficients ξ_j such that the above is maximized. So we just normalize e_0 and thus assume $||\xi|| = 1$ so the maximal is just when $\xi_j = 1$ for the largest $(q_i(\lambda_j))^2 \lambda_j$ and so we obtain

$$||e_i||_A^2 \leq \min_{q_i; q_i(0)=1} \max_{\lambda \in \Lambda(A)} |q_i(\lambda)|^2 ||e_0||_A^2$$

but this is requiring that we know the whole spectrum, yet for a positive definite matrix know all the spectrum is the same as solving the system. So we losen the requirement a little bit and say we want to use the maximum in the whole spectral range, i.e. $\lambda \in [\lambda\min, \lambda\max]$. So the problem boils down to

$$\begin{aligned}||e_i||_A^2 &\leq \min_{q_i;q_i(0)=1} \max_{\lambda\in\Lambda(A)} |q_i(\lambda)|^2||e_0||_A^2 \leq \min_{q_i;q_i(0)=1} \max_{\lambda\in[\lambda\min,\lambda\max]} |q_i(\lambda)|^2||e_0||_A^2 \\ &= ||e_0||_A^2 \max_{\lambda\in[\lambda\min,\lambda\max]} \frac{|p(\lambda)|}{|p(0)|}.\end{aligned}$$

So the problem is to find the best polynomial $p \in P_i$ that is large at 0, some place out side of it's range of spectrum, and small with in the whole range of spectrum. This naturally is just the Chebyshev polynomials. So the first thing we do is we shift the interval to $[-1, 1]$, then it's really Chebyshev. Say that after the shift 0 is mapped to $s_0 = -2\frac{\lambda\min}{\lambda\max - \lambda\min} - 1$ then we can use Jeremy's weak conjecture (not at all a conjecture, proven above) to know that Chebyshev is indeed the best.

All's well and good except we still don't have any idea how fast the convergence rate is. So we compute

$$||e_i||_A \leq T_i\left(-\frac{\lambda\max+\lambda\min}{\lambda\max-\lambda\min}\right)^{-1} ||e_0||_A$$

and knowing Chebyshev we have

$$|T_i(x)| = T_i(|x|)$$

for $|x| \geq 1$ and if we define $\kappa = \frac{\lambda\max}{\lambda\min}$ to be the conditional number of A, then we have the bound

$$||e_i||_A \leq T_i\left(\frac{\kappa+1}{\kappa-1}\right)^{-1} ||e_0||_A \leq cos\left[i\arccos\left(\frac{\kappa+1}{\kappa-1}\right)\right]^{-1} ||e_0||_A$$

where after some hard trial of trigonometric we have

$$||e_i||_A \leq 2\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-i} ||e_0||_A$$

where i is an index, not a complex number! This means $||e_i||$ decays exponentially.

Remark 11.1. *When $\kappa = 1$, above tells us that the decay is infinitely fast. But this makes sense becaue there's only 1 eigenvalue, and that means the matrix is very good.*

Moreover, if there's an outlier of eigenvalue, we just put a 0 there and do the same conjecture argument for the other clustered eigenvalues. This finishes the discussion.

One might ask what happens if A is not symmetric positive definite, then it turns out we can convert the problem to $\max_{z\in\Lambda(A)} \frac{|p_i(z)|}{|p_i(0)|}$ as long as the area of spectrum does not circumscribe 0. This gives rise to Krylov subspaces:

For the same $Ax = b$ system, consider $b, Ab, A^2b, \dots$. It will come to one point that they are linearly dependent, so

$$c_0 b + c_1 Ab + \cdots + c_k A^k b = 0$$

and we can solve

$$b = A\left(-\frac{1}{c_0}\sum_{j=1}^{k} c_j A^j b\right) =: Ax.$$

Same problem comes down to finding such polynomials. We skip here.

11.2. **Trigonometric approximation.**

Def 11.2. *Define the first form of trigonometric polynomials as*

$$T_n = a_0 + \sum_{k=1}^{n} \left(a_k \cos(kx) + b_k \sin(kx)\right)$$

where we note there's really $2n+1$ terms being added, so the dimension of space it spans, with some checking maybe, should be $2n+1$, but we denote this as T_n for obvious reasons: the sum is up to n. There's also a more natural form of course:

$$T_n = \sum_{k=-n}^{n} c_k e^{ikx}$$

but we need the equation to be real valued, so we require $c_{-k} = \overline{c_k}$ because then

$$c_k e^{ikx} + c_{-k} e^{-ikx} = 2\,\mathrm{Re}\left(c_k e^{ikx}\right).$$

Remember as in the deduction of the iterative formula of Chebyshev polynomials that

$$\cos(kx) + \cos((k-2)x) = 2\cos((k-1)x)\cos x$$

which means that $\cos(kx)$ can really be expressed as a polynomial in $\cos x$.

Now as for sin it cannot be written as purely sin polynomial, but we have a close cousin:

$$\sin((k+1)x) - \sin((k-1)x) = 2\cos(kx)\sin x$$

which means we can write

$$\sin((k+1)x) = Q_k(\cos(kx))\sin x$$

which is a k degree polynomial in cos times sin.

This, plus our first definition of T_n means

$$T_n = p(\cos x) + q(\cos x)\sin x$$

where $p \in P_n, q \in P_{n-1}$ and thus T_n is written as an even function plus an odd function.

Now let's say we want to use them to approximate any function, so our first guess is all continuous functions, but that's obviously too big a class since everything we have are periodic. So let's try approximate $C^{2\pi}$, that is, all 2π periodic continuous functions.

We note that this can be perfectly mapped to the unit circle in the complex plane so we will also write $C(\mathbb{T})$ when needed.

Theorem 11.3. *Suppose $f \in C^{2\pi}$ is given and $\varepsilon > 0$, then $\exists$ a triangular polynomial T such that $||f - T|| < \varepsilon$.*

One obvious way to do this is just because trigonometric polynomials form an algebra, so we are done by Stone-Wierstrass, but that's too easy and let's do some work here.

Proof. **Step 1: f even**:

We can restrict ourselves to $[0, \pi]$ if we forget about sin. And this is valid because f is even. So if we let $y = \cos x$ then by normal polynomial approximation we know there exists some $\tilde{p}$ such that

$$|f(\arccos(y)) - \tilde{p}(y)| < \varepsilon.$$

Thus plugging back what y is we get the result.

Step 2: general f:

In this case we first rewrite

$$f = \frac{f(x) + f(-x)}{2} + \frac{f(x) - f(-x)}{2}$$

as is the usual trick. Now we denote approximation polynomials as:

$$P_1 \approx \frac{f(x) - f(-x)}{2} \sin^2(x); \quad P_2 \approx \frac{f(x) - f(-x)}{2} \sin(x)$$

which is valid because both the right hand side are even. But then we note

$$\left| P_1 + (P_2 - \sin x) - f \sin^2(x) \right| < \varepsilon$$

are an approximation. So looks like we're one step away.

Now consider the magic that peels off $\sin^2(x)$. Note

$$\tilde{f}(x) = f(x)\cos^2(x) \Rightarrow \tilde{f}\left(x - \frac{\pi}{2}\right) = f\left(x - \frac{\pi}{2}\right)\sin^2(x)$$

so if we denote $g(x) = f\left(x - \frac{\pi}{2}\right)$ then we create the similar $P_g = P_{1,g} + P_{2,g}$ to get

$$\left| P_g - g(x)\sin^2 x \right| < \varepsilon \Rightarrow \left| P_g(x) - \tilde{f}\left(x - \frac{\pi}{2}\right) \right| < \varepsilon$$

and because P_g is a trigonometric polynomial, so is it's $\frac{\pi}{2}$ shifted version. So

$$\left| P_g\left(x - \frac{\pi}{2}\right) - f(x)\cos^2(x) \right| < \varepsilon$$

and thus we denote $P_3 = P_g\left(x - \frac{\pi}{2}\right)$ to obtain

$$\left| P_1 + P_2 + P_3 - f(x)(\sin^2 + \cos^2) \right| < 2\varepsilon$$

so we are done. □

12. 5/3: Approximation with Trigonometric Polynomials; Gibb's Phenomenon

12.1. **Best approximation with Trigonometric Polynomials.**

This section is a mimic of what we've done before. So many of the results are far from surprising.

Theorem 12.1. *Trig polynomials form a Haar subspace of $C(\mathbb{T})$ of dimension $2n+1$ as long as no $k, k+2\pi$ are chosen as interpolation points.*

Remark 12.2. *The additional requirement is just a condition on that the points should be separated (recall from Stone Wierstrass that the approximating algebra should separate points).*

We use the formula for Vandermonde matrix's determinant to solve this.

Lemma 12.3. *The determinant of the Vandermonde matrix*

$$V = \begin{bmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{bmatrix}$$

is

$$\det V = \prod_{0 \le j < k \le n} (x_k - x_j).$$

Proof. (Lemma)

We use induction. For $n = 1$ and $n = 2$ it is obvious.

Now, we view x_1 as a variable, denote as x and write the determinant as

$$\begin{vmatrix} 1 & x & \dots & x^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{vmatrix} = p(x) \in P_n(x)$$

where we note that for $x = x_i$, $1 \le i \le n$ we have that 2 rows of the determinant is the same so $p(x_i) = 0$. So we can write

$$p(x) = C \prod_{i=1}^{n} (x - x_i)$$

and we note that the coefficient is nothing but

$$C = (-1)^{n+1} \begin{vmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{vmatrix} \overset{IH}{=} (-1)^{n+1} \prod_{1 \le j < k \le n} (x_k - x_j)$$

and thus

$$p(x_0) = \prod_{0 \le j < k \le n} (x_k - x_j).$$

□

Proof. (of Theorem 12.1) One way to solve these kind of problems is just to let the Vandermonde be invertible. So we write out the Vandermonde matrix:

$$V = \begin{bmatrix} e^{-inx_0} & \dots & e^{inx_0} \\ \vdots & & \vdots \\ e^{-inx_{2n}} & \dots & e^{inx_{2n}} \end{bmatrix}$$

and require

$$V \begin{bmatrix} c_{-n} \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_{2n} \end{bmatrix}.$$

Just by what it is we are done if we can show $\det(V) \neq 0$. And we have determinant equality (using lemma 12.3):

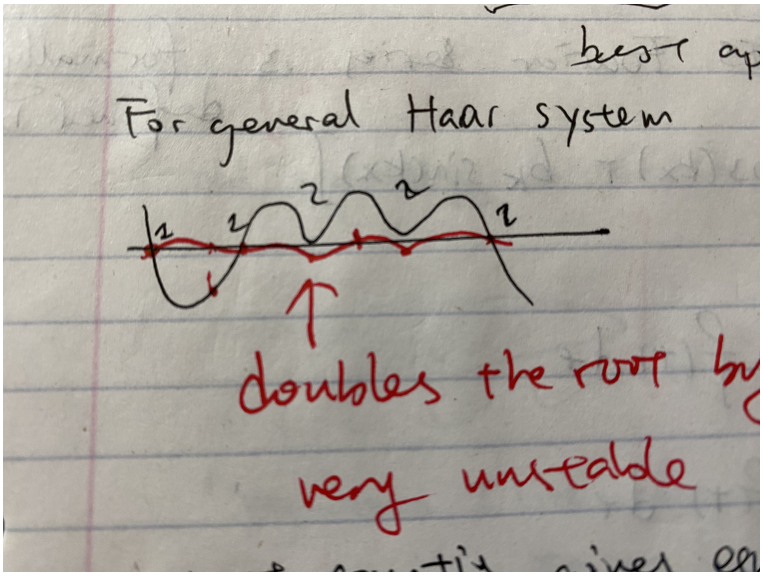
$$\det V = e^{-in(x_0+\cdots+x_{2n})} \begin{vmatrix} 1 & \dots & e^{2inx_0} \\ \vdots & & \vdots \\ 1 & \dots & e^{2inx_{2n}} \end{vmatrix} = e^{-in(x_0+\cdots+x_{2n})} \prod_{0 \le j < k \le 2n} \left(e^{ix_k} - e^{ix_j} \right) \neq 0$$

where the last term is different from 0 because $e^{i\theta}$ has norm 1 and for $x_j \neq x_i + 2k\pi$, non of the multiplied term is 0. □

Theorem 12.4. *If H is an n dimensional Haar system of $C[a, b]$, then it has the equi-oscillation property (the best approximation has dim+1 equi-oscillation points).*

Remark 12.5. *This makes us ponder: what really is behind the equi-oscillation property? The answer should be root counting, and Haar subspace gives us exactly root-counting.*

Remark 12.6. *In the generalization of the proof, we might encounter difficulties such as a zero of the approximating function in the algebra just touches $y = 0$ but does not go across it. In this case we count this point as 2 zeroes. This makes sense because we can do locally small perturbations and then there will be 2 roots created, then we get back. An illustration of this is attached:*



One further question about this is is that an iff condition, i.e. are all systems with equi-oscillation property Haar? The answer is not clear and Jeremy seems to recall larger systems, for instance space consists of functions like $e^{-\alpha x}$. Why do we care about them? The answer is that they are easy to translate: if a function can be represented as

$$f(x) = \sum_{i=1}^{n} e^{-\alpha_i x} w_i$$

then we know

$$f(x+h) = e^{-\alpha_i h} \sum_{i=1}^{n} e^{-\alpha_i x} w_i$$

and this is useful when computing w_i is hard and depends on the choice of x. Then at least if we want to do translation we won't have to suffer all the toils of computing w again. The key idea is sort of like when we do Fourier transform, that we diagonalize the operator.

Below are some concluding properties, very much alike the algebraic polynomial systems.

Corollary 12.7. *For all $f \in C^{2\pi}$, we have*

(1) f has a best approximation $T^* \in T_n$.
(2) $f - T^*$ has an alternating set containing $2n+2$ equi-oscillating points.
(3) T^* is unique.
(4) If $f - T$ has an alternating set of $2n+2$ equi-oscillating points, then $T = T^*$.

12.2. **Gibb's Phenomenon.**

Def 12.8. *Given $f \in C^{2\pi}$, its Fourier series is formally (of course) defined by*

$$f = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[a_k \cos(kx) + b_k \sin(kx) \right]$$

where

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(kt) f(t) dt$$

and

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \sin(kt) f(t) dt.$$

All is formally defined, but maybe let's see what will happen when we just blindly plug in a non-continuous f.

Example 12.9. *Fourier series for the sign function* $f = \operatorname{sgn}(x)$ *on* $[-\pi, \pi]$.

Here we assign $f(0) = 0$.

First, we note that the function is odd, so it's even coefficients, a_k, are 0. And to compute we have

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \sin(kt)\operatorname{sgn}(t) dt = \frac{2}{\pi k}\left(1 - (-1)^k\right)$$

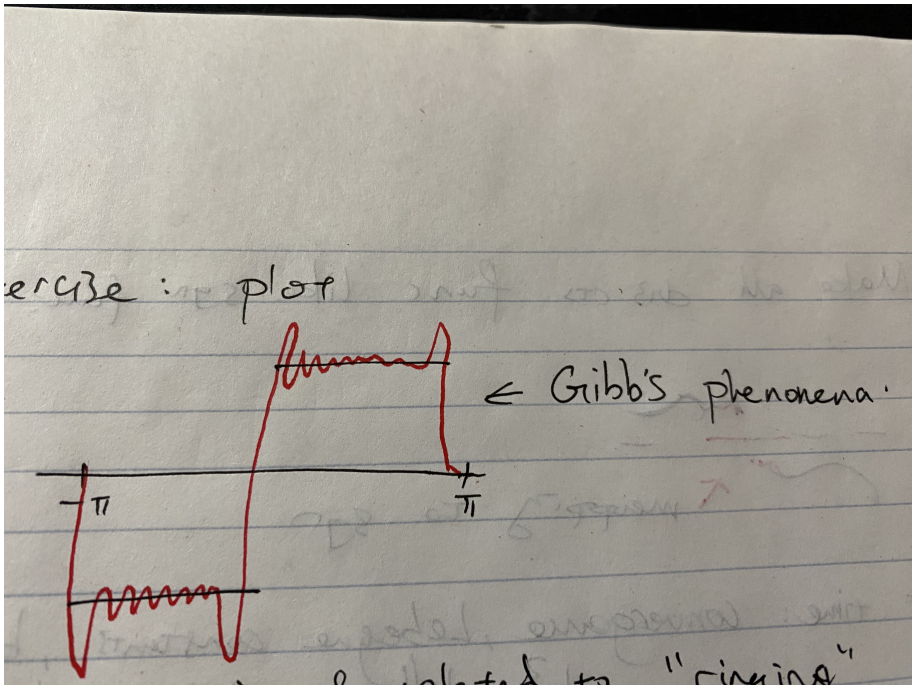
and let

$$S_m(f) := \sum_{k=1}^{m} b_k \sin(kx)$$

then we know

$$S_{2n}(f) := \sum_{k=1}^{n} \frac{4}{(2k-1)\pi} \sin((2k-1)x)$$

which gives us a really good illustration of the Gibb's phenomenon:



which in many fields such as signal processing is called ringing.

Joke 12.10. *When Jeremy was at Yale he worked just across Gibb's grave. "We used to hang out all the time..."*

To be explicit we have

$$\begin{aligned} S_{2n}(f)\left(\frac{\pi}{2n}\right) &= \frac{4}{\pi}\left(\sin\left(\frac{\pi}{2n}\right) + \frac{\sin\left(\frac{3\pi}{2n}\right)}{3} + \cdots + \frac{\sin\left(\frac{(2n-1)\pi}{2n}\right)}{2n-1}\right) \\ &= \frac{\pi}{2n}\frac{4}{\pi}\left(\frac{\sin\left(\frac{\pi}{2n}\right)}{\frac{\pi}{2n}} + \frac{\sin\left(\frac{3\pi}{2n}\right)}{\frac{3\pi}{2n}} + \cdots + \frac{\sin\left(\frac{(2n-1)\pi}{2n}\right)}{\frac{(2n-1)\pi}{2n}}\right) \\ &\approx \frac{2}{\pi}\int_0^\pi \frac{\sin t}{t} dt = 1 + 2 \cdot 0.0894898722... \end{aligned}$$

where the constant is called Wilbraham-Gibbs constant. So we can see that the constant at least does not grow with n.

Why do we care only about the sgn function? Because all discontinuity points can be represented by a local isomorphism with sgn function!

13. 5/8: Fourier Series; Convergence theory

One quick comment is that we know we have Gibb's phenomenon for discontinuity points, but what if I just connect the endpoints and smoothen it? The answer is that for the δ as the interval of decay at endpoints, as long as the frequency $> \delta$, we have a Gibb's-like phenomenon; but if the frequency is less than δ, then we expect exponential decay of Fourier series.

13.1. **Fourier Series.**

Remember last time we've defined Fourier series and Fourier coefficients in both sin and cos or in e^{ikx}. Then clearly by the expression we have

$$\frac{|a_k|}{2}, \frac{|b_k|}{2}, |c_k| \leq ||f||_\infty$$

and we can define partial sums.

Def 13.1. *The n-th partial sum is*

$$S_n(f)(x) := \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) = \sum_{k=-n}^{n} c_n e^{ikn}.$$

Proposition 13.2. $\left\{ \frac{1}{\sqrt{2\pi}}, \frac{\cos(x)}{\sqrt{2\pi}}, \frac{\cos(2x)}{\sqrt{2\pi}}, \dots, \frac{\sin(x)}{\sqrt{2\pi}}, \frac{\sin(2x)}{\sqrt{2\pi}} \right\}$ *is an orthonormal set with respect to L^2 inner product.*

Note that the definition of partial sum is just projection onto the first $2n+1$ dimensional space, so $S_n(1) = 1$ and

$$S_n(\cos(kx)) = \begin{cases} \cos(kx) & k \leq n \\ 0 & \text{else.} \end{cases}$$

Thus $S_n(T)(x) = T(x)$ for all $T \in T_n$. Now we note $S_n : C^{2\pi} \to T_n$ we have

$$S_n^2 = S_n$$

just like we'd expected.

Corollary 13.3. *If $f \in C^{2\pi}$ and $a_k = b_k = 0$ for all k, then $f = 0$.*

Proof. For T a trigonometric polynomial, then we have

$$\int_{-\pi}^{\pi} T f \, dx = 0$$

by projection. But then Wierstrass says $\exists \tilde{T}$ such that $||\tilde{T} - f||_\infty < \varepsilon$ so

$$\int_{-\pi}^{\pi} f^2 dx \leq \int_{-\pi}^{\pi} \tilde{T} f \, dx + \varepsilon \int_{-\pi}^{\pi} |f| dx = 0 + C\varepsilon$$

and thus $||f||_2 = 0$ so $f = 0$. □

One good text book is Schlag's Harmonic Analysis.

Corollary 13.4.

(1) $\frac{1}{\pi}\int_{-\pi}^{\pi}\left[S_n(f)(x)\right]^2 dx = \frac{a_0^2}{2} + \sum_{k=1}^{n}\left(a_k^2 + b_k^2\right) \leq \frac{1}{\pi}\int_{-\pi}^{\pi} f^2 dx.$

(2) $S_n(f) \to f$ *in* L^2 *sense for all* $f \in C^{2\pi}$.

Proof. For the first one just use orthogonality. For the second we note that there exists T such that $||f - T|| \leq \varepsilon$ and we find n large enough so that the tail of T's coordinates does not matter. Then

$$||f - S_n(f)||_2 \leq ||f - T||_2 + ||S_n(f - T)||_2 \leq 2||f - T||_2 \leq 2\sqrt{2\pi}\varepsilon$$

where the last step is because S_n is a projector, so it does not stretch (nor squeeze). □

So note we can write $S_n(f)(x) = \sum_{k=-n}^{n} c_k e^{ikx}$ then we can do some heavy real part computation to get

$$\begin{aligned} S_n(f)(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{k=-n}^{n} e^{ik(x-y)} f(y)dy = \frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\left[\sum_{k=0}^{n} e^{ik(x-y)} - \frac{1}{2}\right] f(y)dy \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\left[\frac{1 - e^{i(n+1)(x-y)}}{1 - e^{i(x-y)}} - \frac{1}{2}\right] f(y)dy = \frac{1}{2\pi}\int_{\pi}^{\pi}\frac{\sin(x-y)\,(n+1/2)}{\sin((x-y)/2)} f(y)dy \end{aligned}$$

where we define

Def 13.5. *The* Dirichlet Kernel *is*

$$D_n(x-y) := \frac{\sin(x-y)\,(n+1/2)}{\sin((x-y)/2)}.$$

One alternative way (more natural even) is to divide $e^{i(x-y)/2}$ to both nominator and denominator, then use Euler's formula to get result.

Proposition 13.6. *(Properties of* D_N*):*

(1) D_n *is continuous and* $\lim_{t\to 0} D_n(t) = 2n + 1.$

(2) $\frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)dt = 1.$

(3) $\frac{|\sin[(n+1/2)t]|}{|t/2|} \leq D_n(t) \leq \frac{\pi}{|t|}.$

(4) $\frac{8}{\pi}\log n \leq ||D_n||_1 \leq 6\pi + \pi\log n.$

Before we prove them, let's answer why we even need (4):

Theorem 13.7. *Given $f \in C^{2\pi}$ the we let T_* to be the best approximation from T_n, then let $\tilde{T}$ to be $S_n(f)$, then*

$$||f - \tilde{T}||_\infty \leq \left(1 + ||D_n||_1\right) ||f - T_*||\infty.$$

The proof of theorem 13.7 is just like every proof about this kind of Lebesgue constant proof, so we skip. Notably $||D_N||_1$ gives a bound for $||S_n||$ so that is why we need the result, which says that it is unbounded, and gives a rate for the unboundedness: since we know the unboundedness is only $\log n$, which is roughly a "constant" function (that's a joke) so if Fourier series is not doing so good, then the best approximation is not doing so good either! (worse than $\dfrac{1}{\log n}$ at least) so we don't even bother.

Now we prove the Proposition 13.6, which is quite basic.

Proof. (Proposition 13.6)

For (1), just use L'Hopital or just write $\sin(x) = x$ near 0. Alternatively, write it as $\displaystyle\sum_{k=-n}^{n} e^{ikx} = 2n + 1$.

For (2), take $f = 1$ then it's obvious.

For (3), use the bound $\sin x \leq x$ and $\sin x \geq \dfrac{1}{\pi}x$ on $[0, \pi]$ to get the result.

For (4), one direction is we just bound above with $1/|x|$ away from 0 and use constant to bound close to 0. So we have

$$\frac{1}{2}||D_N||_1 \leq \int_0^{1/n} (2n+1)dt + \int_{1/n}^{\pi} \frac{\pi}{|t|}dt = O(\log n).$$

Cool thing Jeremy wrote: Exercise: do better.

For the other direction there's not really a quick way so we just do integration piece by piece to get the result... so skip. □

Corollary 13.8. *$\exists f \in C^{2\pi}$ ofr which $||S_n(f)||$ is unbounded in n.*

We give an "unsatisfying proof" of this because we don't give a counterexample but just use uniform boundedness theorem. For a counterexample it's in Schlag's book.

Proof. $S_n : C^{2\pi} \to C^{2\pi}$ is continuous, but on the other hand, assume that

$$\sup_n ||S_n(f)|| < \infty, \forall f \in C^{2\pi}$$

then by Uniform Boundedness we have

$$\sup_n ||S_n|| < \infty.$$

But then we directly find the continuous version of $\text{sgn}(D_n)$ to notice

$$||S_n|| \geq ||D_n|| = \Theta(\log n) \to \infty$$

so we get contradiction. □

Now we try to get a discrete version of Plancherel's. We have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ikx}(f * g)(x)dx = \frac{1}{2\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-ikx} f(y)g(x-y)dydx$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{iky} f(y)dy \int_{-\pi}^{\pi} e^{-ikz} g(z)dz = 2\pi \hat{f}_k \hat{g}_k$$

by a change of variable so

$$\widehat{(f * g)}_k = 2\pi \hat{f}_k \hat{g}_k.$$

Thus

$$S_n(f) = \frac{1}{2\pi} D_n * f \Rightarrow \left(\widehat{S_n(f)}\right)_k = \widehat{(D_n)}_k \hat{f}_k$$

where we know, by definition of what is a Fourier coefficient, that

$$\widehat{(D_n)}_k = 1$$

for all $k = -n, \dots, n$, and 0 else where. So we can view this as a diagonalization:

$$S_n(f) = F^{-1}\,\text{diag}(\hat{D}_n)F(f)$$

where $F : C^{2\pi} \to l^\infty(\mathbb{Z})$. So there's an abrupt decay to 0 at $\pm(n+1)$. And maybe we want some kernel such that the decay is not abrupt, but we need the coefficients to decay linearly, which note that when $n \to 1$ everything's still 1. Thus

$$K_n * e^{ikx} \to e^{ikx}$$

and thus convolving with K_n is just the same as applying

$$\sigma_n(f) := \frac{1}{n}\left(S_0 + S_1 + \dots\right)$$

which as a sum converges in the Cesaro sense.

Where as an example, the partial sum $S = 1 - 1 + 1 - 1 + 1 - \dots$ converges to $\frac{1}{2}$ in the Cesaro sense.

14. 5/10: Fejer Kernel; Fourier Transform

As we've mentioned last time, we define the Fejer kernel to force the series converge (or that the Fourier series converges in Cesaro sense). Moreover, after computation we know

$$K_n = \frac{1}{n}\sum_{k=0}^{n-1} D_k = \frac{\sin^2(n(x-y)/2)}{n\sin^2((x-y)/2)}.$$

Lemma 14.1. *K_n is a non-negative, continuous, even, and*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} K_n(t)dt = \frac{1}{2\pi}||K_n|| = 1.$$

Theorem 14.2. *If $f \in C^{2\pi}$, then $\sigma_n f$ converges uniformly to f.*

Proof. We first note that K_n is an approximate identity, then we have

$$\begin{aligned}\left|f(x) - \frac{1}{2\pi}\int_{-\pi}^{\pi} K_n(t)f(x+t)dt\right| &= \left|\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(x)-f(x+t))K_n(t)dt\right| \\ &\le \frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)-f(x+t)|\,K_n(t)dt \to 0\end{aligned}$$

which we cannot do for D_n since the last step is not valid to get D_n out of abs value. □

14.1. Fourier Transform.

Def 14.3. *The transform of f is*

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\xi x} f(x)dx.$$

Theorem 14.4. *Suppose $f, \hat{f} \in L^1$, then*

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{i\xi x}\hat{f}(\xi)d\xi.$$

Proof.

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{i\xi x}\hat{f}(\xi)d\xi &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\xi(x-y)}f(y)dyd\xi \\ &= \int_{-\infty}^{\infty} f(y)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi(x-y)d\xi}\right]dy = f(x)\end{aligned}$$

since the thing inside the square bracket is just δ. □

Lemma 14.5. *If $f \in L^1$, then $\hat{f} \in C(\mathbb{R})$.*

Proof. This is just because

$$\hat{f}(\xi) - \hat{f}(\xi_0) = \int f(x) \left(e^{-i\xi x} - e^{-i\xi_0 x} \right) dx.$$

□

Lemma 14.6. *(Riemann-Lebesgue Lemma) If $f \in L^1$ then the transform is continuous and $|\hat{f}(\xi)| \to 0$ as $\xi \to \infty$.*

Proof. We can approximate by continuous compactly supported function in L^1. Then

$$\sqrt{2\pi}\hat{f}_\varepsilon(\xi) = \int e^{-i\xi x} f_\varepsilon(x) dx = \int e^{-i\xi x - i\pi} f_\varepsilon \left(x + \frac{\pi}{\xi} \right) dx$$

thus taking average

$$\sqrt{2\pi}\hat{f}_\varepsilon(\xi) = \frac{1}{2} \int e^{-i\xi x} \left(f_\varepsilon(x) - f_\varepsilon \left(x + \frac{\pi}{\xi} \right) \right) dx$$

then apply DCT and we are done. □

Proposition 14.7. *For $j \in \mathbb{Z}^+$, if f has j integrable derivatives, then $\exists C$ such that*

$$|\hat{f}(\xi)| \leq \frac{C}{\left(1 + |\xi|^2\right)^{j/2}}.$$

Moreover, for the converse we have that if $|\hat{f}(\xi)|$ decays like $|\xi|^{-(j+1+\varepsilon)}$ for some $\varepsilon > 0$ then f is continuous and has j continuous derivatives.

Proof. Part 1: Just take integral by parts to get $\frac{1}{|\xi|}$ one at a time.

Converse: We have

$$\frac{d}{dx} \frac{1}{\sqrt{2\pi}} \int e^{i\xi x} \hat{f}(\xi) d\xi = \frac{1}{\sqrt{2\pi}} \int (i\xi) e^{i\xi x} \hat{f}(\xi) d\xi$$

thus as long as the thing inside is well defined we are done. □

Now, as for f compactly supported, we know that $f\hat{f}$ is always finite, thus $\hat{f}(\xi)$ is analytic!

Now, everything above can be easily extended to L^2, the argument is in PDE's notes.

15. 5/15: Bandlimited function; Uncertainty Principle

15.1. **Bandlimited function.** First, we note that $\hat{\hat{f}} = f(-x)$.

Moreover, if f has compact support, then $\hat{f}$ is an entire function on C since we can just write out

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} f(x) e^{-i\xi x} dx.$$

Now we note that there's two kind of situations where the approximation with finite quadratures. First we might expect that there's some narrow bump that our wavelets does not capture; second, we might expect that the original frequency is very high that we've not chosen enough points to match the frequency. For the second question, we try to find a way to make our convergence fast enough, but will not be faster than the limit of frequency.

Def 15.1. *f is <u>bandlimited</u> if* $\operatorname{supp} \hat{f} \in [-\Omega, \Omega]$.

Also, in the later talks we can just assume that we are evaluating f at equispacesd times.

Theorem 15.2. *If $f \in L^1(\mathbb{R})$ and band limited, then*

$$f(t) = \sum_{n \in \mathbb{Z}} f\left(\frac{n\pi}{\Omega}\right) \operatorname{sinc}(\Omega t - n\pi).$$

Proof. First note that if f is band-limited, then we have

$$\hat{f}(\xi) = \begin{cases} \sum_{n \in \mathbb{Z}} \hat{f}(\xi + 2\Omega n) =: \phi(\xi) & |\xi| < \Omega \\ 0 & \text{else} \end{cases}$$

thus

$$\hat{f}(\xi) = \chi_{[-\Omega,\Omega]} \phi(\xi)$$

and because ϕ is a valid periodic function, we get our result. Note that χ is the transform of sinc so we are almost there.

To really get there, we Fourier expand ϕ to get

$$\phi = \sum_n c_n e^{-2\pi i \xi n / 2\Omega}$$

where

$$c_n = \frac{1}{2\Omega} \int_0^{2\Omega} e^{2\pi i \xi n / 2\Omega} \phi(\xi) d\xi = \frac{1}{2\Omega} \int_0^{2\Omega} e^{2\pi i \xi n / 2\Omega} \sum_n \hat{f}(\xi + 2\Omega n) d\xi$$

$$= \frac{1}{2\Omega} \int_{-\infty}^{\infty} e^{2\pi i \xi n / 2\Omega} \hat{f}(\xi) d\xi = \frac{\sqrt{2\pi}}{2\Omega} f\left(\frac{n\pi}{\Omega}\right)$$

and thus

$$\phi = \sum_n \hat{f}(\xi + 2\Omega n) = \frac{\sqrt{2\pi}}{2n} \sum_n f\left(\frac{n\pi}{\Omega}\right) e^{-2\pi i \xi n / 2\Omega}.$$

And now Plancherel gives us

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\xi i x} \hat{f}(\xi) d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} e^{i\xi x} \phi d\xi$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} e^{i\xi x} \frac{\sqrt{2\pi}}{2n} \sum_n f\left(\frac{n\pi}{\Omega}\right) e^{-2\pi i \xi n/2\Omega} d\xi$$

and we pull out the $\sum f$ and see that the others is sinc, sort of. □

In other words, what we did is just extending the compactly supported function to a periodic one, and by extending the period a little bit, we can diminish the low valued ones. But the decay is like sinc and that is roughly $\sum(-1)^n n^{-1}$ order convergence, which is as slow as we can get. This comes from the fact that we are transforming the indicator function. Thus if we smoothen the step function we can get better decay. For decay length $= \lambda$:

- If we only use linear decay, then the decay is $O\left(\frac{1}{x^2}\right)$ as $\lambda \to 0$. Or to be precise

$$f(x) = \sum_n f\left(\frac{n\pi}{\Omega(1+\lambda)}\right) G_\lambda\left(x - \frac{n\pi}{\Omega(1+\lambda)}\right)$$

  where

$$G_\lambda\left(x - \frac{n\pi}{\Omega(1+\lambda)}\right) = \frac{2\sin(x\Omega(1+\lambda/2)\sin(x\Omega\lambda/2))}{\lambda\Omega^2(1+\lambda)x^2}.$$

- In general, we cannot escape the bad term as $\lambda \to 0$ since that means we cannot escape a very very sharp decay.

Note that this λ is limited by the best sound that we can get from this.

Uncertainty principle

Assume that $f \in L^2$ and $||f||_2 = 1$. Let the positional energy be

$$(\Delta x)^2 = \int_{-\infty}^{\infty} x^2 |f|^2 dx - \left(\int_{-\infty}^{\infty} x|f|^2\right)^2$$

and the momentum is

$$(\Delta p)^2 = \int_{-\infty}^{\infty} \xi^2 |\hat{f}(\xi)|^2 d\xi - \left(\int_{-\infty}^{\infty} \xi|\hat{f}(\xi)|^2\right)^2$$

where we require $f'(x) = e^{-ip_0 x} f(x + x_0)$ then the energies are just the first term (last term gone). Then let $u = xf$, $v = \frac{d}{dx} f \Rightarrow \tilde{v} = i\xi\hat{f}$ and thus

$$\int |u|^2 dx = (\Delta x)^2; \int |v|^2 dx = (\Delta p)^2$$

and we have (a^* is $\bar{a}$, for convenience)

$$\begin{aligned}\text{Re}\int \bar{u}v dx &= \frac{1}{2}\int (xf^*)\frac{d}{dx}f + \frac{df^*}{dx}xf dx \\ &\overset{ibp}{=} \frac{1}{1}\int f^* x\frac{d}{dx}f - f^*\frac{d}{dx}(xf)dx = \frac{1}{2}f^*\left[x, \frac{d}{dx}\right]f \\ &= -\frac{1}{2}\int |f|^2 dx = -\frac{1}{2}\end{aligned}$$

after normalization. But then we have

$$\left|\left|\text{Re}\int \bar{u}v dx\right|\right|$$

which means

$$\frac{1}{2} \leq \Delta p \cdot \Delta x$$

and we might want to ask that can we find f compactly supported also compactly supported? The answer is no and the reason is that the transform is entire, so being compactly supported means bounded, thus 0 everywhere.

In other words, if we define the time limiting operator Q_T and band limiting operator P_Ω by

$$Q_T(f)(x) := \begin{cases} f(x) & |x| \leq T \\ 0 & \text{otherwise} \end{cases}; \qquad P_\Omega(\hat{f})(\xi) := \begin{cases} \hat{f}(\xi) & |\xi| \leq \Omega \\ 0 & \text{otherwise} \end{cases}$$

then compact support means for large enough time and frequency we have $Q_T f = f; P_\Omega f = f$ and thus

$$\frac{||P_\Omega Q_T f||^2}{||f||^2} = \frac{\langle P_\Omega Q_T P_\Omega^* f, f\rangle}{||f||^2}$$

and to compute the maximum eigenvalue we get the expression

$$P_\Omega Q_T P_\Omega^*(f)(x) = \frac{\Omega}{\pi}\int_{-T}^{T} \frac{\sin(\Omega(x-t))}{\Omega(x-t)} f(t)dt$$

which is not doable. But then Slepian find a mircle that $P_\Omega Q_T P_\Omega^*$ commutes with

$$(Af)(x) = \frac{d}{dx}(T^2 - x^2)\frac{df}{dx} - \frac{\Omega^2}{\pi^2}f$$

and with this method we've found that up till $\frac{2n\Omega}{\pi}$ we are doing really well to approximate 1, and then there's exponential decay to 0, followed by good approximation of 0. Those eigenvectors are called Prolate spheroidal wave functions.

Remark 15.3. *If they commute then they have the same eigen function is really what the miracle is up to.*

16. 5/17: Reproducing Kernel Spaces

Let the set of bandlimited function be

$$B_\Omega := \left\{ f \in L^2 \middle| \operatorname{supp} \hat{f} \in [-\Omega, \Omega] \right\}$$

then if $f \in B_\Omega$, we can extend f to an entire function on $\mathcal{C}$ which is of exponential type and

$$|f(z)| \leq \frac{1}{\sqrt{2\pi}} ||\hat{f}||_{L^1} e^{\Omega |\operatorname{Im} z|}$$

which means bandlimited functions form a Hilbert space of entire functions. So if $f \in B_\Omega$, $f \in L^1$, then

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} e^{i\xi x} \hat{f}(\xi) d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} e^{i\xi x} \int_{-\infty}^{\infty} e^{-i\xi y} f(y) dy d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) \frac{2\sin(\Omega(x-y))}{x-y} dy =: \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) e_x(y) dy = \langle e_x, f \rangle \end{aligned}$$

so in other words we've found out a function that can be used to evaluate f at a point. So roughly speaking this is a δ function. Note that $e_x = \operatorname{sinc}(\Omega(x-y))$ so it is in B_Ω, being the transform of an indicator function.

So if we set $K(x, y) = e_x(y)$ then

$$f(x) = \int f(y) K(x, y) dy$$

so a function can be evaluated by a function that is in the same space.

Def 16.1. *Hilbert spaces in which any function can be evaluated by some function that is in the same space are called reproducing kernel spaces.*

In other words, for $\mathcal{H}$ a Hilbert space of mappings from X to $\mathbb{R}$, then $\mathcal{H}$ is a reproducing kernel space if $\forall x \in X$, $\exists\, l_x \in \mathcal{H}^*$ with $||l_x|| < \infty$ such that $l_x(f) = f(x)$ for all $f \in \mathcal{H}$.

Example 16.2. *$L^2[-1, 1]$ is not a r.k.s.*

Example 16.3. *Hardy space $H^2(\mathbb{D})$*

Consider all functions that are analytic in the unit disc such that

$$\sup_{0<r<1} \left(\frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^2 d\theta \right)^{\frac{1}{2}} < \infty$$

where we define the inner product as usual:

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(re^{i\theta}) \bar{g}(re^{i\theta}) d\theta$$

then $e_w(z) = \dfrac{1}{1 - z\bar{w}}$ is the evaluation kernel, and the space is called Hardy space $H^2(\mathbb{D})$.

Example 16.4. *Sobolev Spaces.*

Let's take H_0^1 as an example. To be fair the definition is

$$H_0^1([0,1]) := \left\{ f \in L^2[0,1] \middle| f' \in L^1, f(0) = f(1) = 0 \right\}$$

and the induced norm is

$$\langle f, g \rangle_{H_0^1} = \int_0^1 fg + \int_0^1 f'g'.$$

Our goal is to find e_x such that $\langle f, e_x \rangle_{H_0^1} = f(x)$ (for this to make sense, assume dimension is high enough such that f is continuous), and we write out the left hand side:

$$\langle f, e_x \rangle_{H_0^1} = \int_0^1 e_x f \, dy + \int_0^1 f' e_x' \, dy = \int_0^1 f \cdot (e_x - \Delta e_x) dy$$

and this becomes the problem

$$\begin{cases} \Delta e_x - e_x = -\delta_x \\ e_x(0) = e_x(1) = 0 \end{cases}$$

where we get the solution as

$$e_x(y) = \begin{cases} \frac{\sinh(1-x)\sinh(y)}{\sinh(1)} & x > y \\ \frac{\sinh(1-y)\sinh(x)}{\sinh(1)} & y < x. \end{cases}$$

We now introduce some properties of the reproducing kernels.

Def 16.5. *A continuous $\phi : \mathbb{R}^d \to \mathbb{R}$ is positive semi-definite if for all natural numbers $n \in \mathbb{N}$ and distinct points $x_1, \dots, x_n \in \mathbb{R}^d$ and all $\alpha \in \mathbb{R}^n$ we have that the quadratic form is defined by*

$$\sum_{j=1}^n \sum_{k=1}^n \alpha_j \alpha_k \phi(x_j - x_k) \geq 0.$$

And we say it is positive definite if the sign is $>$ for all $\alpha \in \mathbb{R}^n$.

With the same idea,, for a given Ω, we can set

$$F_\phi = \text{span}\{\phi(\cdot, y) | y \in \Omega\}$$

where the bilinear form on F_ϕ is such that

$$\left(\sum_{j=1}^N \alpha_j \phi(\cdot, x_j), \sum_{k=1}^M \beta_k \phi(\cdot, y_k) \right)_\phi := \sum_{j=1}^N \sum_{k=1}^M \alpha_j \beta_k \phi(x_j, y_k)$$

then we have the below theorem.

Theorem 16.6. *(Moore-Aronszajn) If ϕ is symmetric positive definite on X, then there is a unique Hilbert space of functions on X for which ϕ is a representing kernel, i.e. for $f \in F_\phi$ we have*

$$(f, \phi(\cdot, y)) = \sum_{j=1}^{N} \alpha_j \phi(x_j, y) = f(y).$$

Conversely, we have that the reproducing kernel is semi positive definite.

We leave the proof as an exercise and move onto representing theorems.

Suppose that we try to minimize $f \in \mathcal{H}$ such that

$$\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 + \lambda J(f) \right\} \tag{16.1}$$

where $\mathcal{H}$ is a reproducing kernel space, and J is a square semi-norm penalty. Then suppose $\mathcal{H}_0 = \{f | J(f) = 0\}$ and set $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, then set $\phi_1, \ldots, \phi_m$ to be a finite dimensional basis for $\mathcal{H}_0$ and K be the reproducing kernel for $\mathcal{H}_1$.

Now, set $e_i = k(\cdot, x_i)$ where we might interpret y_i as data, where as $f(x_i)$ as sample points.

Theorem 16.7. *(Wabba) The solution to 16.1 is expressible as*

$$f_* = \sum_{j=1}^{m} \alpha_j \phi_j + \sum_{k=1}^{n} \beta_k e_k.$$

Remark 16.8. *The key point here is that we've reduced the infinite dimensional space to only those spanned by $k(\cdot, x_i)$, at the sampling locations.*

Proof. Let

$$f = \sum_{i=1}^{m} a_i \phi_i + \sum_{j=1}^{n} b_j e_j + \psi$$

where $\phi \perp e_j, \phi_i$, then we know

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} \left(y_i - \langle e_i, f \rangle \right)^2 = \sum_{i=1}^{n} \left(y_i - \sum a_j \phi_j(x_i) - \sum b_k e_k(x_i) \right)$$

then the expression to minimize turns out to be

$$F(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{i,j} c_i c_j K(x_i, x_j) + \lambda ||\psi||^2$$

and we see that the ψ does nothing except increase the value. Thus it would rather be 0. □

Now we move to another class of representations.

Suppose $\exists \psi \in L^2$ with

$$C_\psi = 2\pi \int |\xi|^{-1} |\hat{\psi}(\xi)|^2 d\xi < \infty$$

so that if $\psi \in L^1$, then we know that

$$\hat{\psi}(0) = 0 \Rightarrow \int \psi(x)dx = 0.$$

So we extend to

$$\psi^{a,b} = |a|^{-\frac{1}{2}}\psi\left(\frac{x-b}{a}\right)$$

and assume $||\psi||_L^2 = 1$.

Def 16.9. *The continuous wavelet transform is*

$$T^w f(a,b) = \langle f, \psi^{a,b}\rangle.$$

Note that this implies $|T^w f(a,b)| \leq ||f||$.

Proposition 16.10. *For $f, g \in L^2$, we have*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{dadb}{a^2} T^w f(a,b)\overline{T^w g}(a,b) = C_\psi \langle f, g\rangle$$

where we view C_ψ as a weight.

Moreover, if we set $f = g$, then

$$C_\psi^{-1}\int\int \frac{dadb}{a^2}|T^w f|^2 = ||f||^2$$

so that T^w is an isometry from $L^2(\mathbb{R})$ into $L^2\left(\mathbb{R}^2, C_\psi^{-1}\frac{dadb}{a^2}\right)$ where the second term is the weight of the inner product.

Now we further define the norm on the above Hilbert space $\mathcal{H} := L^2\left(\mathbb{R}^2, C_\psi^{-1}\frac{dadb}{a^2}\right)$ to be $|||\cdot|||$, then the image of L^2 is a closed subspace of $\mathcal{H}$.

Then for $F \in \mathcal{H}$ there exists $T^w f = F$ where

$$\begin{aligned} F(a,b) = \langle f, \psi^{a,b}\rangle &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{da'db'}{(a')^2} T^w f(a',b')\overline{T^w \psi^{a,b}}(a',b') \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{da'db'}{(a')^2} F(a',b')K(a,b,a',b') \end{aligned}$$

where $K(a,b,a',b') := \langle \psi^{a',b'}, \psi^{a,b}\rangle$ thus $\mathcal{H}$ is a representing kernel Hilbert space. This helps us get around uncertainty principle.

APPENDIX A. A

APPENDIX B. B

APPENDIX C. C

Acknowledgements.