Approximation Theory and Spectral Methods

Thomas Trogdon
University of Washington
trogdon@uw.edu

Contents

Preface

First and foremost: Very much a work in progress and is no way a replacement for the references listed below. The text focuses on approximation theory for functions of a single variable. Both theoretical and practical aspects are treated. Supplementary references are listed below.

Elementary analysis:

- W Rudin. *Principles of mathematical analysis.* McGraw-Hill, Inc., New York, NY, 3rd edition, 1964

Advanced analysis:

- G B Folland. *Real analysis.* John Wiley and Sons Inc., New York, 1999

Complex analysis:

- M J Ablowitz and A S Fokas. *Complex Variables: Introduction and Applications.* Cambridge University Press, second edition, 2003

Approximation theory and spectral methods:

- L N Trefethen. *Approximation Theory and Approximation Practice, Extended Edition.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1 2019
- L N Trefethen. *Spectral methods in MATLAB.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000
- W Cheney and W Light. *A Course in Approximation Theory*, volume 101 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, Rhode Island, 1 2009
- K Atkinson and W Han. *Theoretical Numerical Analysis.* Springer, New York, NY, 2009
- N I Achieser. *Theory of Approximation.* Dover Publications, 1992
- T J Rivlin. *An Introduction to the Approximation of Functions.* Dover Publications, 1969

Integral equations:

- R Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences.* Springer New York, New York, NY, 2014

Second: I hope you will find this subject truly remarkable.

Notation:

- Subsets of $\mathbb{C}$ or $\mathbb{R}^d$ and denoted by upper-case Greek letters, e.g., Ω, Γ, Σ.

- Function spaces will be denoted by upper-case Roman characters, e.g., $C(\Omega)$, with possible sub/superscripts.

- No distinction will be made in notation for independent variables being scalars or vectors, e.g., $x \in \mathbb{R}$, $y \in \mathbb{R}^3$, etc.

- Operators will be denoted by calligraphic fonts ($\mathcal{K}$) and their discrete approximations by bold characters ($\mathbf{K}$).

- Vectors of expansion coefficients, or function values, will be noted using bold lower-case characters: $\mathbf{u}, \mathbf{v}$, etc. Bold lower-case characters will also be used when speaking abstractly about vectors.

- Different geometric symbols ($\diamond, \triangleleft, \triangleright$) will be used to define, what, in programming terms, we would call anonymous functions, e.g., the function $x \mapsto x^2$ can be defined by $\diamond^2$ or the function $(x, y) \mapsto x^2 + y^2$ can be defined by $\triangleleft^2 + \triangleright^2$ (Note that in this latter case the ordering of the input variables can be ambiguous). The reasoning behind this is if one wants to speak about norm of the function $xf(x)$ it is not mathematically correct to write $\|xf\|$, rather one needs to write $\| \diamond f(\diamond)\|$ or $\| \diamond \cdot f\|$. Some authors use $\cdot$ for this purpose, but as we see from this example, we want to retain it to emphasize multiplication.

- For sets S, $\overline{S}$ denotes the closure (with respect to a clear-from-context topology) and for complex numbers z, $\overline{z}$ denotes the complex conjugate.

- We use $\mathbf{e}_j$ to denote the standard basis vectors with dimension inferred from the context.

What is approximation theory?

Approximation theory is the study of how functions from a given class can be approximated by simpler functions. One may look to establish bounds on the optimal approximation or construct algorithms (hopefully with bounds) to give practical approximations. The first place one sees this, typically, is Fourier series — the approximation of functions by complex exponentials, or equivalently, by trigonometric functions.

In this text, we derive most of our basis functions and approximation schemes from Fourier analysis: Chebyshev polynomials are mapped cosine functions and the rational bases we consider are derived from Möbius transformations of the unit circle. The basis functions that do not come directly from Fourier theory, come from the general construction of orthogonal polynomials.

While we do give some theoretical estimates on best approximation (see Jackson's first-fourth theorems), most approximations we discuss are constructive. Even our proof of Jackson's first theorem is constructive. So, there is always a practical bent to the presentation, but maybe not as much as other authors.

What are spectral methods?

One can understand the historical appearance of "true" spectral methods using symmetric matrices. Let $\mathbf{A}$ be an invertible $N \times N$ real, symmetric matrix and consider solving $\mathbf{A}\mathbf{x} = \mathbf{b}$. If you use the "spectral" decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where

$$\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N\end{bmatrix}, \quad \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N).$$

Then the linear system is solved via

$$\mathbf{x} = \sum_{j=1}^{N} \lambda_j^{-1} \langle \mathbf{b}, \mathbf{u}_j \rangle \mathbf{u}_j.$$

This may appear trivial, but it extends to self-adjoint differential operators provided inner products and (orthonormal) eigenfunctions can be approximated sufficiently well, either numerically or analytically:

$$\mathcal{L}_0 u = f, \quad \mathcal{L}_0 u_j = \lambda_j u_j, \quad u = \sum_{j=1}^{\infty} \lambda_j^{-1} \langle f, u_j \rangle u_j.$$

But the restriction that one can handle approximations of the eigenfunctions, etc., is a large restriction indeed. So, in solving a perturbed problem $\mathcal{L} = \mathcal{L}_0 + \mathcal{E}$ one might try

$$\mathcal{L} u = f, \quad u = \sum_{j=1}^{\infty} a_j u_j.$$

Then the coefficients $\mathbf{a} = (a_j)_{j \geq 1}$ satisfy the infinite-dimensional linear system $\mathbf{L}\mathbf{a} = \mathbf{f}$ where

$$\mathbf{L} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \lambda_3 & \\ & & & \ddots \end{bmatrix} + \begin{bmatrix} \langle \mathcal{E}u_1, u_1 \rangle & \langle \mathcal{E}u_2, u_1 \rangle & \langle \mathcal{E}u_3, u_1 \rangle & \cdots \\ \langle \mathcal{E}u_1, u_2 \rangle & \langle \mathcal{E}u_2, u_2 \rangle & \langle \mathcal{E}u_3, u_2 \rangle & \cdots \\ \langle \mathcal{E}u_1, u_3 \rangle & \langle \mathcal{E}u_2, u_3 \rangle & \langle \mathcal{E}u_3, u_3 \rangle & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \langle f, u_1 \rangle \\ \langle f, u_2 \rangle \\ \langle f, u_3 \rangle \\ \vdots \end{bmatrix}. \tag{0.1}$$

The big leap of spectral methods is to use the eigenfunctions u_j in an entirely different context. For example, the Chebyshev polynomials of the first kind are eigenfunctions of a second-order differential operator. But one may use these polynomials to solve unrelated integral equations.

The issue with (0.1) is that (1) the inner products may be difficult to approximate, (2) one needs to compute/approximate $\mathcal{E}u_j$ and (3) this infinite-dimensional system has to be approximated in some manner. While (3) is unavoidable, (1) and (2) present serious challenges. This is where pseudospectral methods come into play via interpolation operators that in many cases allow one to compute approximations of the action of $\mathcal{E}$ in a simpler way. And pseudospectral methods typically replace the inner products (also referred to as the Galerkin projections) with the condition that the operator equation should hold on a grid. This is known as collocation.

Part I

Function spaces and approximation theory

Chapter 1

Banach and Hilbert spaces

From an approximation-theoretic perspective there are key ideas from analysis:

- compactness,
- completeness, and
- separability.

We begin with a review to show how these ideas arise and are used.

1.1 ▪ Metric spaces

We begin with some important concepts from elementary analysis.

Definition 1.1. *A metric space* (X, d) *satisfies:*

- $d : X \times X \to [0, \infty)$,
- *for* $x, y \in X$, $d(x, y) = 0$ *if and only if* $x = y$,
- *for* $x, y \in X$, $d(x, y) = d(y, x)$, *and*
- *for* $x, y, z \in X$, $d(x, y) \leq d(x, z) + d(z, y)$.

Notation 1.2. *Suppose* (X, d) *is a metric space. Then for* $x \in X$, $\epsilon > 0$,

$$B_\epsilon(x) := \{y \in X \mid d(y, x) < \epsilon\}.$$

Something crucial to numerical, or more generally, finite-dimensional, approximation is the notion of compactness.

Definition 1.3. *A subset* $Y \subset X$ *of a metric space* (X, d) *is said to be compact if* *whenever $\{U_\alpha\}_{\alpha\in A}$ is a collection of open subsets of X such that*

$$Y \subset \bigcup_{a\in A} U_\alpha,$$

then there exists a finite subcollection $\{U_{\alpha_1},\dots U_{\alpha_m}\}$ such that

$$Y \subset \bigcup_{j=1}^{m} U_{\alpha_j}.$$

In other words, a set is compact if and only if every open cover has a finite subcover.

Definition 1.4. *A subset $Y \subset X$ of a metric space (X,d) is said to be sequentially compact if every sequence $(x_j)_{j\geq 1}$, $x_j \in Y$ has a convergent subsequence:*

$$\lim_{k\to\infty} x_{j_k} = x^* \in Y,$$

where $j_k < j_{k+1}$, $\{j_1, j_2, \dots\} \subset \{1, 2, \dots\}$.

Theorem 1.5. *A subset $Y \subset X$ of a metric space (X,d) is compact if and only if it is sequentially compact.*

Proof. Suppose Y is compact. The intuition is that in a compact space, a sequence has nowhere to "escape" to so that it does not have a convergence subsequence.

Let $(x_j)_{j\geq 1}$ be a sequence in Y. We need to show it has a convergent subsequence. If an element $y \in Y$ is repeated an infinite number of times in the sequence, we are done. So, suppose this is not the case. In this setting, the statement of their being a convergent subsequence is equivalent to the sequence having an accumulation point in Y: y is an accumulation point of the sequence if for every $\epsilon > 0$, $B_\epsilon(y)$ contains an element of the sequence (Why is this equivalent?). By contradiction, suppose there is no limit point in Y. Then for every $y \in Y$, there exists $\epsilon_y > 0$ such that $B_{\epsilon_y}(y)$ contains only a finite number of elements in the sequence. Since $Y = \bigcup_{y\in Y} B_{\epsilon_y}(y)$, by compactness, there exists a finite subcover

$$Y = \bigcup_{j=1}^{N} B_{\epsilon_{y_j}}(y_j).$$

This then implies that the sequence has a finite number of distinct elements — at least one must be repeated an infinite number of times, a contradiction. Thus Y is sequentially compact.

Now, suppose Y is sequentially compact. And suppose there exists an open cover of Y that has no finite subcover. The first observation is that sequential compactness implies that given any fixed radius, a finite number of balls will cover Y. Let $\epsilon > 0$ and select $y_1 \in Y$. If $Y \subset B_\epsilon(y)$ then we are done. Inductively,

- if $Y \subset Y_n := \bigcup_{j=1}^{n} B_\epsilon(y_j)$ stop, otherwise

- select $y_{n+1} \in Y \setminus Y_n$.

Claim: There exists $N > 0$ such that $Y = Y_N$. Suppose the sequence is infinite. By sequential compactness, the sequence $(y_j)_{j\geq 1}$ contains a convergent subsequence $(y_{j_k})_{k\geq 1}$. And thus there exists K such that, for $k, \ell > K$, $d(y_{j_k}, y_{j_\ell}) < \epsilon$. But this is in contradiction with the construction, implying the sequence is finite.

Our next task is to show that for any open cover $\{U_\alpha\}_{\alpha\in A}$, there exists $\epsilon > 0$ such that for any $y \in Y$, $B_\epsilon(y) \subset U_\alpha$ for some α. Suppose not. Let take $\epsilon = 1/n$, $n = 1, 2, \ldots$ and for each choice of ϵ there must exist y_n such that $B_{1/n}(y_n)$ is not contained in U_α for all α. Now this sequence, has a convergent subsequence $(y_{n_k})_{k\geq 1}$ that converges to $y \in U_\beta$, for some $\beta \in A$. As U_β is open, $B_{1/n_k}(y_{n_k}) \subset U_\beta$, for k sufficiently large. This gives a contradiction.

We can now complete the proof. Suppose $\{U_\alpha\}_{\alpha\in A}$ is an open cover of Y. Select $\epsilon > 0$ such for every $y \in Y$, $B_\epsilon(y) \subset U_\alpha$ for some α. Since Y can be covered by a finite number of these

$$Y \subset \bigcup_{j=1}^{N} B_\epsilon(y_j),$$

we select $\alpha_j \in A$, to be such that $B_\epsilon(y_j) \subset U_{\alpha_j}$. Then

$$Y \subset \bigcup_{j=1}^{N} U_{\alpha_j},$$

and Y is compact. ■

Theorem 1.6 (Heine–Borel). *A set $\Omega \subset \mathbb{R}^d$, with the standard Euclidean metric, compact if and only if it is closed and bounded.*

1.2 ▪ Completeness and vector spaces

While not entirely accurate, most of the analysis of finite-difference methods can take place in one function space: The space of continuous functions with the max norm.

Definition 1.7. *Given a compact (i.e., closed and bounded) set $\Omega \subset \mathbb{R}^d$, define*

$$C(\Omega) := \{f : \Omega \to \mathbb{C} \mid f \text{ is continuous on } \Omega\}.$$

Recall that the continuous image of a compact set is compact: For $f \in C(\Omega)$, $f(\Omega)$ is compact (and hence closed and bounded). Thus, f is bounded, and the norm

$$\|f\|_\infty := \sup_{x\in\Omega} |f(x)|,$$

is finite. It is then natural to ask if this supremum can be replaced with a maximum. By the definition of supremum, for every $\epsilon > 0$ there exists x_ϵ such that

$$|f(x_\epsilon)| + \epsilon > \|f\|_\infty.$$

Then take $\epsilon = 1, 1/2, 1/3, \ldots$ and the corresponding sequence of x_ϵ's must have a subsequence that converges (because Ω is compact) to $x^* \in \Omega$. By the continuity of f we find

$$|f(x^*)| \geq \|f\|_\infty \Longrightarrow \max_{x \in \Omega} |f(x)| = \|f\|_\infty.$$

With this norm $C(\Omega)$ is a normed vector space. If Ω was not compact, then we would need to restrict to functions for which the norm is finite (and likely cannot be replaced with the max norm).

Definition 1.8. *Given $\Omega \subset \mathbb{R}^d$, define*

$$C(\Omega) := \{f : \Omega \to \mathbb{C} \mid f \text{ is continuous on } \Omega, \quad \|f\|_\infty < \infty\}.$$

Remark 1.9. *Here Ω need not be a subset of $\mathbb{R}^d$. It could be a subset of a different metric space altogether.*

Definition 1.10. *A vector space V over a field $\mathbb{F}$ ($\mathbb{F} = \mathbb{C}, \mathbb{R}$) must satisfy the following for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $a, b \in \mathbb{F}$:*

- $a\mathbf{u} + b\mathbf{v} \in V$,
- $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$,
- $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$,
- *there exists* $\mathbf{0} \in V$ *such that* $\mathbf{u} + \mathbf{0} = \mathbf{u}$,
- *there exists* $-\mathbf{u}$ *such that* $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$,
- $a(b\mathbf{u}) = (ab)\mathbf{u}$,
- $1\mathbf{u} = \mathbf{u}$,
- $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$, *and*
- $(a + b)\mathbf{u} = a\mathbf{u} + b\mathbf{u}$.

Definition 1.11. *A normed vector space $(V, \|\diamond\|)$ over $\mathbb{F}$ is such that V is a vector space over $\mathbb{F}$ and the norm satisfies:*

- $\|\mathbf{u}\| \geq 0$,
- $\|\mathbf{u}\| = 0$ *if and only if* $\mathbf{v} = \mathbf{0}$.
- $\|a\mathbf{u}\| = |a|\|\mathbf{u}\|$, *and*
- $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

Any normed vector space is a metric space with metric $d(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|$.

It is immediate that $(C(\Omega), \| \diamond \|_\infty)$ is a normed vector space over $\mathbb{C}$.

Remark 1.12. *Consider a space spanned by an infinite number of linearly independent vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots\}$. Then consider the sequence*

$$\mathbf{x}_j = \mathbf{u}_j, \quad j = 1, 2, \ldots.$$

For this sequence to converge, there will need to be an asymptotic degeneracy in the spanning vectors.

Definition 1.13. *A sequence $(x_j)_{j\geq 1}$ in a metric space (X, d) is said to be a Cauchy sequence if for every $\epsilon > 0$ there exists $N > 0$ such that for $j, k > N$*

$$d(x_j, x_k) < \epsilon.$$

Exercise 1.1. *Show that if a subsequence of a Cauchy sequence converges, then the original sequence converges.*

Definition 1.14. *A metric space (X, d) is said to be complete if every Cauchy sequence in X converges (in X).*

Theorem 1.15. $\mathbb{R}^d$ *with the standard Euclidean metric is complete.*

Proof. Let $(x_j)_{j\geq 1}$ be a Cauchy sequence in $\mathbb{R}^d$. First, we claim that this sequence is bounded[1]. Set $\epsilon = 1$. Then by the definition of a Cauchy sequence there exists N such that

$$d(x_{N+1}, x_j) < 1, \quad j \geq N+1.$$

Since $\{x_1, \ldots, x_{N+1}\}$ is bounded, the sequence must be. We find that there exists $R > 0$ such that $\{x_1, x_2, \ldots\} \subset \overline{B_R(0)}$. And thus a subsequence must converge in $\mathbb{R}^d$. It follows that if a Cauchy sequence has a convergent subsequence, then the sequence itself must converge. ■

Definition 1.16. *A complete normed vector space is called a Banach space.*

[1] What does boundedness mean in a general metric space?

Theorem 1.17. *If Ω is a subset of a metric space (X, d), then $C(\Omega)$ is a Banach space.*

Proof. We must show that every Cauchy sequence converges. So, let $(f_j)_{j\geq 1}$ be a Cauchy sequence in $C(\Omega)$. Then, in particular, for each $x \in \Omega$, $(f_j(x))_{j\geq 1}$ is a Cauchy sequence in $\mathbb{C}$, which must converge. So, define $f : \Omega \to \mathbb{C}$ via

$$f(x) = \lim_{j\to\infty} f_j(x).$$

It remains to do two things: (1) show f is continuous and (2) show that $\|f - f_j\|_\infty \to 0$ as $j \to \infty$.

(2): Using the triangle inequality

$$|f(x) - f_j(x)| \leq |f(x) - f_k(x)| + |f_k(x) - f_j(x)|.$$

Because the sequence is Cauchy, there exists $N > 0$ such that if $k, j > N$, $|f_k(x) - f_j(x)| < \epsilon$ for all x. Thus

$$|f(x) - f_j(x)| \leq \epsilon + |f(x) - f_k(x)|.$$

Now, the left-hand side has no dependence on k, and we can take the limit as $k \to \infty$:

$$|f(x) - f_j(x)| \leq \epsilon, \quad j > N.$$

This establishes (2).

(1): Let $\epsilon > 0$. For each $x \in \Omega$ must show that there exists $\delta > 0$ such that if $d(x, y) < \delta$ then $|f(x) - f(y)| < \epsilon$. The technique is similar:

$$|f(x) - f(y)| \leq |f(x) - f_k(x)| + |f_k(x) - f_k(y)| + |f(y) - f_k(y)|.$$

Now, let k be large enough so that $\|f - f_k\|_\infty < \epsilon/3$. And choose δ so that $|f_k(x) - f_k(y)| < \epsilon/3$. This shows that f is continuous. ■

Exercise 1.2. *For $C([0, 1])$ show that $\overline{B_1(0)}$ is not compact.*

Definition 1.18. *A metric space (X, d) is said to be separable if there exists a countable dense subset.*

Now, we immediately see an issue with $C(\mathbb{R})$

Proposition 1.19. *$C(\mathbb{R})$ is not separable.*

Proof. Consider just continuous functions that are equal to zero or one at the positive integers (take these functions to be zero for $x \leq 0$). Recall that the set of all infinite binary sequences is uncountable. So for a binary sequence $a = a_1 a_2 a_3 \ldots$, define a function

$$f_a(x) = \sum_j a_j h(x - j), \quad h(x) = \begin{cases} x + 1 & -1 \leq x \leq 0, \\ 1 - x & 0 < x \leq 1, \\ 0 & \text{otherwise}. \end{cases}$$

For $a \neq a'$, we have $\|f_a - f'_a\|_\infty \geq 1/2$. This implies that $C(\mathbb{R})$ contains a uncountable number of disjoint open sets indexed by binary sequences $U_a = B_{1/4}(f_a)$ and it cannot be separable. ■

But this issue does not arise when the domain is compact.

Theorem 1.20 (Weierstrass). *Polynomials are dense in $C([0,1])$ and hence $C([0,1])$ is separable.*

Proof. We give a proof that is attributed to Bernstein and it uses probabilistic ideas. For $f \in C([0,1])$ define the polynomial

$$B_n(x) = B_n(x; f) = \sum_{k=0}^{n} f(x_k)\binom{n}{k} x^k (1-x)^{n-k}, \quad x_k = k/n.$$

The claim is that $\|f - B_n\|_\infty \to 0$ as $n \to \infty$. What we want to show is that as n becomes large, the contributions to $B_n(x)$ comes from only the x_k that are near x (the binomial distribution concentrates on its mean) and then the continuity of f is used locally. To this end, consider $\{p_0, \dots, p_n\}$ such that $\sum_k p_k = 1$. And define a mean and variance

$$\mu = \sum_k x_k p_k, \quad \sigma^2 = \sum_k (x_k - \mu)^2 p_k.$$

Then for $s > 0$

$$\sum_{k: |x_k - \mu|/(s\sigma) \geq 1} p_k = \sum_{k: (x_k - \mu)^2/(s\sigma)^2 \geq 1} p_k \leq \sum_k p_k \frac{(x_k - \mu)^2}{s^2 \sigma^2} = \frac{1}{s^2}.$$

We claim that

$$\begin{aligned} 1 &= \sum_k \binom{n}{k} x^k (1-x)^{n-k}, \\ x &= \sum_k x_k \binom{n}{k} x^k (1-x)^{n-k}, \\ \frac{x(1-x)}{n} &= \sum_k (x_k - \mu)^2 \binom{n}{k} x^k (1-x)^{n-k}. \end{aligned} \tag{1.1}$$

Plugging this all in, we have

$$\sum_{k: |x_k - x| \geq s\sqrt{\frac{x(1-x)}{n}}} \binom{n}{k} x^k (1-x)^{n-k} \leq \frac{1}{s^2}, \quad s > 0.$$

Returning to what we aim to establish:

$$\begin{aligned}|B_n(x) - f(x)| &\leq \sum_k |f(x_k) - f(x)| \binom{n}{k} x^k (1-x)^{n-k} \\ &= \sum_{k:|x_k - x| \geq s\sqrt{\frac{x(1-x)}{n}}} |f(x_k) - f(x)| \binom{n}{k} x^k (1-x)^{n-k} \\ &+ \sum_{k:|x_k - x| < s\sqrt{\frac{x(1-x)}{n}}} |f(x_k) - f(x)| \binom{n}{k} x^k (1-x)^{n-k}.\end{aligned}$$

Now, choose $s = n^{1/4}$ for example. For every $\epsilon > 0$, there exists δ such that $|f(x) - f(y)| < \epsilon/2$ whenever $|x - y| < \delta$ (why?). So, if n is large enough such that $s\sqrt{\frac{x(1-x)}{n}} < \delta$ and $2\|f\|_\infty n^{-1/2} < \epsilon/2$, then

$$|B_n(x) - f(x)| \leq \epsilon/2 + 2\|f\|_\infty n^{-1/2} < \epsilon.$$

And since any polynomial can be approximated uniformly by polynomials with rational coefficients, separability follows.

■

Exercise 1.3. *Establish* (1.1)*.*

While this is encouraging, we will see that $C([0,1])$ is not restrictive enough for practical approximations as this constructive approximation above is not usable, in general, in finite-precision arithmetic as it can converge arbitrarily slowly.

We end this section with a theorem that is of great use in what follows. A series of elements $(\mathbf{v}_k)_{k\geq 1}$ in a vector space is said to be absolutely convergent if

$$\sum_{k=1}^{\infty} \|\mathbf{v}_k\| < \infty.$$

Theorem 1.21. *A normed vector space is a Banach space if and only if every absolutely convergent series converges.*

Proof. First, suppose the space is a Banach space — it is complete. It follows, by repeated use of the triangle inequality, that the sequence $\left(\sum_{k=1}^j \mathbf{v}_k\right)_{j\geq 1}$ is Cauchy and therefore converges.

Conversely, suppose that every absolutely convergent series converges. And suppose that $(\mathbf{u}_j)_{j\geq 1}$ is a Cauchy sequence. For $\epsilon = 2^{-k}$, $k = 1, 2, \ldots$, there exists $j_k > j_{k-1}$ such that if $\ell, j \geq j_k$, $\|\mathbf{u}_j - \mathbf{u}_\ell\| < 2^{-k}$. So, define $\mathbf{v}_k = \mathbf{u}_{j_{k+1}} - \mathbf{u}_{j_k}$. It follows that the sequence $\left(\sum_{k=1}^j \mathbf{v}_k\right)_{j\geq 1}$ converges and

$$\sum_{k=1}^{j} \mathbf{v}_k = \mathbf{u}_{j_{k+1}} - \mathbf{u}_{j_1}.$$

This shows a subsequence of the Cauchy sequence converges, and therefore the sequence itself must converge. ■

1.3 ▪ L^p spaces

We present this section with minimal discussion of measure theory. A full discussion of it is required to establish all the claims in this section. In the end we will largely consider measures on $\mathbb{R}$ that have smooth (away from endpoints) densities.

Definition 1.22. *The Borel σ-algebra (on $\mathbb{R}$) is the smallest σ-algebra that contains the open sets. A Borel measure μ is a measure on this σ-algebra. Such a measure, if it is σ-finite, is uniquely determined by its action on intervals (Carathéodory's extension theorem).*

Definition 1.23. *A function $\rho : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}$ is said to be of bounded variation if for every interval $[a, b] \subset \Omega$ there exists a constant $M(a, b)$ such that for any $a \leq x_0 < x_1 < \cdots < x_n \leq b$*

$$\sum_{j=1}^{n} |\rho(x_j) - \rho(x_{j-1})| \leq M.$$

We can now define the Lebesgue–Stieltjes integral.

Definition 1.24. *Suppose $\rho : [a, b] \to \mathbb{R}$ is a weakly increasing, right-continuous function of bounded variation. The Lebesgue–Stieltjes measure μ_ρ is defined by $\mu_\rho((c, d]) = \rho(c) - \rho(d)$. The Lebesgue–Stieltjes integral with respect to ρ is*

$$\int f(x)\mu_\rho(\mathrm{d}x) = \int f(x)\mathrm{d}\rho(x),$$

for all integrable (and measureable) f.

All measures we will encounter will arise in this fashion even though much of the discussion could be more general. Unlike, the standard Lebesgue integral, if ρ has jump discontinuities (corresponding to point masses in μ_ρ), whether or not the interval of integration has open/closed endpoints is crucial.

Proposition 1.25. *If ρ is absolutely continuous, i.e., $\rho(x) = \int_a^x \rho'(x)\mathrm{d}x$ where ρ' is integrable, then*

$$\int f(x)\mu_\rho(\mathrm{d}x) = \int f(x)\rho'(x)\mathrm{d}x.$$

One of the most useful results from measure-theoretic integration is the following.

Theorem 1.26 (Generalized dominated convergence theorem). *Suppose $f_n, g_n, f, g \in L^1(\mu)$. Suppose as $n \to \infty$, $f_n(x) \to f(x)$, $g_n(x) \to g(x)$ for almost every x (i.e., the set on which there is no convergence is measure zero). Suppose further that $|f_n(x)| \leq g_n(x)$ and*

$$\int g_n(x)\mu(\mathrm{d}x) \to \int g(x)\mu(\mathrm{d}x), \quad n \to \infty.$$

Then

$$\int f_n(x)\mu(\mathrm{d}x) \to \int f(x)\mu(\mathrm{d}x), \quad n \to \infty.$$

The funny, but in the end convenient, thing about such integrals is if $\mu_\rho(\{x : f(x) \neq g(x)\}) = 0$ then $\int f(x)\mu_\rho(\mathrm{d}x) = \int g(x)\mu_\rho(\mathrm{d}x)$. So, really, any norm on these functions that is based on integrating them will see them as the same function. So, we introduce an equivalence relation $\sim$:

$$f \sim g \quad \text{if and only if} \quad \mu_\rho(\{x : f(x) \neq g(x)\}) = 0.$$

We now define the Lebesgue $L^p(\mu)$ spaces.

Theorem-Definition 1.27. *Given a measure μ on $\Omega \subset \mathbb{R}$ (with its associated σ-algebra), for $1 \leq p < \infty$, $L^p(\mu)$ is the normed vector space of equivalence classes of functions $[f]$ (under relation $\sim$) such that*

$$\|[f]\|_p := \left(\int_\Omega |f(x)|^p \mu(\mathrm{d}x) \right)^{1/p} < \infty, \qquad \textit{for any } f \in [f].$$

If $p = \infty$,

$$\|[f]\|_\infty := \inf_{f \in [f]} \|f\|_\infty$$

In this, we often eliminate the brackets $[\cdot]$ and just write $\|f\|_p$ as it rarely causes confusion. But the important point is that if $f_n \to f$ then the limit is only guaranteed to be defined up to a set of measure zero.

The proof of this is essentially done once one establishes Minkowski's inequality below.

One of the most important inequalities concerning $L^p(\mu)$ spaces is the Hölder inequality.

Theorem 1.28 (Hölder inequality). *For $1 \leq p, q \leq \infty$ such that $1 = 1/p + 1/q$, and $f \in L^p(\mu)$, $g \in L^q(\mu)$ then $fg \in L^1(\mu)$ and*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Proof. A proof is based on the elementary inequality: For $a, b \geq 0$, and $0 < \lambda < 1$

$$a^\lambda b^{1-\lambda} \leq \lambda a + (1 - \lambda)b,$$

with equality occuring if and only if $a = b$. To see this, the inequality holds if either a or b is zero. So, suppose that neither vanish and set $\delta = a/b$ to rewrite the inequality.

$$\delta^\lambda \leq \lambda\delta + (1-\lambda), \quad \delta \in (0,\infty). \tag{1.2}$$

We look for critical points of $r(\delta) = \lambda\delta + (1-\lambda) - \delta^\lambda$, $\delta > 0$. We find

$$r'(\delta) = \lambda - \lambda\delta^{\lambda-1}.$$

This vanishes only when $\delta = 1$. Checking the second derivative shows that this is the global minimum:

$$r(\delta) \geq r(1) = 0 \Rightarrow (1.2).$$

Again, if either $\|f\|_p$ or $\|g\|_q$ are zero, the inequality follows trivially. So, suppose they are nonzero and set Then we set $a = (|f(x)|/\|f\|_p)^p$, $b = (|g(x)|/\|g\|_q)^q$, $\lambda = p^{-1}$ to find

$$\frac{|f(x)g(x)|}{\|f\|_p\|g\|_q} \leq p^{-1}(|f(x)|/\|f\|_p)^p + q^{-1}(|g(x)|/\|g\|_q)^q.$$

Upon integration of this expression, the claim follows for $1 < p < \infty$. If $p = 1, q = \infty$, the claim is straightforward to verify.

■

Theorem 1.29 (Minkowski's inequality). *If $1 \leq p \leq \infty$, and $f, g \in L^p(\mu)$, then*

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p.$$

Proof. The case of $p = 1, \infty$ is immediate. So, assume $1 < p < \infty$ and the trick is to use that

$$\begin{aligned}|f(x)+g(x)|^p &\leq (|f(x)| + |g(x)|)|f(x)+g(x)|^{p-1}\\ &\leq |f(x)||f(x)+g(x)|^{p-1} + |g(x)||f(x)+g(x)|^{p-1}.\end{aligned}$$

We then just apply Holder's inequality to each term on the right-hand side, noting that $(p-1)q = p$ when $1/p + 1/q = 1$. This gives

$$\begin{aligned}\int_\Omega |f(x)+g(x)|^p \mu(\mathrm{d}x) &\leq \|f\|_p \||f+g|^{(p-1)q}\|_q + \|g\|_p \||f+g|^{(p-1)q}\|_q\\ &= (\|f\|_p + \|g\|_p)\|f+g\|_p^{p/q}.\end{aligned}$$

This implies that

$$\|f+g\|_p \leq (\|f\|_p + \|g\|_p)^{1/p}\|f+g\|_p^{1/q}.$$

Upon rearranging this

$$\|f+g\|_p^{1-1/q} \leq (\|f\|_p + \|g\|_p)^{1/p},$$

which turs out to be what we need to establish. ■

What we will not prove is the following:

Theorem 1.30. *For $1 \leq p \leq \infty$, $L^p(\mu)$ is a Banach space.*

1.4 ▪ Hilbert spaces

The space $L^2(\mu)$ is special because when $p = q = 2$, $1/p + 1/q = 1$. And for this reason, it becomes a Hilbert space.

Definition 1.31. *A Banach space $(V, \| \diamond \|)$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ is a Hilbert space if the there exists an inner product $\langle \triangleright, \triangleleft \rangle$ that satisfies the following for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, $a, b \in \mathbb{F}$,*

- $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$,
- $\langle a\mathbf{u} + b\mathbf{v}, \mathbf{w} \rangle = a\langle \mathbf{u}, \mathbf{w} \rangle + b\langle \mathbf{v}, \mathbf{w} \rangle$, *and*
- *if* $\mathbb{F} = \mathbb{C}$, $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$.

All inner products satisfy the Cauchy-Schwarz inequality.

Theorem 1.32 (Cauchy-Schwarz inequality). *For all vectors $\mathbf{u}, \mathbf{v}$ in a Hilbert space $(V, \langle \triangleright, \triangleleft \rangle)$ we have*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

Proof. This follows from the non-negativity of the inner product. We can assume that both vectors $\mathbf{u}, \mathbf{v}$ are nonzero. For $a, b \in \mathbb{F}$,

$$0 \leq \langle a\mathbf{u} + b\mathbf{v}, a\mathbf{u} + b\mathbf{v} \rangle = |a|^2 \|\mathbf{u}\|^2 + |b|^2 \|\mathbf{v}\|^2 + a\bar{b}\langle \mathbf{u}, \mathbf{v} \rangle + b\bar{a}\overline{\langle \mathbf{u}, \mathbf{v} \rangle}.$$

Then note that $a\bar{b}\langle \mathbf{u}, \mathbf{v} \rangle + b\bar{a}\overline{\langle \mathbf{u}, \mathbf{v} \rangle} = 2 \operatorname{Re} a\bar{b}\langle \mathbf{u}, \mathbf{v} \rangle$. Choose $a = \|\mathbf{u}\|^{-1}$ and b such that $|b| = \|\mathbf{v}\|^{-1}$ and $a\bar{b}\langle \mathbf{u}, \mathbf{v} \rangle < 0$. Then

$$0 \leq 2 - 2\frac{|\langle \mathbf{u}, \mathbf{v} \rangle|}{\|\mathbf{u}\| \|\mathbf{v}\|},$$

which provides the desired result. ■

The following characterization of norms that are induced from inner products is helpful.

Theorem 1.33. *A Banach space $(V, \| \cdot \|)$ is a Hilbert space if and only if the norm satisfies*

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2.$$

1.4.1 ▪ Orthogonality

Orthogonal vectors

The notion of orthogonality is the primary reason to work with Hilbert spaces. And, in particular, orthonormal systems are particularly nice to work with.

Definition 1.34.

- *A sequence of vectors $(\mathbf{u}_j)_{j\geq 1}$ in a Hilbert space V are said to form an orthonormal system if*

$$\langle \mathbf{u}_j, \mathbf{u}_k \rangle = \delta_{jk}, \quad j, k \geq 1.$$

- *A sequence of vectors $(\mathbf{u}_j)_{j\geq 1}$ in a Hilbert space V are said to form an orthonormal basis if their linear span is dense in V.*

We now have the following characterization of orthonormal systems and bases (see [3]).

Theorem 1.35. *Suppose $(\mathbf{u}_j)_{j\geq 1}$ is an orthonormal system in a Hilbert space V. Then*

(1) Bessel's inequality holds for $\mathbf{v} \in V$

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

(2) For $\mathbf{v} \in V$, the series

$$\sum_{j=1}^{\infty} \langle \mathbf{v}, \mathbf{u}_j \rangle \mathbf{u}_j,$$

converges in V.

(3) If $v = \sum_{j=1}^{\infty} a_j \mathbf{u}_j$ then $a_j = \langle \mathbf{v}, \mathbf{u}_j \rangle$.

(4) A series $\sum_{j=1}^{\infty} a_j \mathbf{u}_j$ converges in V if and only if $\sum_{j=1}^{\infty} |a_j|^2 < \infty$.

Proof. The proof is instructive because it shows how to work with series involving orthonormal vectors. For Bessel's inequality, (1), by induction, one can show that

$$\left\| \mathbf{v} - \sum_{j=1}^{n} \langle \mathbf{v}, \mathbf{u}_j \rangle \mathbf{u}_j \right\|^2 = \|\mathbf{v}\|^2 - \sum_{j=1}^{n} |\langle \mathbf{v}, \mathbf{u}_j \rangle|^2.$$

Since this must be non-negative, we take the limit as $n \to \infty$ to obtain Bessel's inequality.

For (2), we show that it is a Cauchy sequence: For $m > n$

$$\left\|\sum_{j=1}^{n}\langle \mathbf{v},\mathbf{u}_j\rangle\mathbf{u}_j - \sum_{j=1}^{m}\langle \mathbf{v},\mathbf{u}_j\rangle\mathbf{u}_j\right\|^2 = \left\|\sum_{j=n+1}^{m}\langle \mathbf{v},\mathbf{u}_j\rangle\mathbf{u}_j\right\|^2 = \sum_{j=n+1}^{m}|\langle \mathbf{v},\mathbf{u}_j\rangle|^2 \leq \sum_{j=n+1}^{\infty}|\langle \mathbf{v},\mathbf{u}_j\rangle|^2,$$

where the latter series converges due to Bessel's inequality. This shows the sequence is Cauchy and hence it converges.

For (3), consider $\mathbf{v}_n = \sum_{j=1}^{n} a_j\mathbf{u}_j$. By orthogonality, we obtain, for $j \leq n$, $\langle \mathbf{v}_n, \mathbf{u}_j\rangle$. But then, we see that by the Cauchy-Schwarz inequality the inner product is continuous:

$$|\langle \mathbf{v},\mathbf{u}_j\rangle - a_j| = |\langle \mathbf{v},\mathbf{u}_j\rangle - \langle \mathbf{v}_n,\mathbf{u}_j\rangle| \leq \|\mathbf{v}-\mathbf{v}_n\|\|\mathbf{u}_j\| \to 0,$$

as $n \to \infty$. Thus $\langle \mathbf{v},\mathbf{u}_j\rangle = a_j$.

Lastly, for (4), from the proof of (2) we see that $\sum_{j=1}^{\infty}|a_j|^2 < \infty$ is equivalent to the sequence of partial sums being Cauchy in V. ■

Theorem 1.36. *Suppose $(\mathbf{u}_j)_{j\geq 1}$ is an orthonormal system in a Hilbert space V. Then the following are equivalent:*

(1) $(\mathbf{u}_j)_{j\geq 1}$ is an orthonormal basis for V.

(2) Plancharel's identity holds: For any $\mathbf{u},\mathbf{v} \in V$

$$\langle \mathbf{u},\mathbf{v}\rangle = \sum_{j=1}^{\infty}\langle \mathbf{u},\mathbf{u}_j\rangle\overline{\langle \mathbf{v},\mathbf{u}_j\rangle}.$$

(3) Parseval's equality holds: For all $\mathbf{v} \in V$

$$\|\mathbf{v}\|^2 = \sum_{j=1}^{\infty}|\langle \mathbf{v},\mathbf{u}_j\rangle|^2.$$

(4) For $\mathbf{v} \in V$, if $\langle \mathbf{v},\mathbf{u}_j\rangle = 0$ for all j then $\mathbf{v} = 0$.

Proof. (1) ⇒ (2): For $(\mathbf{u}_j)_{j\geq 1}$ to be an orthonormal basis, its linear span must be dense, that is the subspace consisting of all finite linear combinations of these vectors is dense in V. So, that implies that there exists coefficients $a_j(n)$ and an integer $m(n)$ such that for any $\mathbf{v} \in V$

$$\mathbf{v} = \lim_{n\to\infty}\mathbf{v}_n, \quad \mathbf{v}_n := \sum_{j=1}^{m(n)} a_j(n)\mathbf{u}_j.$$

But we know from Theorem 1.35(c) that $a_j(n) = \langle \mathbf{v}_n, \mathbf{u}_j\rangle$. Next, define

$$\tilde{\mathbf{v}}_n := \sum_{j=1}^{m(n)}\langle \mathbf{v},\mathbf{u}_j\rangle\mathbf{u}_j,$$

and consider

$$\|\mathbf{v}_n - \tilde{\mathbf{v}}_n\|^2 = \sum_{j=1}^{m(n)} |\langle \mathbf{v} - \mathbf{v}_n, \mathbf{u}_j\rangle|^2 \leq \|\mathbf{v} - \mathbf{v}_n\|^2,$$

by Bessel's inequality. Thus $\tilde{\mathbf{v}}_n \to \mathbf{v}$ and we can represent

$$\mathbf{v} = \sum_{j=1}^{\infty} \langle \mathbf{v}, \mathbf{u}_j\rangle \mathbf{u}_j, \quad \mathbf{u} = \sum_{j=1}^{\infty} \langle \mathbf{u}, \mathbf{u}_j\rangle \mathbf{u}_j.$$

Truncating these series to n terms, computing the result and sending $n \to \infty$ gives (1) $\Rightarrow$ (2).

(2) $\Rightarrow$ (3): This follows from by taking $\mathbf{u} = \mathbf{v}$.

(3) $\Rightarrow$ (4): This follows directly from (3).

(4) $\Rightarrow$ (1): Suppose that the closure of the linear span does not include a vector $\mathbf{u} \in V$. But we know that

$$\tilde{\mathbf{u}} = \sum_{j=1}^{\infty} \langle \mathbf{u}, \mathbf{u}_j\rangle \mathbf{u}_j,$$

converges in V. And $\langle \mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u}_j\rangle = 0$ for all j. Thus we obtain a contradiction because $\tilde{\mathbf{u}} = \mathbf{u}$ and $\tilde{\mathbf{u}}$ is in the closure of the linear span. ■

Orthogonal subspaces

Another important aspect of Hilbert spaces is that they give access to projections.

Definition 1.37.

- *Two elements* $\mathbf{v}, \mathbf{u}$ *of a Hilbert space* V *are said to be orthogonal if* $\langle \mathbf{u}, \mathbf{v}\rangle = 0$.
- *Let* W *be a subset of a Hilbert space* V*, we define the orthogonal complement of* W *via*

$$W^{\perp} = \{\mathbf{v} \in V \mid \langle \mathbf{v}, \mathbf{u}\rangle = 0 \textit{ for all } \mathbf{u} \in W\}.$$

The follow is a direct consequence of the continuity of the inner product.

Theorem 1.38. *The orthogonal complement of a subset of a Hilbert space is a closed subspace.*

The following shows the existence of a projection operator onto any closed subspace.

Theorem 1.39. *Let W be a closed subspace of a Hilbert space V. Then*

- *For every $\mathbf{v} \in V$ there is a unique closest vector $\mathbf{w} = \mathbf{w}(\mathbf{v}) \in W$ such that*

$$\|\mathbf{v} - \mathbf{w}\| = \inf_{\mathbf{u} \in W} \|\mathbf{v} - \mathbf{u}\|.$$

- *The vector $\mathbf{w}(\mathbf{v})$ is the only vector in W that satsifies $\mathbf{v} - \mathbf{w} \in W^\perp$.*

Proof. For the first claim: By the definition of the infimum, there is a sequence of vectors $\mathbf{u}_j \in W$ such that

$$d \leq \|\mathbf{v} - \mathbf{u}_j\| < d + 1/j, \quad d = \inf_{\mathbf{u} \in W} \|\mathbf{v} - \mathbf{u}\|.$$

We want to show that this sequence converges, and therefore, we must show it is Cauchy. Since we have bounds on $\|\mathbf{v} - \mathbf{u}_j\|$ it makes sense to use Theorem 1.33:

$$\|\mathbf{u}_j - \mathbf{u}_k\|^2 = 2\left[\|\mathbf{v} - \mathbf{u}_j\|^2 + \|\mathbf{v} - \mathbf{u}_k\|^2 - 2\|\mathbf{v} - (\mathbf{u}_j + \mathbf{u}_k)/2\|^2\right]. \tag{1.3}$$

Then we know that $(\mathbf{u}_j - \mathbf{u}_k)/2 \in W$ so this last norm must be at least as large as d. So, we bound

$$\|\mathbf{u}_j - \mathbf{u}_k\|^2 \leq 2(d + 1/j)^2 + 2(d + 1/k)^2 - 4d^2, \tag{1.4}$$

which shows that the sequence is Cauchy and must converge in W because W is closed. Let $\mathbf{w}$ denote the limit. It follows that $\|\mathbf{v} - \mathbf{w}\| = d$. To show uniqueness, suppose $\mathbf{z}$ is another vector in $\mathbf{W}$ satisfying $\|\mathbf{z} - \mathbf{w}\| = d$. We can use Theorem 1.33 again to find

$$\|\mathbf{z} - \mathbf{w}\|^2 = 4d^2 - 4\|\mathbf{v} - (\mathbf{w} + \mathbf{z})/2\|^2 \leq 0.$$

And uniqueness follows.

For the second claim: Here one should think of the first minimization as an objective functional. And then one looks for a linear equation that characterizes the minimizer. So, for $\mathbf{z} \in W$, and $\alpha \in \mathbb{F}$, consider

$$\|\mathbf{v} - \mathbf{w} - \alpha\mathbf{z}\|^2 = \|\mathbf{v} - \mathbf{w}\|^2 + \alpha^2\|\mathbf{z}\|^2 - 2\,\mathrm{Re}\,\bar{\alpha}\langle \mathbf{v} - \mathbf{w}, \mathbf{z}\rangle.$$

If $\mathbf{v} - \mathbf{w}, \mathbf{z}\rangle \neq 0$, α, sufficiently small, can be chosen to decrease the right-hand side below $\|\mathbf{v} - \mathbf{w}\|$, which cannot happen. So, $\mathbf{v} - \mathbf{w} \in W^\perp$. Now, if $\mathbf{z}$ also satisfies $\mathbf{v} - \mathbf{z} \in W^\perp$, we find that because $W^\perp$ is itself a linear subspace, that $\mathbf{z} - \mathbf{w}$ lies in both $W^\perp$ and W —- it must vanish. ■

Definition 1.40.

- *Two subspaces W and Z of a Hilbert space V are said to be orthogonal if*

$$\langle \mathbf{w}, \mathbf{v}\rangle = 0, \text{ for all } \mathbf{v} \in V, \mathbf{w} \in W.$$

- *The direct sum of two orthogonal subspaces W and Z of a Hilbert space V is* *defined by*

$$W \oplus Z = \{\mathbf{w} + \mathbf{z} \mid \mathbf{v} \in V, \mathbf{w} \in W\},$$

Proposition 1.41. *Suppose W and Z are two orthogonal subspaces of a Hilbert space V. For $\mathbf{v} \in W \oplus Z$ there exists a unique representation $\mathbf{v} = \mathbf{w} + \mathbf{z}$, $\mathbf{v} \in V, \mathbf{w} \in W$.*

Proof. By the definition of the direct sum there exists a decomposition $\mathbf{v} = \mathbf{w} + \mathbf{z}$. Suppose that $\mathbf{v} = \mathbf{w}_1 + \mathbf{z}_1$ is another such decomposition. Then we find that

$$\mathbf{w} - \mathbf{w}_1 = \mathbf{z}_1 - \mathbf{z}.$$

The only intersection of orthogonal subspaces is at the origin. ■

1.5 ▪ The dual of a Banach or Hilbert space

Definition 1.42. *For a Banach space V, its dual V^* is the vector space of bounded linear functionals on V. The space V^* is a Banach space with the induced operator norm.*

Example 1.43. Consider $V = C([0,1])$. Then the functionals $\mathcal{E}_x$ defined by $f \mapsto f(x)$ are bounded linear functionals.

When considering the dual of a Hilbert space, the following is important.

Theorem 1.44 (Riesz Representation). *For every bounded linear functional ℓ on a Hilbert space V there exists a unique vector $\mathbf{l}$ such that*

$$\ell(\mathbf{v}) = \langle \mathbf{v}, \mathbf{l} \rangle.$$

The following is immediate.

Corollary 1.45. *For a Hilbert space V, V^* is isometric to V.*

1.6 ▪ Sequence spaces

Sequence spaces can be thought of as function spaces for functions defined on $\mathbb{N}$ or $\mathbb{Z}$. Above, we define $L^p(\mu)$ and by μ the domain of functions becomes clear. For sequences, we no longer us the language of measures (we could!), and we change the notation.

Definition 1.46. *For $1 \leq p < \infty$, the unweighted sequence spaces are defined by*

$$\ell^p(\mathbb{N}) = \left\{ \mathbf{a} = (a_j)_{j \geq 1} \mid \|\mathbf{a}\|_p := \left(\sum_{j=1}^{\infty} |a_j|^p \right)^{1/p} < \infty \right\},$$

$$\ell^p(\mathbb{Z}) = \left\{ \mathbf{a} = (a_j)_{j=-\infty}^{\infty} \mid \|\mathbf{a}\|_p := \left(\sum_{j=-\infty}^{\infty} |a_j|^p \right)^{1/p} < \infty \right\}.$$

Definition 1.47. *For a non-zero (discete) weight function w we define the weighted sequence spaces*

$$\ell^p_w(\mathbb{N}) = \left\{ \mathbf{a} = (a_j)_{j \geq 1} \mid \|\mathbf{a}\|_p := \left(\sum_{j=1}^{\infty} |a_j|^p w(j) \right)^{1/p} < \infty \right\},$$

$$\ell^p_w(\mathbb{Z}) = \left\{ \mathbf{a} = (a_j)_{j=-\infty}^{\infty} \mid \|\mathbf{a}\|_p := \left(\sum_{j=-\infty}^{\infty} |a_j|^p w(j) \right)^{1/p} < \infty \right\}.$$

And lastly, for $p = \infty$ we have a separate definition.

Definition 1.48.

$$\ell^\infty_w(\mathbb{N}) = \left\{ \mathbf{a} = (a_j)_{j \geq 1} \mid \|\mathbf{a}\|_\infty := \sup_j |a_j w(j)| < \infty \right\},$$

$$\ell^\infty_w(\mathbb{Z}) = \left\{ \mathbf{a} = (a_j)_{j=-\infty}^{\infty} \mid \|\mathbf{a}\|_\infty := \sup_j |a_j w(j)| < \infty \right\},$$

where we use $\ell^\infty(\mathbb{N}), \ell^\infty(\mathbb{N})$ if $w(j) = 1$ for all j.

It follows immediately that these are all Banach spaces by taking μ to have point masses at j with strength $w(j)$. And, then, of course $\ell^2_w(\mathbb{Z}), \ell^2_w(\mathbb{N})$ are Hilbert spaces.

1.7 ▪ Linear operators

We discuss linear operators in much more detail below, but we introduce some basic notation here.

Definition 1.49. *Suppose V and W are two vector spaces (over $\mathbb{C}$, for simplicity). A linear operator $\mathcal{L}$ from V to W is a function such that $\mathcal{L}(\mathbf{v}) = \mathcal{L}\mathbf{v} \in Y$ for each $\mathbf{v} \in X$ and satisfies*

$$\mathcal{L}(a\mathbf{v} + b\mathbf{w}) = a\mathcal{L}\mathbf{v} + b\mathcal{L}\mathbf{w},$$

for all $\mathbf{v}, \mathbf{w} \in V$ *and* $a, b \in \mathbb{C}$.

If V *and* W *have norms,* $\| \diamond \|_V, \| \diamond \|_W$, *respectively, then we say* $\mathcal{L}$ *is bounded if*

$$\|\mathcal{L}\|_{V\to W} := \sup_{\mathbf{v}\in V:\mathbf{v}\neq\mathbf{0}} \frac{\|\mathcal{L}\mathbf{v}\|_W}{\|\mathbf{v}\|_V} = \sup_{\mathbf{v}\in V:\|\mathbf{v}\|_V=1} \|\mathcal{L}\mathbf{v}\|_W < \infty.$$

Otherwise, we say $\mathcal{L}$ *is unbounded.*

Chapter 2

Fourier series: The foundation of approximation theory

2.1 ▪ Fourier series

Before, we begin, a technicality should be mentioned. Define

$$\mathbb{T} = [0, 2\pi),$$

with metric $d(\theta, \phi) = \min_{n \in \{0,1,-1\}} |\theta - \phi + 2n\pi|$. Note that for ϵ small,

$$d(0, 2\pi - \epsilon) = \epsilon.$$

so that this topology identifies $2\pi \sim 0$. A continuous function f on $\mathbb{T}$ must satisfy

$$\lim_{\theta \to 0, \theta > 0} f(\theta) = \lim_{\theta \to 2\pi, \theta < 2\pi} f(\theta).$$

The Fourier series for a function $f \in L^2(\mathbb{T})$ is given by

$$\begin{aligned} f(\theta) &= \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} c_j(f)\, \mathrm{e}^{\mathrm{i}j\theta}, \\ c_j(f) &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}j\theta} f(\theta) \mathrm{d}\theta, \quad n = 0, \pm 1, \dots. \end{aligned}$$

Note that $c_j(f)$ is nothing more than the inner product of f with $u_j(\theta) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}j\theta}$. It is clear that $(u_j)_{j=-\infty}^{\infty}$ forms an orthonormal system in the Hilbert space $L^2(\mathbb{T})$. But why does it span?

Theorem 2.1. *The orthonormal system* $\left(\frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}j\diamond}\right)_{j=-\infty}^{\infty}$ *in* $L^2(\mathbb{T})$ *is an orthonormal basis.*

Proof. For $f \in L^2(\mathbb{T})$, consider the function

$$g(z) = \frac{\mathrm{i}\sqrt{2\pi}}{1 - \mathrm{e}^{-2\pi \mathrm{i} z}} \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}z\theta} f(\theta) \mathrm{d}\theta.$$

This is a meromorphic function of z. We then compute

$$\mathrm{Res}_{z=j}g(z) = c_j(f).$$

So, suppose that $c_j(f) = 0$ for all j. We see that $g(z)$ is a holomorphic function. We wish to examine its behavior for large z.

For z in the lower-half plane, $|\,\mathrm{e}^{-\mathrm{i}z\theta}\,| \leq 1$ giving

$$|g(z)| \leq \frac{2\pi\|f\|_2}{|1-\mathrm{e}^{-2\pi\mathrm{i}z}\,|}, \quad \mathrm{Im}\, z \leq 0.$$

Then for $z = x + \mathrm{i}y$

$$\mathrm{Re}(1-\mathrm{e}^{-2\pi\mathrm{i}z}) = \mathrm{Re}(1-\mathrm{e}^{-2\pi\mathrm{i}x}\,\mathrm{e}^{2\pi y}) = 1-\mathrm{e}^{2\pi y}\cos(2\pi x).$$

If we set $x = \pm(j+1/2)$, $j \in \mathbb{N}$, this is bounded below by 1. And if we set $y = -(j+1/2)$, $x \in [-(j+1/2), j+1/2]$ then this is bounded below by $1-\mathrm{e}^{-2\pi(j+1/2)}$. This gives the bound

$$|g(z)| \leq 4\pi\|f\|_2, \quad z \in C_j^-, \quad \mathrm{e}^{-2\pi(j+1/2)} < 1/2, \quad (j \geq 1),$$

where C_j^- is the lower-half of a square with side length $2j+1$ centered at the origin.

For z in the upper-half plane, write

$$g(z) = \frac{\mathrm{i}\sqrt{2\pi}}{\mathrm{e}^{2\pi\mathrm{i}z}-1}\int_0^{2\pi}\mathrm{e}^{\mathrm{i}z(2\pi-\theta)}\,f(\theta)\mathrm{d}\theta.$$

Similar calculations give the same bound on $|g(z)|$ for $z \in C_j^+$ where C_j^+ is the upper-half of a square with side length $2j+1$ centered at the origin.

Cauchy's theorem then gives for z inside $C_j^+ \cup C_j^-$

$$g(z) = \frac{1}{2\pi}\int_{C_j^+\cup C_j^-}\frac{g(z')}{z'-z}\mathrm{d}z,$$
$$|g(z)| \leq 2\|f\|_2\int_{C_j^+\cup C_j^-}\frac{|\mathrm{d}z|}{|z'-z|}.$$

Then by doubling j, for z inside $C_j^+ \cup C_j^-$

$$|g(z)| \leq 2\|f\|_2\int_{C_{2j}^+\cup C_{2j}^-}\frac{|\mathrm{d}z|}{|z'-z|} \leq \|f\|_2\frac{8j+4}{j} \leq 12\|f\|_2.$$

Thus, by Liouville's theorem, g must be a constant.

Then for $z = \mathrm{i}y$, we evaluate

$$\left|\int_0^{2\pi}\mathrm{e}^{-\mathrm{i}z\theta}\,f(\theta)\mathrm{d}\theta\right| \leq \int_0^{2\pi}\mathrm{e}^{-y\theta}\,|f(\theta)|\mathrm{d}\theta \leq \|f\|_2\|\,\mathrm{e}^{-y\diamond}\,\|_2 \to 0,$$

as $y \to 0$. Thus $g(z) = 0$ for all z. From Theorem 1.36(4), it remains to show that $f(\theta) = 0$ for almost every θ. But this follows just by evaluating g on the imaginary axis as the integral must vanish identically there (see the exercise below).

■

Exercise 2.1. *Let $f \in L^1([0,b])$. Show that if*

$$\int_0^b \mathrm{e}^{\alpha x} f(x)\mathrm{d}x = 0,$$

for $\alpha \in [0,\epsilon)$, for any $\epsilon > 0$, then $f = 0$.

We introduce the periodic L^2-based Sobolev spaces.

Definition 2.2. *The Sobolev space $H^s(\mathbb{T})$, $s \geq 0$, is given by*

$$H^s(\mathbb{T}) := \left\{ f \in L^2(\mathbb{T}) : \sum_{j=-\infty}^{\infty} |c_j(f)|^2(1+|j|)^{2s} < \infty \right\},$$

and the norm is given by

$$\|f\|_{H^s}^2 := \sum_{j=-\infty}^{\infty} |c_j(f)|^2(1+|j|)^{2s}.$$

This is simply the space of functions whose Fourier coefficients behave sufficiently well so that the sum is finite. It is a Hilbert space. Note that while $s < 0$ is possible, we will not need this case as it no longer corresponds to a space of functions. It is also important to note that if $f \in H^s(\mathbb{T})$ and $s > 1/2$ then

$$\sum_{j=-\infty}^{\infty} |c_j(f)| = \sum_{j=-\infty}^{\infty} |c_j(f)|(1+|j|)^s(1+|j|)^{-s} \leq \|f\|_{H^s} \sum_{j=-\infty}^{\infty} (1+|j|)^{-2s}.$$

This shows the Fourier series for f is absolutely summable and thus f can be taken to be continuous (why?). The following gives convergence in the H^t norm for $s > 1/2, t < s$.

We consider the truncations of the Fourier series for a function $f \in H^s(\mathbb{T})$ to N terms. To keep the convention the same as the discrete Fourier transform in the next section, define

$$N_+ = \lfloor N/2 \rfloor, \quad N_- = \lfloor (N-1)/2 \rfloor.$$

Note that if $N = 2M+1$ is odd then

$$N_+ + N_- + 1 = M + M + 1 = N.$$

And if $N = 2M$ is even

$$N_+ + N_- + 1 = M + M - 1 + 1 = N.$$

Then define the projection operator $\mathcal{P}_N$ by

$$\mathcal{P}_N f(\theta) = \frac{1}{\sqrt{2\pi}} \sum_{j=-N_-}^{N_+} c_j(f)\, \mathrm{e}^{\mathrm{i}j\theta}\,.$$

The Sobolev spaces $H^s(\mathbb{T})$ are convenient for analyzing the convergence of this approximation.

Theorem 2.3. *Suppose $t < s$ and $f \in H^s(\mathbb{T})$. Then there exists a constant $P_{t,s} > 0$ (independent of f and N) such that*

$$\|f - \mathcal{P}_N f\|_{H^t} \leq N^{t-s} P_{t,s} \|f\|_{H^s}.$$

Proof. The proof is a relatively straightforward calculation:

$$\|f - \mathcal{P}_N f\|_{H^t}^2 = \sum_{j=N_+ + 1}^{\infty} |c_j(f)|^2 (1 + |j|)^{2t} + \sum_{j=-\infty}^{-N_- - 1} |c_j(f)|^2 (1 + |j|)^{2t}.$$

We estimate the first sum:

$$\begin{aligned}\sum_{j=N_+ + 1}^{\infty} |c_j(f)|^2 (1 + |j|)^{2t} &= \sum_{j=N_+ + 1}^{\infty} |c_j(f)|^2 (1 + |j|)^{2s} (1 + |j|)^{2(t-s)} \\ &\leq (2 + N_+)^{2(t-s)} \sum_{j=N_+ + 1}^{\infty} |c_j(f)|^2 (1 + |j|)^{2s}.\end{aligned}$$

Since $N_+ \geq N/2 - 1$ we bound $(2 + N_+)^{2(t-s)} \leq N^{2(t-s)} 2^{2(s-t)}$. Similar calculations follow for the other sum and we find that $P_{s,t} = 2^{s-t}$. ■

2.2 ▪ Discrete Fourier transform

We compute the coefficients $c_j(f)$ using the trapezoidal rule with N points

$$\begin{gathered} c_j(f) \approx \check{c}_j(f) := \frac{\sqrt{2\pi}}{N} \sum_{\ell=1}^{N} f(\check{\theta}_\ell)\, \mathrm{e}^{-\mathrm{i} j \check{\theta}_\ell}, \\ \check{\theta}_\ell = 2\pi \frac{\ell - 1}{N}. \end{gathered}$$

We note that the dependence of $\check{c}_j$ and $\check{\theta}_j$ on N is implicit in all that follows. It is important to note that

$$\exp(-\mathrm{i} j \check{\theta}_\ell) = \exp\left(-2\pi \mathrm{i} \frac{j}{N} (\ell - 1)\right).$$

So, if $j = pN + q$, we find

$$\check{c}_j(f) = \check{c}_q(f).$$

And so, it only makes sense to consider N coefficients $\check{c}_j$. But which ones? Since we aim to compute approximate the sum as j ranges from $-\infty$ to $+\infty$, then it makes sense to compute $\check{c}_j$ for $j = -N_-, -N_- + 1, \ldots, N_+$. Typically, for reasons we will discuss, one uses $N = 2^n$. In matrix-vector notation

$$\begin{bmatrix} \\ \check{c}_j \\ \\ \end{bmatrix}_{j=-N_-}^{N_+} = \frac{\sqrt{2\pi}}{N} \underset{\ell=N \qquad\qquad\qquad\qquad 1}{\begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & \mathrm{e}^{-\mathrm{i} j \check{\theta}_\ell} & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix}_{j=-N_-}^{N_+}} \begin{bmatrix} \\ f(\check{\theta}_\ell) \\ \\ \end{bmatrix}_{\ell=1}^{N}$$

The matrix

$$D_N = \begin{array}{cl} \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & \mathrm{e}^{-\mathrm{i}j\check{\theta}_\ell} & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix} & \begin{array}{l} N_+ \\ \\ \\ j=-N_- \end{array} \\ \begin{array}{ccccc} \ell=N & & & & 1 \end{array} & \end{array} ,$$

is called the discrete Fourier transform (DFT) matrix.

From the coefficients $\check{c}_j$, $j = -N_-, \ldots, N_+$, we construct an approximation to $f : \mathbb{T} \to \mathbb{C}$ via

$$f(\theta) \approx \mathcal{I}_N f(\theta) := \frac{1}{\sqrt{2\pi}} \sum_{j=-N_-}^{N_+} \check{c}_j(f)\, \mathrm{e}^{\mathrm{i}j\theta} .$$

The first, most basic property of $\mathcal{I}_N$ is given by the following proposition and shows that $\mathcal{I}_N$ constructs an interpolation.

Proposition 2.4. *For $\ell = 1, 2, \ldots, N$,*

$$f(\check{\theta}_\ell) = \mathcal{I}_N f(\check{\theta}_\ell).$$

Furthermore, if $f(\theta) = \sum_{j=-N_-}^{N_+} c_j\, \mathrm{e}^{\mathrm{i}j\theta}$, then $\mathcal{I}_N f = f$.

Proof. We begin with the expression

$$\check{c}_j(f) = \frac{1}{N\sqrt{2\pi}} \sum_{\ell=1}^{N} f(\check{\theta}_\ell)\, \mathrm{e}^{-\mathrm{i}j\check{\theta}_\ell},$$

and then evaluate

$$\mathcal{I}_N f(\check{\theta}_m) = \frac{1}{N} \sum_{j=-N_-}^{N_+} \sum_{\ell=1}^{N} f(\check{\theta}_\ell)\, \mathrm{e}^{-\mathrm{i}j\check{\theta}_\ell}\, \mathrm{e}^{\mathrm{i}j\check{\theta}_m} .$$

Then we evaluate

$$\begin{aligned} \sum_{j=-N_-}^{N_+} \mathrm{e}^{-\mathrm{i}j\check{\theta}_\ell}\, \mathrm{e}^{\mathrm{i}j\check{\theta}_m} &= \sum_{j=-N_-}^{N_+} \mathrm{e}^{\mathrm{i}j(\check{\theta}_\ell - \check{\theta}_m)} = \mathrm{e}^{-\mathrm{i}N_-(\check{\theta}_\ell - \check{\theta}_m)} \sum_{j=0}^{N-1} \mathrm{e}^{\mathrm{i}(j+N_-)(\check{\theta}_\ell - \check{\theta}_m)} \\ &= \begin{cases} N & \ell = m, \\ 0 & \ell \neq m. \end{cases} \end{aligned}$$

This establishes the first claim. The second claim follows from the fact that we have just shown the DFT matrix is invertible. ■

Another important property of $\mathcal{I}_N$ is that if $j = pN + q$, $-N_- \leq q \leq N_+$ then

$$\mathcal{I}_N\, \mathrm{e}^{\mathrm{i}j\diamond} = \mathrm{e}^{\mathrm{i}q\diamond} .$$

This then implies that if

$$f(\theta) = \sum_{j=-\infty}^{\infty} c_j(f)\, \mathrm{e}^{\mathrm{i}j\theta},$$

is absolutely convergent, then

$$\mathcal{I}_N f(\theta) = \sum_{j=-N_-}^{N_+} \sum_{k=-\infty}^{\infty} c_{kN+j}(f)\, \mathrm{e}^{\mathrm{i}j\theta}\,.$$

In other words,

$$\check{c}_j(f) = \sum_{k=-\infty}^{\infty} c_{kN+j}(f). \tag{2.1}$$

The previous proposition, while important, says nothing about $|f(\theta) - \mathcal{I}_N f(\theta)|$ for $\theta \neq \check{\theta}_\ell$. The following theorem does and is proved using (2.1) carefully.

Theorem 2.5 (Kress & Sloan). *Suppose $s > 1/2$, $t < s$, and $f \in H^s(\mathbb{T})$. Then there exists constants $D_{t,s}, C_{t,s} > 0$ such that*

$$\|\mathcal{P}_N f - \mathcal{I}_N f\|_{H^t} \leq N^{t-s} D_{t,s} \|f\|_{H^s},$$

and therefore

$$\|f - \mathcal{I}_N f\|_{H^t} \leq N^{t-s} C_{t,s} \|f\|_{H^s}.$$

Proof. We note that $s > 1/2$ is sufficient to ensure that the Fourier series for f is absolutely convergent. We use the notation

$$\sum_{k=a}^{b}{}' = \sum_{k=a, k\neq 0}^{b}.$$

Then compute

$$\begin{aligned} \mathcal{I}_N f(\theta) - \mathcal{P}_N f(\theta) &= \sum_{j=-N_-}^{N_+} \sum_{k=-\infty}^{\infty} c_{kN+j}(f)\, \mathrm{e}^{\mathrm{i}j\theta} - \sum_{j=-N_-}^{N_+} c_j(f)\, \mathrm{e}^{\mathrm{i}j\theta} \\ &= \sum_{j=-N_-}^{N_+} \sum_{k=-\infty}^{\infty}{}' c_{kN+j}(f)\, \mathrm{e}^{\mathrm{i}j\theta}\,. \end{aligned} \tag{2.2}$$

Therefore

$$\|\mathcal{P}_N f - \mathcal{I}_N f\|_{H^t}^2 = \sum_{j=-N_-}^{N_+} \left| \sum_{k=-\infty}^{\infty}{}' c_{kN+j}(f) \right|^2 (1+|j|)^{2t}$$
$$= \sum_{j=-N_-}^{N_+} (1+|j|)^{2t} \left| \sum_{k=-\infty}^{\infty}{}' c_{kN+j}(f)(1+|kN+j|)^s \frac{1}{(1+|kN+j|)^s} \right|^2$$
$$\le \sum_{j=-N_-}^{N_+} (1+|j|)^{2t} \left[\sum_{k=-\infty}^{\infty}{}' |c_{kN+j}(f)|^2 (1+|kN+j|)^{2s} \right] \left[\sum_{k=-\infty}^{\infty}{}' (1+|kN+j|)^{-2s} \right]$$
$$\le \|f\|_s \max_{-N_- \le j \le N} \left[(1+|j|)^{2t} \sum_{k=-\infty}^{\infty}{}' (1+|kN+j|)^{-2s} \right].$$

And it remains to estimate

$$\sum_{k=-\infty}^{\infty}{}' \frac{(1+|j|)^{2t}}{(1+|kN+j|)^{2s}}.$$

We first write,

$$\frac{(1+|j|)^{2t}}{(1+|kN+j|)^{2s}} \le N^{-2s} \frac{(1+N/2)^{2t}}{(1/N+|k+j/N|)^{2s}} \le \frac{N^{2(t-s)}}{|k+j/N|^{2s}}.$$

Now, as $s > 1/2$, we have

$$\sum_{k=1}^{\infty} \frac{1}{(k-1/2)^{2s}} < \infty.$$

Similarly,

$$\sum_{k=-\infty}^{-1} \frac{1}{(k+1/2)^{2s}} < \infty.$$

This gives the existence of the constant $D_{t,s}$ (which actually can be taken to only depend on s). And then the triangle inequality

$$\|f - \mathcal{I}_n f\|_{H^s} \le \|f - \mathcal{P}_n f\|_{H^s} + \|\mathcal{P}_n f - \mathcal{I}_n f\|_{H^s},$$

gives the desired result. ■

We can also ask about how well the DFT approximates the true Fourier coefficients. One result can be derived directly from Theorem 2.5.

Corollary 2.6. *Suppose $f \in H^s(\mathbb{T})$ for $s > 1/2$. Then*

$$|\check{c}_j(f) - c_j(f)| \le N^{-s} C_{0,s} \|f\|_{H^s}.$$

Proof. By the Cauchy-Schwarz inequality

$$|\check{c}_j(f) - c_j(f)| = \left| \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} (f(\theta) - \mathcal{I}_N f(\theta))\, \mathrm{e}^{-\mathrm{i}j\theta}\, \mathrm{d}\theta \right| \le \|f - \mathcal{I}_n f\|_2 \le N^{-s} C_{0,s} \|f\|_{H^s}.$$

■

To further put Theorem 2.5 in context, we have the following, which is an example of a Sobolev embedding result.

Theorem 2.7. *Suppose $f \in H^s(\mathbb{T})$ for $s > r+1/2$ and $r \in \mathbb{N}$. Then f can be taken to be r times continuously differentiable and there exists $A_{r,s} > 0$ such that*

$$\max_{0\le k\le r}\left\|\frac{\mathrm{d}^k f}{\mathrm{d}\theta^k}\right\|_\infty \le A_{r,s}\|f\|_{H^s}.$$

Proof. We first note that if $f \in H^s(\mathbb{T})$ and $s > 1/2$ we have

$$\begin{aligned}\|f\|_\infty &\le \frac{1}{\sqrt{2\pi}}\sum_{j=-\infty}^{\infty}|c_j| = \frac{1}{\sqrt{2\pi}}\sum_{j=-\infty}^{\infty}(1+|j|)^s|c_j|(1+|j|)^{-s}\\ &\le \frac{1}{\sqrt{2\pi}}\sqrt{\sum_{j=-\infty}^{\infty}(1+|j|)^{-2s}}\|f\|_{H^s},\end{aligned}$$

where the first term is finite. Moreover, this shows that f is the uniform limit of continuous functions — it can be taken to be continuous and we suppose we are working with this representative of the equivalence class in what follows. This establishes the result when $r = 0$. For the general case, we claim that

$$\frac{\mathrm{d}^k f}{\mathrm{d}\theta^k}(\theta) = \sum_{j=-\infty}^{\infty}(\mathrm{i}j)^k c_j(f)\,\mathrm{e}^{\mathrm{i}j\theta}.$$

This follows from a standard application of the dominated convergence theorem, see [6, Theorem 2.27]. Then we find that

$$\begin{aligned}\left\|\frac{\mathrm{d}^k f}{\mathrm{d}\theta^k}\right\|_\infty \le \sum_{j=-\infty}^{\infty}|j|^k|c_j(f)| &= \sum_{j=-\infty}^{\infty}\frac{|j|^k}{(1+|j|)^s}(1+|j|)^s|c_j(f)|\\ &\le \sqrt{\sum_{j=-\infty}^{\infty}(1+|j|)^{2(k-s)}}\|f\|_{H^s},\end{aligned}$$

which establishes the result at hand because the sum is finite. ■

This implies that convergence in Sobolev spaces implies convergence of derivatives.

Corollary 2.8. *Suppose $f \in H^s(\mathbb{T})$ for $s > k+1/2$ and $r \in \mathbb{N}$. Then there exists a constant $B_{k,s,t}$ such that for any t satisfying $s > t > k+1/2$*

$$\left\|\frac{\mathrm{d}^k}{\mathrm{d}\theta^k}(f - \mathcal{I}_N f)\right\|_\infty \le B_{k,s,t}N^{t-s}\|f\|_{H^s}.$$

Proof. This follows from

$$\left\| \frac{\mathrm{d}^k}{\mathrm{d}\theta^k}(f - \mathcal{I}_N f) \right\|_\infty \leq A_{k,t} \|f - \mathcal{I}_N f\|_{H^t} \leq A_{k,t} C_{t,s} N^{t-s} \|f\|_{H^s}.$$

■

2.3 ▪ Spaces of continuous functions

Definition 2.9. *A function f on $\mathbb{T}$ is said to be α-Hölder continuous, $0 < \alpha \leq 1$, if there exists $L \geq 0$ such that*

$$|f(\theta) - f(\phi) \leq L\, d(\theta, \phi)^\alpha, \text{ for all } \theta, \phi \in \mathbb{T}.$$

If $\alpha = 1$ we also say that f is Lipschitz continous. The smallest value of L such that this inequality holds is called the Hölder (Lipschitz) constant. We denote by $C^{0,\alpha}(\mathbb{T})$ the (non-separable) Banach space of Hölder continuous functions with norm

$$\|f\|_{C^{0,\alpha}} := \|f\|_\infty + \sup_{\theta \neq \phi} \frac{|f(\theta) - f(\phi)|}{d(\theta, \phi)^\alpha}.$$

Exercise 2.2. *Show that $C^{0,\alpha}(\mathbb{T})$ is a non-separable Banach space. (Hint: Consider $(f_a(\theta) = (\theta - a)^\alpha, \theta > a$ and zero otherwise).*

If f is extended via $f(\theta + 2\pi) = f(\theta)$ to a periodic function on $\mathbb{R}$, this is simply equivalent to

$$|f(\theta) - f(\phi) \leq L|\theta - \phi|^\alpha,$$

for the extended function. We assume that all functions are extended periodically below.

We then note that for $N = 2M + 1$ odd,

$$\mathcal{P}_N f(\theta) = \int_0^{2\pi} f(\phi) D_N(\theta - \phi) \mathrm{d}\phi, \quad D_N(\theta - \phi) = \frac{1}{2\pi} \sum_{j=-M}^{M} \mathrm{e}^{\mathrm{i}j(\theta - \phi)}.$$

The function D_N is called the Dirichlet kernel. And since $\overline{D_N(\theta)} = D_N(\theta)$, we see that

$$D_n(\theta, \phi) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{M} \cos(j(\theta - \phi)).$$

Suppose that we can find $\{\rho_j^{(N)}\}_{j=1}^M$, $\rho_1^{(N)} = 1 + O(N^{-1})$, so that

$$\tilde{D}_N(\theta) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{M} \rho_j^{(N)} \cos(j(\theta - \phi)),$$

satisfies $\tilde{D}_N(\theta) \geq 0$.

Next, suppose $\mu = \mu_N$ is a Borel measure on $\mathbb{R}$ that satisfies:

- Periodicity: $\mu(S) = \mu(S + 2\pi n), \quad n \in \mathbb{Z}$,
- Normality: $\int_{[0,2\pi)} \tilde{D}_N(\theta)\mu(\mathrm{d}\theta) = 1$, and
- Orthogonality: $\int_{[0,2\pi)} \cos\theta \cos(k\theta)\mu(\mathrm{d}\theta) = \delta_{k1}$.

Then suppose f is α-Hölder continuous on $\mathbb{T}$ and define

$$\begin{aligned}\tilde{\mathcal{P}}_N f(\theta) &= \int_{[0,2\pi)} f(\phi)\tilde{D}_N(\theta - \phi)\mu(\mathrm{d}\phi)\\ &= \int_{[0,2\pi)} f(\phi)\tilde{D}_N(\phi - \theta)\mu(\mathrm{d}\phi)\\ &= \int_{[-\theta,2\pi-\theta)} f(\theta + \phi)\tilde{D}_N(\phi)\mu(\mathrm{d}\phi).\end{aligned}$$

The, we claim, by periodicity that

$$\int_{[-\theta,0)} f(\theta+\phi)\tilde{D}_N(\phi)\mu(\mathrm{d}\phi) = \int_{[2\pi-\theta,2\pi)} f(\theta+\phi)\tilde{D}_N(\phi)\mu(\mathrm{d}\phi).$$

Thus

$$\tilde{\mathcal{P}}_N f(\theta) = \int_{[0,2\pi)} f(\theta+\phi)\tilde{D}_N(\phi)\mu(\mathrm{d}\phi).$$

Then supposing that f is α-Hölder continuous with Hölder constant L:

$$\begin{aligned}|f(\theta) - \tilde{\mathcal{P}}_N f(\theta)| &\leq \int_0^{2\pi} |f(\theta) - f(\theta+\phi)|\tilde{D}_N(\phi)\mu(\mathrm{d}\phi)\\ &\leq L\int_0^{2\pi} d(0,\phi)^\alpha \tilde{D}_N(\phi)\mu(\mathrm{d}\phi).\end{aligned}$$

We then claim that

$$d(0,\phi) \leq \frac{\pi}{\sqrt{2}}\sqrt{1-\cos\phi}.$$

We apply the Hölder inequality with $p = 2/\alpha$, $q = (1-\alpha/2)^{-1}$ to find

$$\int_0^{2\pi} d(0,\phi)^\alpha \tilde{D}_N(\phi)\mu(\mathrm{d}\phi) \leq \left[\int_0^{2\pi} \frac{\pi^2}{2}(1-\cos\phi)\tilde{D}_N(\phi)\mu(\mathrm{d}\phi)\right]^{\alpha/2} = \frac{\pi^\alpha}{2^{\alpha/2}}\left(1-\rho_1\right)^{\alpha/2}.$$

We arrive at the following

Theorem 2.10. *Suppose ρ_j are chosen such that $\tilde{D}_N$ is non-negative and μ satisfyies the periodicity, normality and orthogonality properities. Then if f is α-Hölder continuous*

$$\|f - \tilde{\mathcal{P}}_N f\|_\infty \leq L\frac{\pi^\alpha}{2^{\alpha/2}}\left(1-\rho_1\right)^{\alpha/2},$$

where L is the Hölder constant for f.

Before we give the two main instances in which this theorem is applied, we note that this all has a, potentially desirable, outcome that if $f(\theta) \geq 0$ for all θ, then $\tilde{P}_N f(\theta) \geq 0$ for all θ.

Lemma 2.11. *Define*

$$\rho_k = \frac{(N-k+2)\cos\left(\frac{k\pi}{N+2}\right) + \sin\left(\frac{k\pi}{N+2}\right)\cot\left(\frac{\pi}{N+2}\right)}{N+2}, \qquad k = 0, 1, \ldots, N.$$

then $\tilde{D}_N$ is non-negative and

$$\frac{\pi}{\sqrt{2}}(1-\rho_1)^{1/2} \leq \frac{\pi^2}{2}(N+2)^{-1}.$$

Proof. Let $\{c_\ell\}_{\ell=0}^N$ be any real numbers. Then

$$\left(\sum_{\ell=0}^N c_\ell \exp(\mathrm{i}\ell\theta)\right)\left(\sum_{\ell=0}^N c_\ell \exp(-\mathrm{i}\ell\theta)\right) = \left|\sum_{\ell=0}^N c_\ell \exp(\mathrm{i}\ell\theta)\right|^2 \geq 0.$$

Expanding and using that $\exp(\mathrm{i}k\theta) + \exp(-\mathrm{i}k\theta) = 2\cos(k\theta)$ we find

$$\left(\sum_{\ell=0}^N c_\ell \exp(\mathrm{i}\ell\theta)\right)\left(\sum_{\ell=0}^N c_\ell \exp(-\mathrm{i}\ell\theta)\right) = \sum_{k=0}^N c_k^2 + 2\sum_{p=1}^n \sum_{k=0}^{N-p} c_k c_{k+p}\cos(p\theta).$$

Because $\tilde{D}_N$ must have the constant term equal to $1/2$ we require $c_0^2 + \ldots + c_N^2 = 1/2$.

Let

$$c_\ell = c\sin\left(\frac{\ell+1}{N+2}\pi\right), \qquad \ell = 0, \ldots, N,$$

where

$$c^2 = \frac{1}{2\sum_{\ell=0}^N \sin^2\left(\frac{\ell+1}{N+2}\pi\right)} = \frac{1}{N+2}.$$

Then setting $\rho_0 = 1$ and

$$\rho_k = 2\sum_{\ell=0}^{N-k} c_\ell c_{c+k}$$

we obtain $\tilde{D}_N$.

Next, we show that these damping coefficients are equal to those described above.

Following [?] we have that

$$\begin{aligned}
2\sum_{\ell=0}^{N-k} c_\ell c_{\ell+k} &= 2c^2 \sum_{\ell=0}^{N-k} \sin\left(\frac{\ell+1}{N+2}\pi\right)\sin\left(\frac{\ell+k+1}{N+2}\pi\right)\\
&= 2c^2 \sum_{\ell=1}^{N-k+1} \sin\left(\frac{\ell}{N+2}\pi\right)\sin\left(\frac{\ell+k}{N+2}\pi\right)\\
&= c^2 \sum_{\ell=1}^{N-k+1}\left(\cos\left(\frac{k}{N+2}\pi\right) - \cos\left(\frac{2\ell+k}{N+2}\pi\right)\right)\\
&= c^2\left((N-k)\cos\left(\frac{k}{N+2}\pi\right) - \operatorname{Re}\sum_{\ell=1}^{N-k+1}\exp\left(\mathrm{i}\frac{2\ell+k}{N+2}\pi\right)\right)\\
&= c^2\left((N-k+1)\cos\left(\frac{k}{N+2}\pi\right) - \sin\left(\frac{\ell}{N+2}\pi\right)\cot\left(\frac{\pi}{N+2}\right)\right)
\end{aligned}$$

This gives the claimed expression for ρ_k.

Using this expression, it is easy to verify that $\rho_1 = \cos(\pi/(N+2))$. Thus

$$(1-\rho_1)^{1/2} = \left(1-\cos\left(\frac{\pi}{N+2}\right)\right)^{1/2} = \sqrt{2}\sin\left(\frac{\pi}{2N+4}\right) \le \sqrt{2}\frac{\pi}{2N+4}$$

so

$$\frac{\pi}{\sqrt{2}}(1-\rho_1)^{1/2} \le \frac{\pi^2}{2N+4}$$

■

Theorem 2.12.

- *If μ is standard Lebesgue measure, then periodicity, normality and orthogonality hold.*
- *If $\mu = \mu_N = \frac{1}{N}\sum_{\ell\in\mathbb{Z}} \delta_{\check{\theta}_\ell}$ then periodicity, normality and orthogonality hold.*

In both cases, if f is α-Hölder continuous

$$\|f - \tilde{\mathcal{P}}_N f\|_\infty \le L\left(\frac{\pi^2}{2N+4}\right)^\alpha.$$

In defining E_N to be the linear span of $(\mathrm{e}^{\mathrm{i}j\diamond})_{j=-N}^N$ we arrive at Jackson's first theorem.

Theorem 2.13 (Jackson's first theorem). *Suppose that $f \in C^{0,\alpha}(\mathbb{T})$ with Hölder constant L. Then if N is odd*

$$\inf_{g\in E_N} \|f-g\|_\infty \le L\left(\frac{\pi^2}{2N+4}\right)^\alpha.$$

If N is even we use $\inf_{g\in E_N}\|f-g\|_\infty \le \inf_{g\in E_{N-1}}\|f-g\|_\infty$.

Remark 2.14. *What is often referred to as Jackson's first theorem in the literature has a slightly better constant than this: $\pi/(2(N+1))$. We use this proof because it is constructive.*

To get to Jackson's second theorem, we define a space of smoother functions.

Definition 2.15. *Define $C^{k,\alpha}(\mathbb{T})$ to be the space of complex-valued, k-times continuously differentiable functions such that $f^{(k)} \in C^{0,\alpha}(\mathbb{T})$ with norm*

$$\|f\|_{C^{k,\alpha}} := \sum_{j=0}^{k-1} \|f^{(j)}\|_\infty + \|f^{(k)}\|_{C^{0,\alpha}}.$$

Theorem 2.16 (Jackson's second theorem). *Suppose that $C^{k,\alpha}(\mathbb{T})$ and $f^{(k)}$ has Hölder constant L. Then if N is odd*

$$\inf_{g\in E_N} \|f-g\|_\infty \leq L\left(\frac{2\pi^2}{2N+4}\right)^{k+\alpha}.$$

If N is even we use $\inf_{g\in E_N}\|f-g\|_\infty \leq \inf_{g\in E_{N-1}}\|f-g\|_\infty$.

Proof. The main idea is that we can apply Jackson's first theorem to $f^{(k)}$, which should then give higher rates of convergence for, in effect, integrals of this function. But it is difficult to try to obtain an approximant of $f^{(k)}$ and then show how its primitives approximate f. The issue is that the primitive of arbitrary element of E_N is not in E_N — it must have zero average. So, we take a different approach. We see that for any $h \in E_N$, if f is continuously differentiable, we have

$$\inf_{g\in E_N}\|f-g\|_\infty = \inf_{g\in E_N}\|(f-h)-g\|_\infty \leq L\frac{\pi^2}{2N+4},$$

where L is the *Lipschitz* constant for $f-h$, $L \leq \|f'-h'\|_\infty$. Thus

$$\inf_{g\in E_N}\|f-g\|_\infty \leq \frac{\pi^2}{2n+4}\inf_{h\in E_N}\|f'-h'\|_\infty = \frac{\pi^2}{2N+4}\inf_{h\in E_N^0}\|f'-h\|_\infty,$$

where E_N^0 is the linear span of $(\mathrm{e}^{\mathrm{i}j\diamond})_{j=-N,j\neq 0}^N$. But now if we add the constraint that $\int_0^{2\pi} f(\theta)\mathrm{d}\theta = 0$, and define $g_0 = \frac{1}{2\pi}\int_0^{2\pi} g(\theta)\mathrm{d}\theta$:

$$\|f-(g-g_0)\|_\infty - |g_0| \leq \|f-g\|_\infty.$$

But

$$|g_0| = \left|\frac{1}{2\pi}\int_0^{2\pi}(g(\theta)-f(\theta))\mathrm{d}\theta\right| \leq \|f-g\|_\infty. \tag{2.3}$$

We find $\|f - (g - g_0)\|_\infty \leq 2\|f - g\|_\infty$. This results in

$$\inf_{g \in E_N^0} \|f - g\|_\infty \leq \frac{2\pi^2}{2N+4} \inf_{h \in E_N^0} \|f' - h\|_\infty .$$

Inductively, we obtain

$$\begin{aligned}
\inf_{g \in E_N} \|f - g\| &\leq \frac{\pi^2}{2N+4} \inf_{h \in E_N^0} \|f' - h\|_\infty \\
&\leq \frac{\pi^2}{2n+4} \frac{2\pi^2}{2N+4} \inf_{h \in E_N^0} \|f'' - h\|_\infty \\
&\leq \cdots \leq \frac{1}{2} \left(\frac{2\pi^2}{2N+4} \right)^k \inf_{h \in E_N^0} \|f^{(k)} - h\|_\infty .
\end{aligned}$$

To finish the proof, we note that we can find $\tilde{h} \in E_N^0$, $c \in \mathbb{C}$ such that

$$\|f^{(k)} - \tilde{h} - c\|_\infty \leq L \left(\frac{\pi^2}{2N+4} \right)^\alpha ,$$

where L is the α-Hölder constant for $f^{(k)}$. Then

$$\inf_{h \in E_N^0} \|f^{(k)} - h\|_\infty \leq \|f^{(k)} - \tilde{\|}_\infty \leq L \left(\frac{\pi^2}{2N+4} \right)^\alpha + |c| .$$

Then, we estimate $|c|$ via

$$|c| = \left| \frac{1}{2\pi} (\tilde{h} + c - f) \mathrm{d}\theta \right| \leq L \left(\frac{\pi^2}{2N+4} \right)^\alpha .$$

Thus

$$\inf_{g \in E_N} \|f - g\| \leq L \left(\frac{2\pi^2}{2N+4} \right)^{k+\alpha} .$$

■

2.3.1 ▪ From best approximation to practical approximation: Lebesgue constants

We equip the finite-dimensional space E_N with the $\| \diamond \|_\infty$ norm. And we interpret $\mathcal{P}_N$, $\mathcal{I}_N$ as linear operators from $C(\mathbb{T})$ to E_N.

Determining bounds on the best approximation is useful to understand how well a given approximation scheme (e.g., $\mathcal{I}_N$) is performing. But it also gives immediate bounds.

Theorem 2.17. *Suppose that $f \in C^{k,\alpha}(\mathbb{T})$ then there exist constants $C_{k,\alpha}, D_{k,\alpha}$ such that for $N > 1$*

$$\|f - \mathcal{P}_n f\|_\infty \leq C_{k,\alpha} \frac{\log N}{N^{k+\alpha}}, \quad \|f - \mathcal{I}_n f\|_\infty \leq D_{k,\alpha} \frac{\log N}{N^{k+\alpha}} .$$

Proof. Let

$$d_N(f) = \inf_{p \in E_N} \|f - p\|_\infty. \tag{2.4}$$

Then for any $\epsilon > 0$ there exists[2] $p_\epsilon \in E_N$ such that $\|f - p_\epsilon\| - \epsilon < d_N(f)$. Then consider

$$\|f - \mathcal{P}_N f\|_\infty \leq \|f - \mathcal{P}_N p_\epsilon\|_\infty + \|\mathcal{P}_N p_\epsilon - \mathcal{P}_N f\|_\infty.$$

Then because $\mathcal{P}_N p_\epsilon = p_\epsilon$ we find

$$\|f - \mathcal{P}_N f\|_\infty \leq (1 + \|\mathcal{P}_N\|_{C(\mathbb{T}) \to E_N})(d_N(f) + \epsilon).$$

Since ϵ is arbitrary, we send it to zero, and it remains to bound $\ell_N := \|\mathcal{P}_N\|_{C(\mathbb{T}) \to E_N}$. It follows that

$$\ell_N = \int_0^{2\pi} |D_N(\theta)| \mathrm{d}\theta.$$

Suppose $N = 2M + 1$

$$\begin{aligned} D_N(\theta) &= \sum_{j=-M}^{M} (\mathrm{e}^{\mathrm{i}\theta})^j = \mathrm{e}^{-M\mathrm{i}\theta} \sum_{0=j}^{2M} (\mathrm{e}^{\mathrm{i}\theta})^j = \mathrm{e}^{-M\mathrm{i}\theta} \frac{1 - \mathrm{e}^{N\mathrm{i}\theta}}{1 - \mathrm{e}^{\mathrm{i}\theta}} \\ &= \frac{\mathrm{e}^{-(M+1/2)\mathrm{i}\theta} - \mathrm{e}^{(M+1/2)\mathrm{i}\theta}}{\mathrm{e}^{-\mathrm{i}\theta/2} - \mathrm{e}^{\mathrm{i}\theta/2}} = \frac{\sin((M+1/2)\theta)}{\sin(\theta/2)}. \end{aligned}$$

If $N = 2M$, we obtain

$$D_N(\theta) = \frac{\sin((M-1/2)\theta)}{\sin(\theta/2)} + \mathrm{e}^{\mathrm{i}M\theta}.$$

By careful analysis of integrals of $\frac{\sin x}{x}$ one obtains

$$\ell_N = O(\log N).$$

Then, going back to Watson [23] and Hardy [8] one also has

$$\|\mathcal{I}_N\|_{C(\mathbb{T}) \to E_N} = O(\log N).$$

More refined estimates are possible, but unnecessary for our purposes. The claims follow. ■

Remark 2.18. *The norms $\|\mathcal{I}_N\|_{C(\mathbb{T}) \to E_N}$, $\|\mathcal{P}_N\|_{C(\mathbb{T}) \to E_N}$ are referred to as the Lebesgue constants for Fourier interpolation and for Fourier projections, respectively.*

Exercise 2.3. *Suppose $K \in C(\mathbb{T})$. Show that*

$$\mathcal{K}f(\theta) := \int_0^{2\pi} K(\theta - \phi) f(\phi) \mathrm{d}\phi$$

is a bounded linear operator on $C(\mathbb{T})$ and

$$\|\mathcal{K}\|_{C(\mathbb{T}) \to C(\mathbb{T})} = \int_0^{2\pi} |K(\theta)| \mathrm{d}\theta.$$

[2] Actually, there exists a unique p such that $d_N(f) = \|f - p\|_\infty$ but we will not need that here.

2.4 ▪ Spaces of analytic functions

As indicated by Jackson's theorem, if f is infinitely differentiable then one can approximate f with a Fourier-like series that should converge at a rate that is faster than any polynomial rate. This *does not* imply it is exponential! But with the assumption of analyticity of f in an appropriate annulus, we get an exponential rate of convergence.

Define $\mathbb{U} = \{z \in \mathbb{C} \mid |z| = 1\}$, with counter-clockwise orientation. Then we have the following space of analytic functions.

Definition 2.19. *For $\rho > 1$, set $A_\rho(\mathbb{U}) = \{z \in \mathbb{C} \mid \rho^{-1} < |z| < \rho\}$. Then define*

$$B_\rho(\mathbb{U}) := \left\{ f : A_\rho \to \mathbb{C} \mid f \text{ is analytic and } \|f\|_\infty = \sup_{z \in A_\rho} |f(z)| < \infty \right\}.$$

We can then establish the following.

Theorem 2.20. *Suppose $g \in B_\rho(\mathbb{U})$, $\rho > 1$, with $M = \sup_{z \in A_\rho(\mathbb{U})} |g(z)|$. Then on $[0, 2\pi)$ for $f(\theta) = g(\mathrm{e}^{\mathrm{i}\theta})$*

$$\|f - \mathcal{P}_N f\|_\infty \leq \frac{2M\rho^{-N_+}}{1 - 1/\rho},$$

and

$$\|f - \mathcal{I}_N f\|_\infty \leq \frac{4M\rho^{-N_+}}{1 - 1/\rho}.$$

Proof. First, we compute, for $z = \mathrm{e}^{\mathrm{i}\theta}$

$$c_j(f) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}j\theta} f(\theta) \mathrm{d}\theta = \frac{1}{2\pi\mathrm{i}} \int_{\mathbb{U}} z^{-j} g(z) \frac{\mathrm{d}z}{z}.$$

If $j \geq 0$, we deform this integral to $\rho'\mathbb{U} := \{z \in \mathbb{C} \mid |z| = \rho'\}$, $1 < \rho' < \rho$. Then we bound

$$|c_j(f)| \leq M(\rho')^{-j-1}$$

If $j \leq 0$, we compute with $\overline{c_j(f)}$, obtaining the same bound. Since this bound holds for every $\rho' < \rho$, it also holds for ρ. Then

$$|f(\theta) - \mathcal{P}_N f(\theta)| \leq \sum_{j > N_+} M\rho^{-|j|-1} + \sum_{j < -N_-} M\rho^{-|j|-1} \leq \frac{2M\rho^{-N_+}}{1 - 1/\rho}.$$

For the second claim, we simply apply (2.2) to obtain

$$|\mathcal{P}_N f(\theta) - \mathcal{I}_N f(\theta)| \leq \frac{2M\rho^{-N_+}}{1 - 1/\rho},$$

and the claim follows.

■

For $g \in B_\rho(\mathbb{U})$ and $1 < \rho' < \rho$, consider the following integral

$$h_N(z) = \frac{1}{2\pi \mathrm{i}} \int_{\partial A'_\rho} \frac{z^{N_+ + 1} - z^{-N_-}}{\zeta^{N_+ + 1} - \zeta^{-N_-}} \frac{g(\zeta)}{\zeta - z} \mathrm{d}z.$$

Lemma 2.21. *If For $z = \mathrm{e}^{\mathrm{i}\theta}$, $\theta \in [0, 2\pi)$, $h_N(z) - g(z) = -\mathcal{I}_N f(\theta)$, $f(\theta) = g(\mathrm{e}^{\mathrm{i}\theta})$.*

Proof. It suffices to show that $h_N(\mathrm{e}^{\mathrm{i}\check{\theta}_\ell}) = 0$ and that $h_N(z) + g(z)$ is in the linear span of $(z^j)_{j=-N_-}^{N_+}$. The first claim follows immediately. For the second, suppose $z \neq z_\ell := \mathrm{e}^{\mathrm{i}\check{\theta}_\ell}$ for all ℓ. Then by a residue calculation at simple poles

$$h_N(z) = f(z) + \sum_{\ell=1}^{N} \frac{z^{N_+ + 1} - z^{-N_-}}{(N_+ + 1) z_\ell^{N_+} + N_- z_\ell^{-N_- - 1}} \frac{g(z_\ell)}{z_\ell - z}.$$

From this, we see that $(h_n(z) - g(z)) z^{N_-}$ is entire and is $O(z^{N_+ + N_-})$ at infinity, meaning it is a polynomial of degree at most $N_+ + N_-$. Thus

$$h_n(z) - g(z) = \sum_{j=N_-}^{N_+} a_j z^j,$$

for some coefficients a_j. ■

This representation of $\mathcal{I}_N$ can also be used to estimate $|f(\theta) - \mathcal{I}_N f(\theta)|$, but it tends to give looser bounds than bounding individual coefficients.

Chapter 3

Orthogonal polynomials: Approximation theory on intervals

3.1 ▪ From Fourier series to polynomial expansions

Fourier expansions provide the theoretical foundation for the approximation of smooth (more than just continuous) functions with polynomials. Define $\mathbb{I} := [-1, 1]$. For $f : \mathbb{I} \to \mathbb{C}$, set $g(\theta) = f(\cos\theta)$. Now, if f is continuous on $\mathbb{I}$, then g is continuous on $\mathbb{T}$.

Lemma 3.1. *Suppose $f : \mathbb{I} \to \mathbb{C}$ is continuous. Then for $g(\theta) = f(\cos\theta)$, for $M > 1$*

$$\mathcal{P}_{2M+1} g(\theta) = \frac{c_0(g)}{\sqrt{2\pi}} + \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{M} c_j(g)\cos(j\theta),$$

$$\mathcal{I}_{2M} g(\theta) = \frac{\check{c}_0(g)}{\sqrt{2\pi}} + \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{M-1} \check{c}_j(g)\cos(j\theta) + \frac{1}{\sqrt{2\pi}} \check{c}_M(g)\, \mathrm{e}^{\mathrm{i}M\theta}.$$

Proof. We simply compute

$$c_j(g) = \frac{1}{2\pi}\int_0^{2\pi} g(\theta)\, \mathrm{e}^{-\mathrm{i}j\theta}\, \mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} f(\cos\theta)[\cos(j\theta) - \mathrm{i}\sin(j\theta)]\mathrm{d}\theta$$
$$= \frac{1}{2\pi}\int_0^{2\pi} f(\cos\theta)\cos(j\theta)\mathrm{d}\theta = c_{-j}(f).$$

Then by the aliasing formula (2.1), supposing that g has an absolutely convergent Fourier series

$$\check{c}_{-j}(g) = \sum_{k=-\infty}^{\infty} c_{kN-j}(g) = \sum_{k=-\infty}^{\infty} c_{-kN-j}(g) = \sum_{k=-\infty}^{\infty} c_{kN+j}(g) = \check{c}_j(g).$$

Then the general case can be seen by approximating f, uniformly, with polynomials. This shows that

$$\mathcal{I}_{2M} g(\theta) = \frac{\check{c}_0(g)}{\sqrt{2\pi}} + \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{M-1} \check{c}_j(g)\cos(j\theta) + \frac{1}{\sqrt{2\pi}} \check{c}_M(g)\, \mathrm{e}^{\mathrm{i}M\theta}. \quad (3.1)$$

■

This allows us to define polynomial approximations to f. But we first note that in (3.1), if g is real-valued, we can obtain a better approximation by taking the real part of this expression.

Definition 3.2. *Suppose $f : \mathbb{I} \to \mathbb{C}$ is continuous. Then for $g(\theta) = f(\cos\theta)$, define*

$$\begin{aligned} \mathcal{P}_N^{\mathrm{T}} f(x) &= \mathcal{P}_{2N+1} g(\arccos x), \\ \mathcal{I}_N^{\mathrm{T}} f(x) &= \mathcal{I}_{2N} g(\arccos x) - \frac{\mathrm{i}}{\sqrt{2\pi}} \check{c}_N(g) \sin(N\theta), \end{aligned}$$

for $-1 \leq x \leq 1$.

Exercise 3.1. *Show that $T_k(x) = \cos(k \arccos x)$, for $k = 0, 1, 2, \ldots$ and $-1 \leq x \leq 1$ is a polynomial of degree k. T_k is the kth Chebyshev polynomial of the first kind.*

Exercise 3.2. *Verify that for both choices of μ in Theorem 2.12, if $g(\theta) = f(\cos\theta)$, for f continuous, and $N = 2M + 1$, then*

$$\tilde{\mathcal{P}}_N g(\arccos x), \quad -1 \leq x \leq 1,$$

is a polynomial in x.

3.2 ▪ Jackson's theorems for polynomials

We begin with a basic observation. Suppose that $f : \mathbb{I} \to \mathbb{C}$ is α-Hölder continuous with Hölder constant L. Then set $g(\theta) = f(\cos(\theta))$ which automatically extends periodically. Then

$$|g(\theta) - g(\phi)| = |f(\cos\theta) - f(\cos\phi)| \leq L|\cos\theta - \cos\phi|^\alpha.$$

But then, by the mean-value theorem $|\cos\theta - \cos\phi| \leq |\theta - \phi|$ and we find that g is α-Hölder continuous with the same parameters as f. Let P_N be the linear span of polynomials of degree at most N. The following is an immediate consequence of Jackson's first theorem (Theorem 2.13).

Theorem 3.3 (Jackson's third theorem). *Suppose that $f \in C^{0,\alpha}(\mathbb{I})$ with Hölder constant L. Then*

$$\inf_{g \in P_N} \|f - g\|_\infty \leq L \left(\frac{\pi^2}{4N + 6} \right)^\alpha.$$

While it does follow that if $f \in C^{k,\alpha}(\mathbb{I})$ then $g \in C^{k,\alpha}(\mathbb{T})$, the relationship between derivatives is more complicated. But we still can upgrade Jackson's third theorem.

Theorem 3.4 (Jackson's fourth theorem). *Suppose $f \in C^{k,\alpha}(\mathbb{I})$ and $f^{(k)}$ has α-Hölder constant L. Then for $N > k$*

$$\inf_{g \in P_N} \|f - g\|_\infty \leq \frac{L\pi^{2(k+\alpha)}}{(4N+6)(4N+4)\cdots(4N-4k+10)(4N-4k+6)^\alpha}.$$

Proof. By Theorem 3.3, for $N > k$,

$$\inf_{p \in P_{N-k}} \|f^{(k)} - p\|_\infty \leq L \left(\frac{\pi^2}{4N - 4k + 6} \right)^\alpha .$$

Following the proof of Theorem 2.16, for $p \in P_N$

$$\inf_{g \in P_N} \|f - g\|_\infty \leq \inf_{g \in P_N} \|f - p - g\|_\infty \leq \frac{\pi^2}{4N + 6} \|f' - p'\|_\infty.$$

Taking an infimum over p, we find

$$\inf_{g \in P_N} \|f - g\|_\infty \leq \frac{\pi^2}{4N + 6} \inf_{p \in P_{N-1}} \|f' - p\|_\infty.$$

Therefore

$$\inf_{g \in P_N} \|f - g\|_\infty \leq \frac{\pi^2}{4N + 6} \frac{\pi^2}{4N - 4 + 6} \inf_{p \in P_{N-2}} \|f'' - p\|_\infty.$$

And continuing, we establish the theorem. ■

3.3 ▪ Properties of Chebyshev expansions

In this section we analyze the operators $\mathcal{P}_N^{\mathrm{T}}$, $\mathcal{I}_N^{\mathrm{T}}$ in more detail.

3.3.1 ▪ Properties of $\mathcal{P}_N^{\mathrm{T}}$

We first recall that for $f : \mathbb{I} \to \mathbb{C}$, $g(\theta) = f(\cos\theta)$

$$c_0(g) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} g(\theta) \mathrm{d}\theta.$$

Set $x = \cos\theta$, then $\mathrm{d}x = -\sin\theta \mathrm{d}\theta$. We have $\sin\theta = (\operatorname{sign}\sin\theta)\sqrt{1 - x^2}$. Breaking up the integral as $\int_0^{2\pi} = \int_0^\pi + \int_\pi^{2\pi}$ one finds

$$c_0(g) = \sqrt{\frac{2}{\pi}} \int_{-1}^1 \frac{f(x)}{\sqrt{1 - x^2}} \mathrm{d}x. \tag{3.2}$$

Similarly, for $j > 0$

$$c_j(g) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}j\theta} g(\theta) \mathrm{d}\theta = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \cos(j\theta) g(\theta) \mathrm{d}\theta = \sqrt{\frac{2}{\pi}} \int_{-1}^1 T_j(x) \frac{f(x)}{\sqrt{1 - x^2}} \mathrm{d}x.$$

By choosing $f(x) = T_k(x)$ we see that $(T_j)_{j=0}^\infty$ forms an orthogonal system with respect to the inner product

$$\langle f, h \rangle_{\mathrm{T}} := \int_{-1}^1 f(x) \overline{h(x)} \mu_{\mathrm{T}}(\mathrm{d}x), \quad \mu_{\mathrm{T}}(\mathrm{d}x) = \frac{\mathbb{1}_{[-1,1]}(x)}{\pi\sqrt{1 - x^2}} \mathrm{d}x. \tag{3.3}$$

And on closer examination, we find that

$$1,\ \sqrt{2}T_1(x),\ \sqrt{2}T_2(x),\ \ldots,$$

is an orthonormal system. And now, we since we have seen that polynomials are dense in $C(\mathbb{I})$. And a key step in establishing the Lebesgue integral demonstrates that piecewise constant functions are dense in $L^2(\mu)$. By the monotone convergence theorem, a step function can be approximated by a continuous function. This implies this orthonormal system is indeed an orthonormal basis. This is the first example of a sequence of orthogonal polynomials that we will encounter.

Below, we will encounter other classes of "classical" orthogonal polynomials. Only in rare cases are these polynomials orthonormal. So, we use an _ to denote normalized polynomials. For example,

$$\underline{T}_j(x) = \begin{cases} 1 & j = 0, \\ \sqrt{2}T_j & j > 0. \end{cases}$$

And for these to be truly normalized, we need to specify a normalization for the weight function $\frac{1}{\pi\sqrt{1-x^2}}$. And for the underlined polynomials, we assume it to be a probability measure so that the first orthonormal polynomial is simply 1. And this results in

$$\mathcal{P}_N^{\mathrm{T}} f(x) = \sum_{j=0}^{N} \langle f, \underline{T}_j \rangle_T \underline{T}_j(x).$$

3.3.2 ▪ Properties of $\mathcal{I}_N^{\mathrm{T}}$

We know $\mathcal{I}_N g$ interpolates g. Because we removed a term, it might not seem immediate that $\mathcal{I}_N^{\mathrm{T}} f$ interpolates f at some points. Recall that

$$\mathcal{I}_{2N} g(\check{\theta}_\ell) = g(\check{\theta}_\ell), \quad \check{\theta}_\ell = 2\pi \frac{\ell - 1}{2N}, \quad \ell = 1, 2, \ldots, 2N.$$

But we note that $\sin(M\check{\theta}_\ell) = 0$ and therefore

$$\mathcal{I}_N^{\mathrm{T}} f(\check{x}_\ell) = f(\check{x}_\ell), \qquad \check{x}_\ell = \cos\left(2\pi \frac{\ell - 1}{2N}\right), \quad \ell = 1, 2, \ldots, 2N + 1.$$

Here we also note that $\ell = 1, 2, \cdots, N + 1$ is sufficient to enumerate the entire set.

And since degree N polynomial approximations at $N + 1$ points are unique, we see that $\mathcal{I}_N^{\mathrm{T}}$ is a projection.

Remark 3.5. *One could define $\mathcal{I}_N^{\mathrm{T}}$ in terms of $\mathcal{I}_{2N+1}$. This keeps more symmetry in the coefficients, but it loses two things:*

1. *The FFT is most efficient for vectors in a dimension that is a power of 2, or more specifically, when the dimension has a factorization using only small primes.*

2. *The interpolation points would no longer be symmetric about $x = 0$.*

3.3.3 ▪ Convergence

We equip the space of polynomials of degree N with the $\| \diamond \|_\infty$ norm on $\mathbb{I}$. We then see directly that

$$\begin{aligned}\|\mathcal{P}_N^{\mathrm{T}}\|_{\mathcal{C}(\mathbb{I})\to P_N} &= \|\mathcal{P}_{2N+1}\|_{\mathcal{C}(\mathbb{T})\to E_{2N+1}},\\ \|\mathcal{I}_N^{\mathrm{T}}\|_{\mathcal{C}(\mathbb{I})\to P_N} &\leq \|\mathcal{I}_{2N}\|_{\mathcal{C}(\mathbb{T})\to E_{2N}}.\end{aligned}$$

The following is immediate from Theorem 3.4:

Theorem 3.6. *Suppose $f \in C^{k,\alpha}(\mathbb{I})$ and $f^{(k)}$ has α-Hölder constant L. Then for $N > k \geq 0$, $N > 1$, there exists constants $C_{k,\alpha}, D_{k,\alpha}$ such that*

$$\begin{aligned}\|\mathcal{P}_N^{\mathrm{T}} f - f\|_\infty &\leq \frac{LC_{k,\alpha} \log N \pi^{2(k+\alpha)}}{(N+3)(N+2)\cdots(N-k+2)(N-k+3)^\alpha},\\ \|\mathcal{I}_N^{\mathrm{T}} f - f\|_\infty &\leq \frac{LD_{k,\alpha} \log N \pi^{2(k+\alpha)}}{(N+3)(N+2)\cdots(N-k+2)(N-k+3)^\alpha}.\end{aligned}$$

The next natural question is to ask how analyticity affects convergence. Here we only consider convergence in the $\| \diamond \|_\infty$ norm. We have seen that the convergence of Fourier series/interpolants is exponential if the function being approximated, when mapped to a function on $\mathbb{U}$, has an analytic extension that is analytic and bounded in an appropriate annulus $A_\rho(\mathbb{U})$. So, for $z = \mathrm{e}^{\mathrm{i}\theta}$, $\cos\theta = \frac{1}{2}(z + z^{-1})$. And, we need

$$h(z) := f\left(\frac{1}{2}(z + z^{-1})\right), \tag{3.4}$$

to be bounded and analytic in $A_\rho(\mathbb{U})$. But this raises the question, what is the image of $A_\rho(\mathbb{U})$ under the mapping $z \mapsto \frac{1}{2}(z + z^{-1})$.

Definition 3.7. *For $\rho > 1$ set*

$$A_\rho(\mathbb{I}) = \left\{\frac{1}{2}(z + z^{-1}) \mid z \in A_\rho(\mathbb{U})\right\}.$$

This is the Bernstein ellipse with parameter ρ. It is an ellipse with horizontal major axis, centered at the origin containing $\mathbb{I}$. The space of functions that are bounded and analytic in $A_\rho(\mathbb{I})$ is denoted by $B_\rho(\mathbb{I})$ with the usual $\| \diamond \|_\infty$ norm.

The following is immediate.

Theorem 3.8. *Suppose $f \in B_\rho(\mathbb{I})$, $\rho > 1$, with $M = \sup_{z \in B_\rho} |f(z)|$. Then on $\mathbb{I}$*

$$\|f - \mathcal{P}_N^{\mathrm{T}} f\|_\infty \leq \frac{2M\rho^{-N}}{1 - 1/\rho},$$

and

$$\|f - \mathcal{I}_N^{\mathrm{T}} f\|_\infty \leq \frac{4M\rho^{-N+1}}{1 - 1/\rho}.$$

3.4 ▪ General theory of orthogonal polynomials

In looking at the interpolation points from the previous section $\check{x}_\ell = \cos\left(\pi\frac{\ell-1}{N}\right)$, $\ell = 1, 2, \ldots, N+1$, we can see that they are the extrema of T_N on $\mathbb{I}$. Alternatively, we could have used

$$2\pi\frac{\ell - 1/2}{N},$$

to compute approximate Fourier coefficients, giving a new set of interpolation points. This would have resulted in new interpolation points on $\mathbb{I}$:

$$\cos\left(\pi\frac{\ell - 1/2}{N}\right),$$

which can be seen to be the roots of T_N. The convergence theory is the same in either case, and in situations where one wants to avoid evaluating the function being approximated at ±1, due to, for example, the cancelation of singularities, this can be desirable. Furthermore, the use of the zeros of orthogonal polynomials as grids for interpolation, integration and differentiation is of great use, far beyond Chebyshev polynomials of the first kind.

We now introduce the theory of orthogonal polynomials on $\mathbb{I}$. The general theory of orthogonal polynomials is far too wide-ranging for us to include in a reasonable number of pages. Interested readers are first referred to [7] and [19]. As we have seen, since polynomials are dense in $L^2(\mu)$ one can perform the Gram-Schmidt process on the monomials $\{1, \diamond, \diamond^2, \ldots\}$ using the inner product

$$\langle f, g\rangle_\mu := \int_{\mathbb{R}} f(x)\overline{g(x)}\mu(\mathrm{d}x),$$

to obtain an orthonormal basis for $L^2(\mu)$. Often this process is described by first constructing the monic orthogonal basis $(\pi_k(\diamond;\mu))_{k\geq 0}$ satisfying

- $\pi_k(x;\mu) = x^k + O(x^{k-1}), \quad x \to \infty$ and
- $\int \pi_k(x;\mu)\pi_j(k;\mu)\mu(\mathrm{d}x) = 0$ for $j \neq k$.

Lemma 3.9. *Suppose the measure μ satisfies (1) $\int |x|^k \mu(\mathrm{d}x) < \infty$ for all $k \geq 0$ and (2) the support of μ contains an infinite number of points. Then the monic polynomials $\pi_k(x;\mu)$ exist for all $k \geq 0$. If the support only contains n distinct points then $\pi_k(x;\mu)$ exist only for $0 \leq k \leq n-1$.*

Proof. The sequence can only fail to exist if $x^k \in \operatorname{span}\{\pi_0, \ldots, \pi_{k-1}\}$ in $L^2(\mu)$. This can clearly happen if μ is discrete with delta masses at a finite number of points. Next, by Theorem 1.39, there is a unique choice of coefficients a_j that minimize

$$\left\| \diamond^k - \sum_{j=0}^{k-1} a_j \pi_j \right\|_{L^2(\mu)}.$$

But with the supposition that μ has an infinite number of points in its support, there exists $y \in \mathbb{R}$, in the support of μ, such that $E(y) := y^k - \sum_{j=0}^{k-1} a_j\pi_j(y) \neq 0$. A point in the support of a measure is characterized by the fact that every open neighborhood of it has positive measure. So, we can choose a ball $B_\epsilon(y)$ such that $|E(y)| \geq \delta$ on $B_\epsilon(y)$ for some $\delta > 0$, by continuity. Then

$$\|E\|_{L^2(\mu)}^2 \geq \int_{B_\epsilon(y)} |E(y)|^2 \mu(\mathrm{d}y) \geq \delta^2 \mu(B_\epsilon(y)) > 0.$$

Thus $x^k \not\in \operatorname{span}\{\pi_0, \ldots, \pi_{\mathrm{k}-1}\}$ in $L^2(\mu)$ and the monic polynomials exist. ■

Then the orthonormal polynomials p_k are simply defined by

$$p_k(x;\mu) = \frac{\pi_k(x;\mu)}{\|\pi_k\|_{L^2(\mu)}}, \quad k = 0, 1, 2, \ldots.$$

We now identify a number of useful properties of orthogonal polynomials.

3.4.1 ▪ The three-term recurrence

Arguably the most fundamental aspect of orthogonal polynomials is the three-term recurrence that they satisfy.

Theorem 3.10. *Let μ be a Borel measure on $\mathbb{R}$, with an infinite number of points in its support.*

(a) Suppose $(q_j)_{j\geq 0}$ is a sequence of orthogonal polynomials for μ. Then there exists constants A_j, B_j, C_j such that

$$xq_j(x) = A_j q_j(x) + B_j q_{j-1}(x) + C_j q_{j+1}(x), \quad j \geq 0,$$

with the convention $q_{-1} \equiv 0$.

(b) If the polynomials $q_j = p_j$ are orthonormal then $C_j = B_{j+1}$. In this case we write,

$$xp_j(x;\mu) = a_j(\mu)p_j(x;\mu) + b_{j-1}(\mu)p_{j-1}(x;\mu) + b_j(\mu)p_{j+1}(x;\mu), \quad j \geq 0,$$

for sequence $a_j(\mu), b_j(\mu)$, $j \geq 0$, $b_j \neq 0$.

(c) If the polynomials are constructed by normalizing the monic polynomials using positive coefficients, then $b_j(\mu) > 0$ for all j.

Proof. The proof of (a) is left as an exercise. To prove (b) we note that

$$B_j = \langle \diamond q_j(\diamond), q_{j-1}\rangle_\mu,$$
$$C_j = \langle \diamond q_j(\diamond), q_{j+1}\rangle_\mu.$$

And then we just have to note that the polynomials are real-valued. Part (c) follows from examining the leading coefficients of the polynomials. ■

Exercise 3.3. *Establish Theorem 3.10(a).*

The recurrence coefficients are often arranged in a Jacobi operator.

Definition 3.11. *A Jacobi operator is a semi-infinite matrix*

$$\mathcal{J} = \begin{bmatrix} a_0 & b_0 & & \\ b_0 & a_1 & b_1 & \\ & b_1 & a_2 & \ddots \\ & & \ddots & \ddots \end{bmatrix},$$

where $b_j > 0$ for $j \geq 0$. Finite truncations are referred to as Jacobi matrices:

$$\mathbf{J}_N := \mathcal{J}_{1:N,1:N} = \begin{bmatrix} a_0 & b_0 & & & \\ b_0 & a_1 & b_1 & & \\ & b_1 & a_2 & \ddots & \\ & & \ddots & \ddots & b_{N-2} \\ & & & b_{N-2} & a_{N-1} \end{bmatrix}.$$

Then we define the Jacobi operator associated to a measure.

Definition 3.12. *Let μ be a Borel measure on $\mathbb{R}$, with an infinite number of points in its support. Then define*

$$\mathcal{J}(\mu) = \begin{bmatrix} a_0(\mu) & b_0(\mu) & & \\ b_0(\mu) & a_1(\mu) & b_1(\mu) & \\ & b_1(\mu) & a_2(\mu) & \ddots \\ & & \ddots & \ddots \end{bmatrix}.$$

3.4.2 ▪ Gaussian quadrature

We now include a brief discussion of the development of quadrature rules. A quadrature rule on $\mathbb{R}$ consists of a set of nodes $x_1 < x_2 < \cdots < x_n$ and weights w_j, $j = 1, 2, \ldots, N$ such that, informally,

$$\int_{\mathbb{R}} f(x)\mu(\mathrm{d}x) \approx \sum_j w_j f(x_j).$$

We write

$$E_N(f) = E_N(f; (x_j), (w_j)) = \int_{\mathbb{R}} f(x)\mu(\mathrm{d}x) - \sum_j w_j f(x_j).$$

A quadrature formula is said to have degree of exactness d if

$$E_N(p) = 0, \quad \forall p \in P_d = \mathrm{span}\{1, \diamond, \ldots, \diamond^{\mathrm{d}}\}.$$

A quadrature rule is said to be *interpolatory* if it has degree of exactness at least $d = N - 1$ but we will do much better than that.

While there are many ways to motivate the following, the definition of a Gaussian quadrature rule for a measure μ comes from the following observation from inverse spectral theory. For convenience, suppose that μ has compact support, then

$$\int_{\mathbb{R}} \frac{\mu(\mathrm{d}x)}{x - z} = \mathbf{e}_1^T (\mathcal{J}(\mu) - z)^{-1} \mathbf{e}_1.$$

If we instead considered a finite truncation $\mathbf{J}_N$, using its eigenvalue decomposition

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{bmatrix}, \quad \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N),$$

we find

$$\mathbf{e}_0^T (\mathbf{J}_N - z)^{-1} \mathbf{e}_0 = \sum_{j=1}^{N} \frac{w_j}{\lambda_j - z}, \quad w_j = |\mathbf{u}_{1j}|^2.$$

We recognize the latter as

$$\sum_{j=1}^{N} \frac{w_j}{\lambda_j - z} = \int_{\mathbb{R}} \frac{\mu_k(\mathrm{d}x)}{x - z}, \quad \mu_N = \sum_{j=1}^{N} |u_{1j}|^2 \delta_{\lambda_j},$$

where u_{ij} is the (i,j) entry of $\mathbf{U}$. This leads us to investigate properties of the quadrature rule defined by the eigenvalues and squared-modulii of the first components of the normalized eigenvectors of $\mathbf{J}_N$. We call this μ_N the Nth-order Gaussian quadrature rule for μ.

Theorem 3.13. *A Gaussian quadrature rule μ_N for a probability measure μ, with compact support*[a]*, has degree of exactness $2N-1$.*

[a] This theorem still holds for many measures without compact support but many technicalities arise.

Proof. Using the partial sum of the geometric series one can see that both of these functions have Laurent series at infinity and equating coefficients (and using Neumann series **??**) we see that for all $j \in \mathbb{N}$

$$\int_{\mathbb{R}} x^j \mu(\mathrm{d}x) = \mathbf{e}_1^T \mathcal{J}(\mu)^j \mathbf{e}_1, \quad \int_{\mathbb{R}} x^j \mu_k(\mathrm{d}x) = \mathbf{e}_1^T \mathbf{J}_N^j \mathbf{e}_1.$$

So, it remains to show that $\mathbf{e}_1^T \mathcal{J}(\mu)^j \mathbf{e}_1 = \mathbf{e}_1^T \mathbf{J}_N^j \mathbf{e}_1$ for $j = 1, 2, \ldots, 2N-1$. One way to see this is that multiplication by a tridiagonal matrix acts like a diffusion. If a vector is non-zero at one entry, then one should expect it to also be non-zero at neighboring entries after multiplications. So, one can see that the non-zero entries of $\mathcal{J}(\mu)^j \mathbf{e}_1$ and $\mathbf{J}_N^j \mathbf{e}_1$ will agree for $j \leq N-1$. The first N entries will agree for $j \leq N$, but $\mathcal{J}(\mu)^N \mathbf{e}_1$ will, in general, have a non-zero entry a position $k+1$. This is the first error introduced. Then it takes a remaining $N-1$ multiplications for this to propagate back to the first entry — the entry we care about, hence $2N-1$. ∎

3.4.3 ▪ Interpolation

Given a measure μ with $\operatorname{supp}(\mu) = \mathbb{I}$, and its Jacobi operator $\mathcal{J}(\mu)$, the Gaussian quadrature rules associated to it provide a natural way to discretize the inner product $\langle \triangleleft, \triangleright \rangle_\mu$. So, define

$$\langle f, g \rangle_{\mu,N} = \int_{\mathbb{R}} f(x)\overline{g(x)} \mu_N(\mathrm{d}x) = \sum_{j=1}^{N} f(\lambda_j)\overline{g(\lambda_j)} w_j.$$

To ensure that this makes sense we need the following lemma.

Lemma 3.14. *The zeros of $p_j(x;\mu)$ for $j \geq 1$ lie in the smallest closed interval that contains the support of μ.*

Proof. Let $[a,b]$ denote the smallest interval that contains the support of μ. Now suppose that p_j has a root outside $[a,b]$. Let $y_1,\ldots,y_m$ be the points within $[a,b]$ at which p_j changes sign. Necessarily, $m < j$. Then it follows that

$$p_j(x;\mu)\prod_{k=1}^{m}(x-y_k),$$

is either strictly positive or strictly negative on $[a,b]$. If $m=0$, then the empty product is taken to be 1. Thus

$$0 \neq \int_{\mathbb{R}} p_j(x;\mu)\prod_{k=1}^{m}(x-y_k)\mu(\mathrm{d}x).$$

But this is a contradiction as p_j is in the orthogonal complement of the span of lower degree polynomials. The conclusion follows. ■

Thus to discuss the interpolation of a function using roots of orthogonal polynomials, we will need to require that the function is defined on the smallest closed interval that contains the support of μ (the convex hull of the support). We find direct parallels to the discrete Fourier transform. Recall that for $f \in L^2(\mu)$ we have the convergent expansion

$$f(x) = \sum_{j=0}^{\infty} \langle f, p_j(\diamond;\mu)\rangle_\mu \, p_j(x;\mu).$$

So, suggestively define

$$\mathcal{I}_N^\mu f(x) = \sum_{j=0}^{N-1} \langle f, p_j(\diamond;\mu)\rangle_{\mu,N} \, p_j(x;\mu).$$

At this point, it may not be clear how to even compute with such a representation. For now, we establish the following:

Theorem 3.15. *Consider a probability measure μ with* $\operatorname{supp}(\mu) = [-1,1]$ *and its Jacobi operator $\mathcal{J}(\mu)$, let λ_j, $j = 1,2,\ldots,N$ denote the eigenvalues of $\mathbf{J}_N$. Then*

$$\mathcal{I}_N^\mu f(\lambda_j) = f(\lambda_j), \quad j = 1,2,\ldots,N.$$

Proof. We first have to establish a connection between the polynomials themselves and the weights in the quadrature rule. Note that the three-term recurrence implies that

$$\mathcal{J}(\mu)\begin{bmatrix} p_0(x;\mu) \\ p_1(x;\mu) \\ \vdots \end{bmatrix} = x\begin{bmatrix} p_0(x;\mu) \\ p_1(x;\mu) \\ \vdots \end{bmatrix},$$

which resembles an eigenvalue equation. It follows that $p_1, \ldots, p_N$ satisfy the relation

$$\mathbf{J}_N \begin{bmatrix} p_0(x;\mu) \\ p_1(x;\mu) \\ \vdots \\ p_{N-1}(x;\mu) \end{bmatrix} = x \begin{bmatrix} p_0(x;\mu) \\ p_1(x;\mu) \\ \vdots \\ p_{N-1}(x;\mu) \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ b_{N-1} p_N(x;\mu) \end{bmatrix}.$$

From this it follows that a root of $p_N(x;\mu)$ is an eigenvalue of $\mathbf{J}_N$. If p_N has distinct roots, then this is all the eigenvalues. Suppose that p_N has a repeated root. We first obtain $\mathbf{J}_N \mathbf{v} = \lambda \mathbf{v}$. Then this relation can be differentiated with respect to x to obtain a generalized eigenvector $\mathbf{w}$ satisfying $\mathbf{J}_N \mathbf{w} = \lambda \mathbf{w} + \mathbf{v}$. Then from $\langle \mathbf{J}_N \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{J}_N \mathbf{w} \rangle$ we obtain the contradiction that $\mathbf{v} = \mathbf{0}$. Therefore, the orthogonal matrix of eigenvectors is given by

$$\mathbf{U} = \underbrace{\begin{array}{c} \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & p_j(\lambda_\ell;\mu) & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix}_{j=N-1}^{0} \\ \begin{array}{ccc} \ell=1 & \qquad\qquad & N \end{array} \end{array}}_{\mathbf{P}} \mathbf{D}, \tag{3.5}$$

where $\mathbf{D}$ is chosen to normalize the columns. Since $p_0 = 1$

$$w_j = \left(\sum_{\ell=0}^{N-1} p_\ell(\lambda_j;\mu)^2 \right)^{-1}, \quad \mathbf{D} = \operatorname{diag}(\sqrt{w}_1, \sqrt{w}_2, \ldots, \sqrt{w}_N). \tag{3.6}$$

Upon setting $c_j = \langle f, p_j(\diamond;\mu) \rangle_{\mu,N}$, we find

$$\begin{bmatrix} \\ c_j \\ \\ \end{bmatrix} = \mathbf{P}\mathbf{D}^2 \begin{bmatrix} \\ f(\lambda_j) \\ \\ \end{bmatrix} = \mathbf{U}\mathbf{D} \begin{bmatrix} \\ f(\lambda_j) \\ \\ \end{bmatrix}.$$

Then because $\mathbf{U}$ must be orthogonal, we find

$$\begin{bmatrix} \\ f(\lambda_j) \\ \\ \end{bmatrix} = \mathbf{D}^{-1}\mathbf{U}^T \begin{bmatrix} \\ c_j \\ \\ \end{bmatrix} = \mathbf{P}^T \begin{bmatrix} \\ c_j \\ \\ \end{bmatrix} = \begin{array}{c} \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & p_j(\lambda_\ell;\mu) & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix}_{\ell=N}^{1} \\ \begin{array}{ccc} j=0 & \qquad\qquad & N-1 \end{array} \end{array} \begin{bmatrix} \\ c_j \\ \\ \end{bmatrix}.$$

The right-hand side of this is precisely the vector made up of the evaluations of $\mathcal{I}_N^\mu f$ at the λ_j's. ■

3.4.4 ▪ From Jacobi matrix to transform

The proof of the previous theorem actually gives access to a reasonable computation method for computing the coefficients c_j in the interpolant $\mathcal{I}_N^\mu f$. The first thing to note is that symmetric, tridiagonal eigenvalue problems are, apart from diagonal ones, the easiest eigenvalues problems to solve numerically. So, computing $\mathbf{U}$ in (3.5) can be considered a solved problem (Wilkinson shifts!), and in particular, computing the Gaussian quadrature rule is straightforward. Then all one needs to compute is the diagonal matrix $\mathbf{D}$, but that just contains the quadrature weights.

3.4.5 ▪ The evaluation of the interpolant

The simplest algorithm to evaluate $\mathcal{I}_N^\mu f$ at a point x is to run the three-term recurrence and use it to compute the polynomials using

$$p_{j+1}(x;\mu) = \frac{1}{b_j}\left[(x-a_j)p_j(x;\mu) - b_{j-1}p_{j-1}(x;\mu)\right].$$

and then to simply sum the series

$$\sum_{j=0}^{N-1} c_j p_j(x;\mu).$$

Each step of the iteration requires 5 FLOPs (floating point operations) giving $5(N-1)$ total FLOPs to run the recurrence out to p_{N-1}. And then evaluating the sum requires $2N-1$ FLOPs. So, evaluation at a point with this method requires $7(N-1)+1$ FLOPs.

This can be improved using what is known as the Clenshaw algorithm. We use the linear system perspective from [14]. The observation is that the orthogonal polynomials satisfy the lower-triangular linear system

$$\underbrace{\begin{bmatrix} 1 & & & & \\ a_0 - x & b_0 & & & \\ b_0 & a_1 - x & b_1 & & \\ & b_1 & a_2 - x & b_2 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}}_{\mathbf{L}(x)} \begin{bmatrix} p_0(x;\mu) \\ p_1(x;\mu) \\ p_2(x;\mu) \\ \vdots \end{bmatrix} = \mathbf{e}_1.$$

And as such a system is solved by forward substitution, we can truncate this system to any square size to obtain

$$\underbrace{\begin{bmatrix} 1 & & & & & \\ a_0 - x & b_0 & & & & \\ b_0 & a_1 - x & b_1 & & & \\ & b_1 & a_2 - x & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{N-3} & a_{N-2} - x & b_{N-2} \end{bmatrix}}_{\mathbf{L}_N(x)} \begin{bmatrix} p_0(x;\mu) \\ p_1(x;\mu) \\ p_2(x;\mu) \\ \vdots \\ p_{N-1}(x;\mu) \end{bmatrix} = \mathbf{e}_1.$$

If $\mathbf{c} = (c_j)_{j\geq 0}$ is the vector of N coefficients, in terms of what is above, we want to compute

$$\begin{bmatrix} p_0(x;\mu) & p_1(x;\mu) & \cdots & p_{N-1}(x;\mu) \end{bmatrix} \mathbf{c} = (\mathbf{L}_N(x)^{-1}\mathbf{e}_1)^T\mathbf{c} = \mathbf{e}_1^T\mathbf{L}_N(x)^{-T}\mathbf{c}.$$

This is the essence of the Clenshaw algorithm — solving the upper-triangular system $\mathbf{L}_N(x)^T\mathbf{s} = \mathbf{c}$ and then retaining just the first entry. Each step in the backward substitution used to solve this linear system requires no more than 5 FLOPs and solving this system is no more expensive than evaluating $p_{N-1}(x;\mu)$.

3.5 ▪ Connection coefficients

Consider two sequences of orthogonal polynomials: $(p_j(x;\mu))_{j\geq 0}$ and $(p_j(x;\nu))_{j\geq 0}$. Since for each N, the polynomials through degree N are linearly independent and span the same finite-dimensional space. So, there exists an upper-triangular change-of-basis matrix such that

$$\begin{bmatrix} p_0(x;\mu) & p_1(x;\mu) & \cdots \end{bmatrix} = \begin{bmatrix} p_0(x;\nu) & p_1(x;\nu) & \cdots \end{bmatrix} \underbrace{\begin{bmatrix} u_{00} & u_{10} & \cdots & \\ & u_{11} & u_{12} & \cdots \\ & & \ddots & \ddots \end{bmatrix}}_{\mathcal{U}_{\nu\to\mu}}.$$

The elements of the semi-infinite, upper-triangular matrix $U_{\nu\to\mu}$ are referred to as the connection coefficients. By orthogonality, we have

$$\begin{aligned} p_j(x;\mu) &= \sum_{k=0}^{j} u_{jk} p_k(x;\nu), \\ u_{jk} &= \int_{\mathbb{R}} p_j(x;\mu) p_k(x;\mu) \nu(\mathrm{d}x). \end{aligned}$$

For specific choices of measures μ, ν this matrix $\mathcal{U}_{\nu\to\mu}$ can be sparse, and in particular, banded with a very small bandwidth. This is discussed more below.

The matrix of functions

$$\begin{bmatrix} p_0(x;\mu) & p_1(x;\mu) & \cdots \end{bmatrix}$$

is sometimes called a quasi-matrix.

3.6 ▪ Jacobi polynomials

In this section we highlight properties of the orthogonal polynomials with respect to the two-parameter family of weight functions

$$w_{\alpha,\beta}(x) := Z_{\alpha,\beta}^{-1}(1-x)^\alpha(1+x)^\beta \mathbb{1}_{[-1,1]}(x), \quad \alpha,\beta > -1. \tag{3.7}$$

Here $Z_{\alpha,\beta}$ is the normalization constant so that

$$\mu(\mathrm{d}x) = w_{\alpha,\beta}(x)\mathrm{d}x,$$

is a probability measure on $\mathbb{R}$. It can be computed as

$$Z_{\alpha,\beta} = \frac{{}_2F_1(1,-\alpha,2+\beta,-1)}{1+\beta} + \frac{{}_2F_1(1,-\beta,2+\alpha,-1)}{1+\alpha},$$

in terms of the hypergeometric function ${}_2F_1$ [13]. We abuse notation and use $p_j(x;\alpha,\beta)$ to refer to the jth orthonormal polynomial. This is referred to as an orthonormal Jacobi polynomial. The classical notation [13] is for unnormalized, and not monic, Jacobi polynomials is $P_j^{(\alpha,\beta)}(x)$ such that

$$m_j(\alpha,\beta) := \int_{-1}^{1} P_j^{(\alpha,\beta)}(x)^2(1-x)^\alpha(1+x)^\beta \mathrm{d}x = \frac{2^{\alpha+\beta+1}\Gamma(j+\alpha+1)\Gamma(j+\beta+1)}{(2j+\alpha+\beta+1)\Gamma(j+\alpha+\beta+1)j!},$$

where $\Gamma(\diamond)$ is the Gamma function [13]. Set $d_j = d_j(\alpha,\beta) = 2j+\alpha+\beta$. The polynomials satisfy the three-term recurrence relation

$$\begin{aligned} &2j(j+\alpha+\beta)(d_j-2)P_j^{(\alpha,\beta)}(x) \\ &\quad = (d_j-1)\left[d_j(d_j-2)x+\alpha^2-\beta^2\right]P_{j-1}^{(\alpha,\beta)}(x) - 2d_j(d_j-\beta-1)(d_j-\alpha-1)P_{j-2}^{(\alpha,\beta)}(x). \end{aligned}$$

3.6.1 ▪ The Jacobi operator

We wish to find the associated Jacobi matrix. Divide by $d_j(d_j-1)(d_j-2)m_{j-1}^{1/2}$ to find

$$\frac{2j(j+\alpha+\beta)\sqrt{m_j}}{d_j(d_j-1)\sqrt{m_{j-1}}}p_j(x;\alpha,\beta) = \left[x + \frac{\alpha^2-\beta^2}{d_j(d_j-2)}\right]p_{j-1}(x;\alpha,\beta) \\ - \frac{2(j+\beta-1)(j+\alpha-1)\sqrt{m_{j-2}}}{(d_j-1)(d_j-2)\sqrt{m_{j-1}}}p_{j-2}(x;\alpha,\beta).$$

This can be further simplified using

$$\begin{aligned}\frac{m_n}{m_{n-1}} &= \frac{2n+\alpha+\beta-1}{2n+\alpha+\beta+1}\frac{(n-1)!}{n!}\frac{\Gamma(n+\alpha+1)}{\Gamma(n+\alpha)}\frac{\Gamma(n+\beta+1)}{\Gamma(n+\beta)}\frac{\Gamma(n+\alpha+\beta)}{\Gamma(n+\alpha+\beta+1)}\\ &= \frac{2n+\alpha+\beta-1}{2n+\alpha+\beta+1}\frac{(n+\alpha)(n+\beta)}{n(n+\alpha+\beta)} = \frac{d_n-1}{d_n+1}\frac{(n+\alpha)(n+\beta)}{n(n+\alpha+\beta)}\end{aligned}$$

Define

$$b_{j-1} = \frac{2\sqrt{j}\sqrt{(j+\alpha)(j+\beta)}\sqrt{j+\alpha+\beta}}{d_j\sqrt{d_j^2-1}},$$

$$a_{j-1} = \frac{\beta^2-\alpha^2}{d_j(d_j-2)}.$$

Note that

$$\frac{2j(j+\alpha+\beta)\sqrt{m_j}}{d_j(d_j-1)\sqrt{m_{j-1}}} = b_{j-1}$$

and

$$\begin{aligned}\frac{2(j+\beta-1)(j+\alpha-1)\sqrt{m_{j-2}}}{(d_j-1)(d_j-2)\sqrt{m_{j-1}}} &= \frac{2(j+\beta-1)(j+\alpha-1)}{(d_j-1)(d_j-2)}\frac{\sqrt{d_j-1}}{\sqrt{d_j-3}}\frac{\sqrt{j-1}\sqrt{j+\alpha+\beta-1}}{\sqrt{j+\alpha-1}\sqrt{j+\beta-1}}\\ &= \frac{2\sqrt{j+\beta-1}\sqrt{j+\alpha-1}\sqrt{j-1}\sqrt{j+\alpha+\beta-1}}{(d_j-2)\sqrt{d_j-1}\sqrt{d_j-3}} = b_{j-2},\end{aligned}$$

and therefore we have determined the three-term recurrence coefficients for normalized Jacobi polynomials.

3.6.2 ▪ Derivatives

We also want to understand how taking a derivative transforms Jacobi polynomials. To do this, consider the monic polynomials $\pi_j(x;\alpha,\beta)$ with respect to the Jacobi measure (3.7). Then, for $j > k$, by integration by parts, using that π_k is orthogonal to all lower degree polynomials

$$\int_{\mathbb{R}} \pi_j'(x;\alpha,\beta)\pi_k'(x;\alpha,\beta)(1-x)^{\alpha+1}(1+x)^{\beta+1}\mathrm{d}x = 0.$$

Upon examining the leading coefficient, we have

$$\pi_j'(x;\alpha,\beta) = j\pi_{j-1}(x;\alpha+1,\beta+1).$$

Similarly, there must exist $c_j = c_j(\alpha,\beta)$ such that

$$p_j'(x;\alpha,\beta) = c_j p_{j-1}(x;\alpha+1,\beta+1).$$

And if $h_j(\alpha,\beta)$ is such that

$$p_j(x;\alpha,\beta) = h_j(\alpha,\beta)\pi_j(x;\alpha,\beta),$$

then

$$c_j(\alpha,\beta) = j\frac{h_j(\alpha,\beta)}{h_{j-1}(\alpha+1,\beta+1)}.$$

3.6.3 ▪ More on Chebyshev polynomials of the first kind

The Chebyshev polynomials T_j of the first kind, orthogonal with respect to μ_{T}, can be obtained by setting $\alpha = \beta = -1/2$. This gives

$$\begin{aligned} a_j &= 0, \quad j \geq 0,\\ b_1 &= \frac{1}{\sqrt{2}},\\ b_j &= \frac{1}{2}, \quad j \geq 1. \end{aligned}$$

We immediately see that

$$x_j = \cos\left(\pi\frac{j-1/2}{N}\right), \quad j = 1,2,\ldots,N,$$

give the roots of T_N and hence give the Guassian quadrature nodes. Then, in light of (3.6), we see that

$$T_0(x_j)^2 + 2\sum_{k=1}^{N-1} T_k(x_j)^2 = 1 + 2\sum_{k=1}^{N-1}\cos^2\left(k\pi\frac{j-1/2}{N}\right) = N + \sum_{k=1}^{N-1}\cos\left(k\pi\frac{2j-1}{N}\right).$$

Here we recall that $p_0(x;-1/2,-1/2) = T_0(x)$, $p_j(x;-1/2,-1/2) = \sqrt{2}T_j(x)$, $j \geq 1$. For this last sum we recognize, upon setting $z = \exp\left(\pi\frac{2j-1}{N}\right)$,

$$\sum_{k=1}^{N-1}\cos\left(2k\pi\frac{j-1/2}{N}\right) = \operatorname{Re}\sum_{k=1}^{N-1} z^k = \operatorname{Re}\frac{1-z^N}{1-z} - 1,$$

and

$$\operatorname{Re}\frac{1-z^N}{1-z} = 2\operatorname{Re}\frac{1}{1-z} = 2\frac{\operatorname{Re} z - |z|^2}{|1-z|^2} = 1.$$

This gives the following.

Proposition 3.16. *The Nth Gaussian quadrature rule for μ_{T} is given by*

$$\frac{1}{N}\sum_{j=1}^{N}\delta_{x_j}, \quad x_j = \cos\left(\pi\frac{j-1/2}{N}\right), \quad j = 1,2,\ldots,N.$$

3.6.4 ▪ Chebyshev polynomials of the second kind

The Chebyshev polynomials of the second kind, denoted by U_j, are found by setting $\alpha = \beta = 1/2$. We find

$$\begin{aligned} a_j &= 0, \quad j \geq 0, \\ b_j &= \frac{1}{2}, \quad j \geq 0. \end{aligned}$$

And we find that

$$U_j(\cos\theta) = \frac{\sin(j+1)\theta}{\sin\theta}. \tag{3.8}$$

With this normalization, these polynomials form an orthonormal system with respect to the inner product with (probability) weight

$$\mu_{\mathrm{U}}(\mathrm{d}x) = \frac{1}{2\pi}\sqrt{1-x^2}\mathbb{1}_{[-1,1]}(x)\mathrm{d}x.$$

Exercise 3.4. *Establish* (3.8)*.*

And we have the following,

Proposition 3.17. *The Nth Gaussian quadrature rule for μ_{U} is given by*

$$\sum_{j=1}^{N} w_j \delta_{x_j}, \quad x_j = \cos\left(\frac{j}{N+1}\pi\right), \quad w_j = \frac{2}{N+1}\sin^2\left(\frac{i}{N+1}\pi\right), \quad j = 1, 2, \ldots, N.$$

Exercise 3.5. *Prove* (3.17)*.*

3.6.5 ▪ Chebyshev polynomials of the third kind

The Chebyshev polynomials of the third kind, denoted by V_j, are orthonormal polynomials and are found by setting $\alpha = -\beta = -1/2$. This gives

$$\begin{aligned} a_0 &= \frac{1}{2}, \\ a_j &= 0 \quad j \geq 1, \\ b_j &= \frac{1}{2}, \quad j \geq 0. \end{aligned}$$

3.6.6 ▪ Chebyshev polynomials of the fourth kind

The Chebyshev polynomials of the fourth kind, denoted by W_j, are orthonormal polynomials and are found by setting $\alpha = -\beta = 1/2$. This gives

$$\begin{aligned} a_0 &= -\frac{1}{2}, \\ a_j &= 0 \quad j \geq 1, \\ b_j &= \frac{1}{2}, \quad j \geq 0. \end{aligned}$$

3.6.7 ▪ Legendre polynomials

Legendre polynomials are found by setting $\alpha = \beta = 0$ and are denoted by $P_j(x)$. With the classical definition, they are not orthonormal. They are normalized so that

$$\int_{-1}^{1} P_j(x)^2 \mathrm{d}x = \frac{2}{2j+1}.$$

Therefore

$$p_j(x;0,0) = \frac{1}{\sqrt{2j+1}} P_j(x).$$

This gives the recurrence coefficients

$$\begin{aligned} a_j &= 0 \quad j \geq 0, \\ b_j &= \frac{j+1}{\sqrt{(2j+2)(2j+3)}}, \quad j \geq 0. \end{aligned}$$

3.6.8 ▪ Ultraspherical polynomials

The ultraspherical (Gegenbauer) polynomials are found by setting $\alpha = \beta = \lambda - 1/2$ for $\lambda > -1/2$ and denoted by $C_j^{(\lambda)}(x)$. These are also not, classically, orthonormal. They satisfy

$$\int_{-1}^{1} C_j^{(\lambda)}(x)^2 (1-x^2)^{\lambda - 1/2} \mathrm{d}x = \frac{2^{1-2\lambda}\pi\Gamma(j+2\lambda)}{(j+\lambda)\Gamma(\lambda)^2 j!},$$

and

$$C_j^{(\lambda)}(x) = \frac{2^j (\lambda)_j}{j!} x^j + O(x^{j-1}),$$

where

$$(a)_j = \frac{\Gamma(a+j)}{\Gamma(a)}, \tag{3.9}$$

is the Pochhammer symbol.

3.6.9 ▪ Derivatives

We obtain the relationship between derivatives, with the convention $C_{-1}^{(\lambda)} = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}x} C_j^{(\lambda)}(x) = j \frac{2^j (\lambda)_j}{j!} \frac{(j-1)!}{2^{j-1}(\lambda+1)_{j-1}} C_{j-1}^{(\lambda+1)}(x) = 2\lambda C_{j-1}^{(\lambda+1)}(x).$$

3.6.10 ▪ Connection coefficients

We see from above that differentiation is sparse mapping $C_j^{(\lambda)}$ to a multiple of $C_{j-1}^{(\lambda+1)}$. And fortunately, the basis conversion from $(C_j^{(\lambda)})_{j\geq 0}$ to $(C_j^{(\lambda+1)})_{j\geq 0}$ is also sparse. Let $\mathbf{c}$ be a (column) vector of coefficients in the expansion:

$$\sum_j c_j p_j(x;\mu) = \begin{bmatrix} p_0(x;\mu) & p_1(x;\mu) & \cdots \end{bmatrix} \mathbf{c} = \begin{bmatrix} p_0(x;\nu) & p_1(x;\nu) & \cdots \end{bmatrix} \mathcal{U}_{\nu\to\mu}\mathbf{c}.$$

So, we see that $\mathbf{d} = \mathcal{U}_{\nu\to\mu}\mathbf{c}$ gives the coefficients in the basis $(p_j(x;\nu))_{j\geq 0}$. We have defined this for the orthonormal polynomials (with respect to a probability measure), and for good reason. It gives a canonical definition of the connection coefficients. But we can also consider

$$\begin{bmatrix} C_0^{(\lambda)}(x) & C_1^{(\lambda)}(x) & \cdots \end{bmatrix} = \begin{bmatrix} C_0^{(\lambda+1)}(x) & C_1^{(\lambda+1)}(x) & \cdots \end{bmatrix} \underbrace{\begin{bmatrix} \tilde{u}_{00} & \tilde{u}_{10} & \cdots & \\ & \tilde{u}_{11} & \tilde{u}_{12} & \cdots \\ & & \ddots & \ddots \end{bmatrix}}_{\tilde{\mathcal{U}}_{\lambda\to\lambda+1}}.$$

We find

$$\int_{-1}^{1} C_k^{(\lambda)}(x) C_j^{(\lambda+1)}(x)(1-x^2)^{\lambda+\frac{1}{2}}\mathrm{d}x = \tilde{u}_{kj}\int_{-1}^{1} C_j^{(\lambda+1)}(x)^2(1-x^2)^{\lambda+\frac{1}{2}}\mathrm{d}x.$$

Then for $j < k-2$, since $(1-x^2)C_j^{(\lambda+1)}(x)$ is a polynomial of degree $< j$, $\tilde{u}_{kj} = 0$. This implies the connection coefficient matrix is banded, with upper-bandwidth two (a matrix with upper- and lower-bandwidth zero is a diagonal matrix). Then it also follows from the symmetry of the weight function that $C_j^{(\lambda)}(x)$ is even/odd if j is even/odd. And therefore $\tilde{u}_{k,k-1} = 0$. To compute a formula for the remaining coefficients, we use the observation that if

$$\begin{aligned} p_j(x;\mu) &= c_j x^j + O(x^{j-1}), \\ h_j(x) &= d_j x^j + O(x^{j-1}), \end{aligned}$$

Then, by orthogonality with lower-degree polynomials

$$\int_{\mathbb{R}} p_j(x;\mu)h_j(x)\mu(\mathrm{d}x) = \frac{d_j}{c_j}\int_{\mathbb{R}} p_j(x;\mu)^2\mu(\mathrm{d}x).$$

And thus the leading coefficient is all that is needed to compute such an integral. The following holds

$$C_j^{(\lambda)}(x) = \begin{cases} \frac{\lambda}{\lambda+j}\left(C_j^{(\lambda+1)}(x) - C_{j-2}^{(\lambda+1)}(x)\right) & k \geq 2, \\ \frac{\lambda}{\lambda+1}C_1^{(\lambda+1)}(x) & k = 1, \\ C_0^{(\lambda+1)}(x) & k = 0, \end{cases} \quad \lambda \geq 1. \tag{3.10}$$

Exercise 3.6. *Verify* (3.10).

3.7 ▪ Polynomial interpolation at arbitrary nodes

Of course, interpolation can be discussed with no reference to orthogonal polynomials. While there is a lot of literature and perspectives here, such as Newton's divided differences, we only discuss Lagrange interpolation and the barycentric formula.

3.7.1 ▪ Lagrange interpolation

Let $P = \{x_1, \ldots, x_N\}$, $x_1 < x_2 < \cdots < x_N$ be a set of interpolation points in $[-1,1]$ and f be a function defined at these points. We wish to construct a function h that interpolates this function with a degree $N-1$ polynomial. The Lagrange basis polynomials ℓ_j are defined by

$$\ell_j(x) = \frac{\prod\limits_{\substack{i=1\\i\neq j}}(x-x_i)}{\prod\limits_{\substack{i=1\\i\neq j}}(x_j-x_i)}$$

Then it follows that this is a degree $N-1$ polynomial that satisfies $\ell_j(x_i)=0$ if $i\neq j$ and $\ell_j(x_j)=1$. Thus

$$h(x) = \mathcal{I}^P f(x) = \mathcal{I}^P_{N-1} f(x) := \sum_{j=1}^N f(x_j)\ell_j(x),$$

is the desired polynomial. For a given measure μ, if one can compute $\nu_j = \int_{\mathbb{R}} \ell_j(x)\mu(\mathrm{d}x)$ for each j in some reasonable manner, one arrives at an interpolatory quadrature rule

$$\sum_{j=1}^N f(x_j)\nu_j.$$

The norm $\|\mathcal{I}^P_N\|_{C([-1,1])\to P_{N-1}}$ is the Lebesgue constant associated with the set P. The determination of a set of nodes that minimizes this remains an open problem.

3.7.2 ▪ Barycentric interpolation formula

Barycentric interpolation is yet a different viewpoint on interpolation and is derived from the Lagrange interpolant. Let $\nu(x)=\nu_N(x)=\prod_{j=1}^N(x-x_j)$ be the node polynomial. Then

$$\ell_j(x) = \ell_j(x)\frac{\nu(x)}{\nu(x)} = \frac{\nu(x)\frac{1}{x-x_j}}{\prod\limits_{\substack{i=1\\i\neq j}}(x_j-x_i)}.$$

So, if we define $\omega_j^{-1} = \prod_{\substack{i=1\\i\neq j}}(x_j-x_i)$, then we write

$$\ell_j(x) = \nu(x)\frac{\omega_j}{x_j-x},$$

giving

$$\mathcal{I}^P f(x) = \nu(x)\sum_{j=1}^N f_j\frac{\omega_j}{x_j-x}.$$

The numerical stability of this formula can be improved by using the identity

$$1 = \nu(x)\sum_{j=1}^N \frac{\omega_j}{x_j-x}.$$

Therefore

$$\mathcal{I}^P f(x) = \frac{\displaystyle\sum_{j=1}^{N} f_j \frac{\omega_j}{x_j - x}}{\displaystyle\sum_{j=1}^{N} \frac{\omega_j}{x_j - x}}.$$

We note that while the introduction of removable singularities might seem dangerous, it has been shown that this formula has good numerical stability properties [10]. Furthermore, it can be evaluated in $O(N)$ FLOPs, whereas the Lagrange form appears to require $O(N^2)$ FLOPs. The improvement comes from the fact that the weights ω_j can be precomputed.

3.7.3 ▪ Runge phenomenon

The Runge phenomenon is essentially the fact that $\|\mathcal{I}^P\|_{C([-1,1]) \to P_{N-1}}$ grows exponentially for equally spaced interpolation nodes as the spacing tends to zero.

3.8 ▪ Convergence of general orthogonal polynomial interpolants

3.8.1 ▪ Convergence of orthogonal polynomial interpolants — Jacobi polynomials

To understand the convergence for smooth, but not necessarily analytic functions it suffices to estimate the norms (i.e., the Lebesgue constants)

$$\|\mathcal{I}_N^\mu\|_{C(\mathbb{I}) \to P_{N-1}}.$$

It was shown by Szegő [19, pg. 336] (see also Fejér and Shohat) that for Jacobi parameters $\alpha, \beta > -1$, if $\mu(\mathrm{d}x) = w_{\alpha,\beta}(x)\mathrm{d}x$ is as in (3.7), then

$$\|\mathcal{I}_N^\mu\|_{C(\mathbb{I}) \to P_{N-1}} = O(\max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}). \tag{3.11}$$

3.8.2 ▪ Convergence for analytic functions using general orthogonal polynomials*

In this section we determine bounds on the convergence rate for interpolants using the zeros of general orthogonal polynomials. The main assumption we make is that the Jacobi matrix $\mathcal{J}(\mu)$ is sufficiently close to $\mathcal{J}(\mu_{\mathrm{T}})$.

The first general result we need to address convergence is a classical eigenvalue perturbation result.

Lemma 3.18 (Weilandt–Hoffman inequality). *Let $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{C}^{N\times N}$ be Hermitian and use $\lambda_j(\mathbf{A}_i)$ to denote the jth eigenvalue (in increasing order) of $\mathbf{A}_i$. Then*

$$\sum_{j=1}^{N} |\lambda_j(\mathbf{A}_1) - \lambda_j(\mathbf{A}_2)|^2 \leq \|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{F}}^2,$$

where $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{Tr}\mathbf{A}^*\mathbf{A}$.

Proof. The first claim is that for every permutation σ of N integers

$$\sum_{j=1}^{N} |\lambda_j(\mathbf{A}_1) - \lambda_j(\mathbf{A}_2)|^2 \leq \sum_{j=1}^{N} |\lambda_j(\mathbf{A}_1) - \lambda_{\sigma(j)}(\mathbf{A}_2)|^2.$$

This follows from the so-called rearrangement inequality [9], which states that, in particular,

$$\sum_j \lambda_j(\mathbf{A}_1)\lambda_j(\mathbf{A}_2) \geq \sum_j \lambda_j(\mathbf{A}_1)\lambda_{\sigma(j)}(\mathbf{A}_2). \tag{3.12}$$

The rearrangement inequality can be established by straightforward induction. We also take the matrices to have distinct eigenvalues, in which case equality in the rearrangement inequality is only attained for the identity permutation. We will argue below why it suffices to consider the case of distinct eigenvalues.

Diagonalize $\mathbf{A}_i = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^*$, where $\mathbf{\Lambda}_i$ has its diagonal entries in increasing order. Then we need to show that

$$\mathrm{Tr}(\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)^2 \leq \mathrm{Tr}(\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^* - \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^*)^2.$$

Expanding the left-hand side, we have

$$\mathrm{Tr}\left[\mathbf{\Lambda}_1^2 + \mathbf{\Lambda}_2^2 - 2\mathbf{\Lambda}_1\mathbf{\Lambda}_2\right].$$

For the right-hand side:

$$\mathrm{Tr}\left[\mathbf{\Lambda}_1^2 + \mathbf{\Lambda}_2^2 - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^*\mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^* - \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^*\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^*\right].$$

Using the cyclic property of the trace the inequality is therefore equivalent to

$$\mathrm{Tr}\,\mathbf{\Lambda}_1\mathbf{\Lambda}_2 \geq \mathrm{Tr}\,\mathbf{U}_2^*\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^*\mathbf{U}_2\mathbf{\Lambda}_2.$$

Set $\mathbf{U} = \mathbf{U}_1^*\mathbf{U}_2$ and

$$F(\mathbf{U}) := \mathrm{Tr}\,\mathbf{U}^*\mathbf{\Lambda}_1\mathbf{U}\mathbf{\Lambda}_2.$$

It suffices to show the right-hand side is maximized, over all unitary matrices, when $\mathbf{U}$ is a diagonal multiple of $\mathbf{I}$. Since F is a continuous function on the set of unitary matrices, which is compact, F attains its maximum, at say, $\mathbf{V}$, $F(\mathbf{V}) = \max_{\mathbf{U}\ |\ \mathbf{U}^*\mathbf{U}=\mathbf{I}} F(\mathbf{U})$. Then given $\epsilon > 0$ and a Hermitian matrix $\mathbf{A}$,

$$\mathbf{U} = \mathbf{U}_\epsilon = \mathbf{V}(\mathrm{i} - \epsilon\mathbf{A})^{-1}(\mathrm{i} + \epsilon\mathbf{A}),$$

is also a unitary matrix. We now expand $\mathbf{U}$ as $\epsilon \to 0$ giving

$$\mathbf{U} = \mathbf{V} - 2\mathrm{i}\epsilon\mathbf{V}\mathbf{A} + O(\epsilon^2). \tag{3.13}$$

And therefore

$$F(\mathbf{U}) = F(\mathbf{V}) + 2\mathrm{i}\epsilon\mathrm{Tr}\left[\mathbf{A}\mathbf{V}^*\mathbf{\Lambda}_1\mathbf{V}\mathbf{\Lambda}_2 - \mathbf{V}\mathbf{\Lambda}_1\mathbf{V}\mathbf{A}\mathbf{\Lambda}_2\right] + O(\epsilon^2).$$

For this to be the maximum, the $O(\epsilon)$ term must vanish identically, for any choice of $\mathbf{A}$. We find

$$\operatorname{Tr}\mathbf{A}\left[\mathbf{V}^*\mathbf{\Lambda}_1\mathbf{V}\mathbf{\Lambda}_2 - \mathbf{\Lambda}_2\mathbf{V}^*\mathbf{\Lambda}_1\mathbf{V}\right] = 0, \text{ for all } \mathbf{A}, \quad \mathbf{A} = \mathbf{A}^*.$$

Since the matrix in brackets is Hermitian, we can choose $\mathbf{A}$ to equal to it and find that

$$\mathbf{V}^*\mathbf{\Lambda}_1\mathbf{V}\mathbf{\Lambda}_2 - \mathbf{\Lambda}_2\mathbf{V}^*\mathbf{\Lambda}_1\mathbf{V} = \mathbf{0}.$$

This is stating that the matrices $\mathbf{V}^*\mathbf{\Lambda}_1\mathbf{V}$, $\mathbf{\Lambda}_2$ commute. Therefore, their eigenspaces must coincide. This implies that $\mathbf{V}$ is a diagonal (with modulus one diagonal entries) multiple of a permutation matrix. The value of F is independent of this diagonal matrix. So let $\mathbf{P}$ be a permutation matrix and we find

$$F(\mathbf{DP}) = F(\mathbf{P}) = \operatorname{Tr}(\mathbf{\Lambda}_1\mathbf{P})^*(\mathbf{P}\Lambda_2) = \sum_j \lambda_j(\mathbf{A}_1)\lambda_{\sigma(j)}(\mathbf{A}_2).$$

And then we see that $\mathbf{P} = \mathbf{I}$, again by the (strict) rearrangement inequality, since $\mathbf{V}$ is a maximizer. This establishes the inequality in the case of distinct eigenvalues. For any Hermitian matrix with repeated eigenvalues, we can approximate it with a matrix with distinct eigenvalues where the eigenvalues lie within ϵ of original eigenvalues. Since the lemma holds for this matrix, we send ϵ to zero to obtain the desired result. ■

Exercise 3.7. *Establish* (3.13).

Exercise 3.8. *Show that* $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{Tr}\mathbf{B}^*\mathbf{A}$ *defines an inner product on* $\mathbb{C}^{N\times N}$. *Hint: Show* $\operatorname{Tr}\mathbf{B}^*\mathbf{A} = \sum_{i,j}\bar{b}_{ij}a_{ij}$.

3.8.3 ▪ Hermite interpolation formula and some potential theory

The key supposition in the Hermite formula is that the function being interpolated is analytic in an open region Ω that contains the unit interval $\mathbb{I}$. We aim to construct a degree $N-1$ interpolant at the nodes $P = \{x_1, x_2, \ldots, x_N\}$, $x_1 < x_2 < \cdots < x_N$, $x_j \in \mathbb{I}$. While the following holds for more general regions, we just state it for Bernstein ellipses.

Proposition 3.19 (Hermite interpolation formula). *Suppose f is analytic in $B_\rho(\mathbb{I})$. For any $1 < \rho' < \rho$,*

$$f(x) - \mathcal{I}^P(x) = \frac{1}{2\pi \mathrm{i}}\int_{\partial B_{\rho'}(\mathbb{I})} \frac{\nu_N(x)}{\nu_N(z)}\frac{f(z)}{z-x}\mathrm{d}z, \quad x \in \mathbb{I}, \quad \nu_N(x) = \prod_{j=1}^N (x - x_j).$$

To verify this formula it is easy to first check that the right-hand side vanishes at x_j for each j. And then a residue calculation confirms that the right-hand side is indeed equal to $f(x)$ plus a polynomial.

Exercise 3.9. *Prove Proposition 3.19.*

This gives the basic estimate

$$|f(x) - \mathcal{I}^P(x)| \leq \frac{\|f\|_{L^1(\partial B_{\rho'}(\mathbb{I}))}}{2\pi} \frac{\max_{x\in\mathbb{I}} |\nu_N(x)|}{\min_{z\in\Gamma} |\nu_N(z)|}. \tag{3.14}$$

Next, define the discrete measure

$$\mu_N = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_j}, \tag{3.15}$$

so that

$$|\nu_N(z)| = \exp\left(-N \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_N(\mathrm{d}x) \right).$$

For $z \notin \mathbb{I}$, $\log \frac{1}{|z-x|}$ is an infinitely smooth function of x (with norms depending on z). If the x_j's are chosen via Proposition 3.16, then

$$\int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_N(\mathrm{d}x) = \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_{\mathrm{T}}(\mathrm{d}x) + o(1).$$

But we note that, by the mean-value theorem log, for $0 < x, y$ satisfies

$$|\log x - \log y| \leq \frac{1}{\min\{x,y\}} |y - x|.$$

Since $|z - x| \geq 1/2(\rho' - 1/\rho') =: d(\rho')$ for $x \in \mathbb{I}$, $z \in B_{\rho'}(\mathbb{I})$, we have

$$|\log|z-x| - \log|z-x'|| \leq d(\rho')^{-1} ||z-x| - |z-x'|| \leq d(\rho')^{-1}|x - x'|.$$

Therefore, the error in the best approximation of $\log \frac{1}{|z-x|}$ with a degree N polynomial is $O(N^{-1})$, uniformly for $z \in B_{\rho'}(\mathbb{I})$:

$$\max_{z\in B_{\rho'}(\mathbb{I})} \left| \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_N(\mathrm{d}x) - \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_{\mathrm{T}}(\mathrm{d}x) \right| = O(N^{-1}).$$

This allows one to estimate the denominator in our error bound, at least for this specific choice of nodes. To estimate it for a general set of nodes, we must make some assumptions.

Lemma 3.20. *Suppose that μ is such that* $\operatorname{supp}(\mu) = \mathbb{I}$ *and*

$$\sup_{N\geq 1} \|\mathbf{J}_N(\mu) - \mathbf{J}_N(\mu_{\mathrm{T}})\|_{\mathrm{F}} < \infty.$$

Then the node polynomial $\nu_N(x) = \prod_{j=1}^{N}(x - \lambda_j)$ for the eigenvalues λ_j of $\mathbf{J}_N(\mu)$ satisfies

$$-\frac{1}{N} \log|\nu_N(z)| = \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_{\mathrm{T}}(\mathrm{d}x) + O(N^{-1/2}),$$

uniformly for $z \in B_\rho(\mathbb{I})$, for ρ fixed.

Proof. As above let x_j be chosen as in Proposition 3.16. In light of the previous calculations, it suffices to show that

$$\left| \frac{1}{N} \sum_{j=1}^{N} f(x_j) - \frac{1}{N} \sum_{j=1}^{N} f(\lambda_j) \right| = \frac{L}{\sqrt{N}} \|\mathbf{J}_N(\mu) - \mathbf{J}_N(\mu_{\mathrm{T}})\|_{\mathrm{F}},$$

where L is the Lipschitz constant for f. And indeed, this follows from the Weilandt–Hoffman inequality:

$$\left|\frac{1}{N}\sum_{j=1}^{N} f(x_j) - \frac{1}{N}\sum_{j=1}^{N} f(\lambda_j)\right| \leq \frac{L}{N}\sum_{j=1}^{N}|x_j-\lambda_j| \leq \frac{L}{\sqrt{N}}\left(\sum_{j=1}^{N}|x_j-\lambda_j|^2\right)^{1/2}.$$

■

The next estimate is on $\nu_N(x)$ for $x \in [-1,1]$.

Lemma 3.21. *Suppose that* μ *is such that* $\operatorname{supp}(\mu) = \mathbb{I}$ *and*

$$\sup_{N\geq 1}\|\mathbf{J}_N(\mu) - \mathbf{J}_N(\mu_{\mathrm{T}})\|_{\mathrm{F}} < \infty.$$

Then the node polynomial $\nu_N(x) = \prod_{j=1}^{N}(x-\lambda_j)$ *for the eigenvalues* λ_j *of* $\mathbf{J}_N(\mu)$ *satisfies*

$$-\frac{1}{N}\log|\nu_N(x)| \geq \int_{\mathbb{R}} \log\frac{1}{|x-s|}\mu_{\mathrm{T}}(\mathrm{d}s) + O(N^{-1/6}),$$

uniformly for $x \in \mathbb{I}$.

Proof. We do not try to show convergence of the left-hand side to the right because ν_N has zeros whereas the right-hand side does not. We consider the expression

$$-\frac{1}{N}\log|\nu_N(x)| = \int_{\mathbb{R}} \log\frac{1}{|x-s|}\mu_N(\mathrm{d}s), \quad \mu_N = \frac{1}{N}\sum_{j=1}^{N}\delta_{\lambda_j},$$

where we assign the value of $+\infty$ if $x = \lambda_j$ for any j. Then we define

$$\log_M(x) = \begin{cases} \log M & x \geq M, \\ \log(x) & \text{otherwise.}\end{cases}$$

We have

$$\begin{aligned}
&\int_{\mathbb{R}} \log\frac{1}{|x-s|}\mu_N(\mathrm{d}s) - \int_{\mathbb{R}} \log\frac{1}{|x-s|}\mu_{\mathrm{T}}(\mathrm{d}s)\\
&= \int_{\mathbb{R}}\left[\log\frac{1}{|x-s|} - \log_M\frac{1}{|x-s|}\right]\mu_N(\mathrm{d}s)\\
&- \int_{\mathbb{R}}\left[\log\frac{1}{|x-s|} - \log_M\frac{1}{|x-s|}\right]\mu_{\mathrm{T}}(\mathrm{d}s)\\
&+ \int_{\mathbb{R}}\log_M\frac{1}{|x-s|}(\mu_N(\mathrm{d}s) - \mu_{\mathrm{T}}(\mathrm{d}s))\\
&\geq -\int_{\mathbb{R}}\left[\log\frac{1}{|x-s|} - \log_M\frac{1}{|x-s|}\right]\mu_{\mathrm{T}}(\mathrm{d}s)\\
&+ \int_{\mathbb{R}}\log_M\frac{1}{|x-s|}(\mu_N(\mathrm{d}s) - \mu_{\mathrm{T}}(\mathrm{d}s)).
\end{aligned}$$

Then we note that $\log_M(1/|x|)$ has Lipschitz constant at most M, and then it follows that we can approximate it with polynomials of degree N, uniformly with error $O(M/N)$. Let $\tilde{p}_{N,x}(s)$ be this approximation. Then

$$\begin{aligned}
&\int_{\mathbb{R}} \log_M \frac{1}{|x-s|}(\mu_N(\mathrm{d}s) - \mu_{\mathrm{T}}(\mathrm{d}s)) \\
&\quad = \int_{\mathbb{R}} \log_M \frac{1}{|x-s|}(\mu_{N,\mathrm{T}}(\mathrm{d}s) - \mu_{\mathrm{T}}(\mathrm{d}s)) \\
&\quad + \int_{\mathbb{R}} \log_M \frac{1}{|x-s|}(\mu_{N,}(\mathrm{d}s) - \mu_{N,\mathrm{T}}(\mathrm{d}s)) \\
&\quad = \int_{\mathbb{R}} \left[\log_M \frac{1}{|x-s|} - \tilde{p}_{N,x}(s)\right](\mu_{N,\mathrm{T}}(\mathrm{d}s) - \mu_{\mathrm{T}}(\mathrm{d}s)) \\
&\quad + \int_{\mathbb{R}} \tilde{p}_{N,x}(s)(\mu_{N,\mathrm{T}}(\mathrm{d}s) - \mu_{\mathrm{T}}(\mathrm{d}s)) \\
&\quad + \int_{\mathbb{R}} \log_M \frac{1}{|x-s|}(\mu_{N,}(\mathrm{d}s) - \mu_{N,\mathrm{T}}(\mathrm{d}s)).
\end{aligned}$$

Here $\mu_{N,\mathrm{T}} = N^{-1}\sum_j \delta_{x_j}$ where the x_j are again as in Proposition 3.16. The first term here is $O(M/N)$, the second vanishes identically if $M \leq 2N-1$. The last is $O(M/\sqrt{N})$ by the Weilandt–Hoffman inequality.

It remains to estimate

$$\begin{aligned}
&\int_{\mathbb{R}} \left[\log \frac{1}{|x-s|} - \log_M \frac{1}{|x-s|}\right] \mu_{\mathrm{T}}(\mathrm{d}s) \\
&\quad = \int_{|x-s|\leq M^{-1}} \log \frac{1}{M|x-s|}\mu_{\mathrm{T}}(\mathrm{d}s)
\end{aligned}$$

If we just consider

$$\int_{x\leq s\leq M^{-1}+x} \log \frac{1}{M|x-s|}\mu_{\mathrm{T}}(\mathrm{d}s),$$

then this is clearly maximized, for $M > 1$, when $x = -1$. And so,

$$\int_{|x-s|\leq M^{-1}} \log \frac{1}{M|x-s|}\mu_{\mathrm{T}}(\mathrm{d}s) \leq 2\int_{|s+1|\leq M^{-1}} \log \frac{1}{M|s+1|}\mu_{\mathrm{T}}(\mathrm{d}s) = O(M^{-1/2}).$$

Choosing $M = N^{1/3}$ gives the result. ■

Theorem 3.22. *Suppose f is analytic and bounded in $B_\rho(\mathbb{I})$ for $\rho > 1$. Suppose also that μ is supported on $\mathbb{I}$ such that*

$$\sup_{N\geq 1} \|\mathbf{J}_N(\mu) - \mathbf{J}_N(\mu_{\mathrm{T}})\|_{\mathrm{F}} < \infty.$$

Then

$$\limsup_{N\to\infty} \frac{1}{N} \log \|f - \mathcal{I}_N^\mu f\|_\infty \leq -\log \rho.$$

Proof. Using (3.14), for $1 < \rho' < \rho$, and applying Lemmas 3.20 and 3.21, we have

$$\frac{1}{N} \log \|f - \mathcal{I}_N^\mu f\|_\infty \leq - \max_{z \in \partial B_{\rho'}(\mathbb{I})} \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_{\mathrm{T}}(\mathrm{d}x) \\ + \min_{x \in \mathbb{I}} \int_{\mathbb{R}} \log \frac{1}{|x-s|} \mu_{\mathrm{T}}(\mathrm{d}s) + O(N^{-1/6}).$$

It can then be shown that [14]

$$\pi \int_{\mathbb{R}} \log \frac{1}{|z-x|} \mu_{\mathrm{T}}(\mathrm{d}x) = -\log |J_+^{-1}(z)| - \log 2,$$

where J_+^{-1} is an inverse of $z \mapsto 1/2(z + 1/z)$, $J_+^{-1}(z) = z + \sqrt{z-1}\sqrt{z+1}$. The result follows. ■

Part II

Operator equations and spectrum approximation

Chapter 4

Bounded and unbounded linear operators

The purpose of this chapter is to introduce the basics of operators on Banach spaces. Much of the focus is then concerning operators on Hilbert spaces. We include a rather direct approach to working with unbounded operators and full generality is not the purpose here. The goal is to build up enough theory to understand approximate solutions of operator equations of the form

$$\mathcal{L}\mathbf{u} = \mathbf{b}.$$

A key concept is that a linear operator $\mathcal{L}$ has, generally speaking, its own domain, denoted $D(\mathcal{L})$. We always require $D(\mathcal{L})$ to be dense in the ambient space. So, given two operators $\mathcal{L}, \mathcal{M}$, the sum $\mathcal{L} + \mathcal{M}$ is only defined as an operator on the intersection of the two domains $D(\mathcal{L}) \cap D(\mathcal{M})$. The case where this issue is easily overlooked is that of bounded linear operators.

4.1 ▪ Bounded linear operators

Definition 4.1. *Suppose V and W are Banach spaces. The space $L(V, W)$ is the space of bounded linear operators $\mathcal{L}$ from V to W ($D(\mathcal{L}) = V$) satisfying*

$$\|\mathcal{L}\|_{V\to W} := \sup_{\mathbf{v}\in V:\mathbf{v}\neq\mathbf{0}} \frac{\|\mathcal{L}\mathbf{v}\|_W}{\|\mathbf{v}\|_V} = \sup_{\mathbf{v}\in V:\|\mathbf{v}\|_V=1} \|\mathcal{L}\mathbf{v}\|_W < \infty.$$

In the previous definition V and W need not be complete, but we will not consider such a setting and the following actually only needs W to be a Banach space.

Proposition 4.2. *$L(V, W)$ is a Banach space.*

We have some other elementary properties. Suppose $\mathcal{L} \in L(V, W)$:

- Suppose $\mathcal{M} \in L(W, X)$, then $\mathcal{M}\mathcal{L} \in L(V, X)$ and

$$\|\mathcal{M}\mathcal{L}\|_{V\to X} \leq \|\mathcal{L}\|_{V\to W}\|\mathcal{M}\|_{W\to X}.$$

- Define $N(\mathcal{L}) = \{\mathbf{v} \in V \mid \mathcal{L}\mathbf{v} = \mathbf{0}\}$. Then $N(\mathcal{L})$ is a closed subspace of V.
- Define $R(\mathcal{L}) = \{\mathcal{L}\mathbf{v} \in V \mid \mathbf{v} \in V\}$. Then $R(\mathcal{L})$ is a subspace of W that may not be closed.
- A linear operator is bounded if and only if it is continuous at $\mathbf{0}$.

Definition 4.3. *An operator $\mathcal{L} \in L(V,W)$ is said to be injective (or one-to-one), if*

$$\mathbf{v}_1 \neq \mathbf{v}_2 \quad \Rightarrow \quad \mathcal{L}v_1 \neq \mathcal{L}v_2,$$

for all $\mathbf{v}_1, \mathbf{v}_2 \in V$.

Evidently, for an injective operator, there exists an inverse on its range. We require more for an operator to be called invertible.

Definition 4.4.

- *An operator $\mathcal{L} \in L(V,W)$ is said to be surjective if $\mathcal{L}V = W$.*
- *An injective and surjective operator (i.e., bijective) operator is said to be invertible.*

This implies that there exists an operator $\mathcal{L}^{-1}$ that maps W to V such that $\mathcal{L}^{-1}\mathcal{L} = \mathrm{id}$. An important theorem in functional analysis is the open mapping theorem.

Theorem 4.5 (Open mapping). *Suppose that $\mathcal{L} \in L(V,W)$ is surjective, then $\mathcal{L}\Omega$, $\Omega \subset V$ is open whenever Ω is open (i.e. $\mathcal{L}$ is an open map).*

The following is a consequence of the open mapping theorem shows that invertible bounded linear operators always have bounded inverse.

Corollary 4.6. *If $\mathcal{L} \in L(V,W)$ is invertible then $\mathcal{L}^{-1} \in L(W,V)$.*

We also note an important fact. If $\mathcal{L} \in L(V,W)$ is invertible then set

$$\mathcal{L}\mathcal{L}^{-1}\mathbf{v} = \mathbf{w}.$$

And then

$$\mathcal{L}^{-1}\mathbf{v} = \mathcal{L}^{-1}\mathbf{w}.$$

As $\mathcal{L}^{-1}$ is injective, $\mathbf{v} = \mathbf{w}$. And therefore $\mathcal{L}^{-1}$ is also the right inverse of $\mathcal{L}$.

Lemma 4.7. *Suppose $\mathcal{L} \in L(V)$. If there exists $k \geq 1$ such that $\|\mathcal{L}^k\| < 1$ then $\mathrm{id} - \mathcal{L}$ is invertible and*

$$(\mathrm{id} - \mathcal{L})^{-1} = \sum_{j=0}^{\infty} \mathcal{L}^j, \quad \mathcal{L}^0 := \mathrm{id}\,.$$

Proof. We first show that the series above converges. For $N = pk + q$, for integers p, q, $0 \leq q < k$

$$\sum_{j=0}^{N} \mathcal{L}^j = \left[\sum_{\ell=0}^{k-1} \mathcal{L}^\ell\right] \sum_{j=0}^{p-1} \mathcal{L}^{kj} + \left[\sum_{\ell=0}^{q} \mathcal{L}^\ell\right] \mathcal{L}^{pk} \tag{4.1}$$

Since $\mathcal{L}$ is bounded, we have for any $0 \leq q < k$, and a constant $M > 0$

$$\left\| \sum_{\ell=0}^{q} \mathcal{L}^\ell \right\| \leq M.$$

Thus implies that for any fixed q, the series (4.1), using the convergence of the geometric series, is absolutely convergent as $p \to \infty$. Then, we note for $N = pk + q, N' = pk + q'$ we have

$$\left\| \sum_{j=0}^{N} \mathcal{L}^j - \sum_{j=0}^{N'} \mathcal{L}^j \right\| \leq M \|\mathcal{L}^k\|^p,$$

and the series must all converge to the quantity. Lastly, we see that

$$\left(\sum_{j=0}^{\infty} \mathcal{L}^j\right) (\mathrm{id} - \mathcal{L}) = \lim_{N\to\infty} \left(\sum_{j=0}^{N} \mathcal{L}^j\right) (\mathrm{id} - \mathcal{L}) = \lim_{N\to\infty} (\mathrm{id} - \mathcal{L}^{N+1}) = \mathrm{id},$$

where the limits refer to convergence in $L(V)$. ■

The following perturbation theorem is essential.

Theorem 4.8. *Suppose that $\mathcal{L} \in L(V, W)$ is invertible. If $\mathcal{M} \in L(V, W)$ is such that $\|\mathcal{L} - \mathcal{M}\|_{V\to W} < \|\mathcal{L}^{-1}\|_{W\to V}^{-1}$, then $\mathcal{M}$ is also invertible. Furthermore, we have the following estimates*

$$\|\mathcal{M}^{-1}\|_{W\to V} \leq \frac{\|\mathcal{L}^{-1}\|_{W\to V}}{1-\rho},$$
$$\|\mathcal{M}^{-1} - \mathcal{L}^{-1}\|_{W\to V} \leq \frac{\rho}{1-\rho}\|\mathcal{L}^{-1}\|_{W\to V},$$

where $\rho = \|\mathcal{L}^{-1}\|_{W\to V}\|\mathcal{L} - \mathcal{M}\|_{V\to W}$

Proof. Consider

$$\mathcal{M} = \mathcal{L} - (\mathcal{L} - \mathcal{M}) = \mathcal{L}(\mathrm{id} - \mathcal{L}^{-1}(\mathcal{L} - \mathcal{M})).$$

Then set $\mathcal{L}_0 = \mathcal{L}^{-1}(\mathcal{L} - \mathcal{M})$ and we bound

$$\|\mathcal{L}_0\|_V \leq \|\mathcal{L}^{-1}\|_{W\to V}\|\mathcal{L} - \mathcal{M}\|_{V\to W} =: \rho < 1.$$

Therefore, by Lemma 4.7, $\mathrm{id} - \mathcal{L}^{-1}(\mathcal{L} - \mathcal{M})$ is invertible and therefore $\mathcal{M}$ is the composition of invertible operators and is therefore invertible.

Then, we bound

$$\|(\mathrm{id} - \mathcal{L}^{-1}(\mathcal{L} - \mathcal{M}))^{-1}\|_V \leq \sum_{j=0}^{\infty} \rho^j = \frac{1}{1-\rho}.$$

Thus

$$\|\mathcal{M}^{-1}\|_{W\to V} \leq \frac{\|\mathcal{L}^{-1}\|_{W\to V}}{1-\rho}.$$

And

$$\mathcal{M}^{-1} - \mathcal{L}^{-1} = (\mathrm{id} - \mathcal{L}^{-1}(\mathcal{L} - \mathcal{M}))^{-1}\mathcal{L}^{-1} - \mathcal{L}^{-1} = \left(\sum_{j=1}^{\infty} \mathcal{L}_0^j\right) \mathcal{L}^{-1}.$$

Thus

$$\|\mathcal{M}^{-1} - \mathcal{L}^{-1}\|_{W\to V} \leq \frac{\rho}{1-\rho}\|\mathcal{L}^{-1}\|_{W\to V}.$$

■

We find the following by setting $\mathcal{L} = -z\,\mathrm{id}$ and $\mathcal{M} = \mathcal{L} - z$.

Corollary 4.9. *Suppose $\mathcal{L} \in L(V)$. Then for $|z| > \|\mathcal{L}\|$, $\mathcal{L} - z$ is invertible.*

4.1.1 ▪ Bounded operators on Hilbert spaces

Suppose $\mathcal{L} \in L(V)$ where V is a Hilbert space. A key aspect of a Hilbert space is that it gives a cleaner notion of the adjoint of an operator. When the operator is bounded, many technicalities do not arise, when compare to unbounded operators.

For a fixed $\mathbf{w} \in V$, consider the bounded linear functional

$$\ell_{\mathbf{w}}(\mathbf{v}) = \langle \mathcal{L}\mathbf{v}, \mathbf{w}\rangle.$$

By the Riesz representation theorem there exists $\mathbf{z} \in V$ such that

$$\ell_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{z}\rangle.$$

And we are led to define $\mathcal{L}^*\mathbf{w} = \mathbf{z}$. And it follows immediately that $\|\mathcal{L}^*\|_V < \infty$. The operator $\mathcal{L}^*$ is called the adjoint of $\mathcal{L}$

Definition 4.10. *An bounded operator $\mathcal{L}$ on a Hilbert space V is said to be self-adjoint if $\mathcal{L}^* = \mathcal{L}$.*

Proposition 4.11. *Suppose $\mathcal{L}$ is a self-adjoint operator on a Hilbert space V. Then $\mathcal{L} - \gamma$ is invertible for any $\gamma \in \mathbb{C} \setminus \mathbb{R}$, $\alpha \neq 0$.*

Proof. Without loss of generality we can assume $\gamma = \alpha + \mathrm{i}\beta, \beta > 0$. And by replacing $\mathcal{L}$ with $\mathcal{L} - \alpha$ we can assume $\alpha = 0$. We begin with a straightforward calculation, using the self-adjointness. For $\beta' > 0$

$$\begin{aligned}\|(\mathcal{L} - \mathrm{i}\beta')\mathbf{v}\|^2 &= \langle(\mathcal{L} - \mathrm{i}\beta')\mathbf{v}, (\mathcal{L} - \mathrm{i}\beta')\mathbf{v}\rangle \\ &= \|\mathcal{L}\mathbf{v}\|^2 + \beta'^2\|\mathbf{v}\|^2 + \langle\mathcal{L}\mathbf{v}, \mathrm{i}\beta'\mathbf{v}\rangle + \langle\mathrm{i}\beta'\mathbf{v}, \mathcal{L}\mathbf{v}\rangle \\ &= \|\mathcal{L}\mathbf{v}\|^2 + \beta'^2\|\mathbf{v}\|^2 \geq \beta'^2\|\mathbf{v}\|^2.\end{aligned}$$

This implies that $\mathcal{L} - \mathrm{i}\beta'$ is injective, but most importantly, if this operator is invertible then $\|(\mathcal{L} - \mathrm{i}\beta')^{-1}\| \leq |\beta'|^{-1}$. And we know that if $\beta' > \|\mathcal{L}\|$ this operator is invertible by Corollary 4.9. So, now take β' to be such a value, and take β to satisfy $0 < \beta < \beta'$. Then

$$\|\mathcal{L} - \mathrm{i}\beta' - (\mathcal{L} - \mathrm{i}\beta)\| = |\beta - \beta'|.$$

Then by Theorem 4.8, $\mathcal{L} - \mathrm{i}\beta$ is invertible if

$$|\beta' - \beta| < \beta'.$$

or $\beta > 0$. ■

4.2 ▪ Unbounded linear operators

We use the term *unbounded linear operators* to refer to "not necessarily bounded" operators. And for simplicity, we only consider unbounded operators $\mathcal{L}$ where the domain $D(\mathcal{L})$ and range $R(\mathcal{L})$ are both subspaces of the same Hilbert space V.

Example 4.12. The most common case of an unbounded operator is the derivative operator. Let $\mathcal{D} = \frac{\mathrm{d}}{\mathrm{d}\theta}$ and set $V = L^2(\mathbb{T})$ and $D(\mathcal{D}) = H^1(\mathbb{T})$.

And recall that our definition of an invertible operator is one that is injective and surjective, from its domain onto V. Then we do our gut-check calculation

$$\begin{aligned}&\mathcal{L}^{-1}\mathbf{v} \in D(\mathcal{L}), \\ &\mathcal{L}\mathcal{L}^{-1}\mathbf{v} =: \mathbf{w}, \\ &\mathcal{L}^{-1}\mathbf{v} = \mathcal{L}^{-1}\mathbf{w},\end{aligned}$$

And by injectivity of $\mathcal{L}^{-1}$, $\mathbf{w} = \mathbf{v}$ and $\mathcal{L}^{-1}$ is a right inverse of $\mathcal{L}$.

Definition 4.13. *An unbounded linear operator $\mathcal{L}$ with domain $D(\mathcal{L})$ is said to be closed if the graph*

$$G(\mathcal{L}) = \{(\mathbf{v}, \mathcal{L}\mathbf{v}) \mid \mathbf{v} \in D(\mathcal{L})\},$$

is closed in the product topology of $V \times V$.

In other words, this definition is stating that if $(\mathbf{v}_n, \mathbf{w}_n)$, for $n = 1, 2, \ldots$ is a sequence in $G(\mathcal{L})$ that converges[3] $(\mathbf{v}_n, \mathbf{w}_n) \to (\mathbf{v}, \mathbf{w})$ then $\mathbf{v} \in D(\mathcal{L})$ and $\mathbf{w} = \mathcal{L}\mathbf{v}$.

[3] This convergence means that each of $\mathbf{v}_n, \mathbf{w}_n$ converge in V.

Proposition 4.14. *Let $\mathcal{D} = \frac{\mathrm{d}}{\mathrm{d}\theta}$ and set $V = L^2(\mathbb{T})$ with $D(\mathcal{D}) = H^1(\mathbb{T})$. The $\mathcal{D}$ is closed*

Exercise 4.1. *Prove Proposition 4.14.*

The importance of closed operators is that they, in some sense, have the largest domain possible. And having this property is important when determining if the operator is self-adjoint.

Proposition 4.15. *Suppose $\mathcal{L}$ is a closed operator on a Banach space V, and $\mathcal{M} \in L(V)$ is invertible. Then $\mathcal{ML}$ is closed.*

Proof. Suppose $(\mathbf{u}_n, \mathcal{ML}\mathbf{u}_n)$ converges. Thus $(\mathbf{u}_n, \mathcal{L}\mathbf{u}_n)$ converges and since $\mathcal{L}$ is closed, $\mathbf{u}_n \to \mathbf{u} \in D(\mathcal{L}) = D(\mathcal{ML})$. ■

Proposition 4.16. *Suppose $\mathcal{L}$ is an unbounded operator on a Banach space V. If $z - \mathcal{L}$ is invertible for some $z \in \mathbb{C}$ then $\mathcal{L}$ is closed.*

Proof. Let $\mathbf{v}_n$ be such that $\mathbf{v}_n \to \mathbf{v}$ in V and $\mathcal{L}\mathbf{v}_n \to \mathbf{w}$ in V. Then we write

$$\mathbf{v}_n = (z - \mathcal{L})^{-1}(z - \mathcal{L})\mathbf{v}_n = (z - \mathcal{L})^{-1}(z\mathbf{v}_n - \mathcal{L}\mathbf{v}_n).$$

The right-hand side here converges to $(z - \mathcal{L})^{-1}(z\mathbf{v} - \mathbf{w})$ and this implies that $\mathbf{v} \in D(\mathcal{L})$. Then we check

$$(z - \mathcal{L})\mathbf{v} = (z\mathbf{v} - \mathbf{w}),$$

which implies that $\mathcal{L}\mathbf{v} = \mathbf{w}$ and therefore $\mathcal{L}$ is closed. ■

4.2.1 ▪ Symmetric and self-adjoint unbounded operators

Given an unbounded operator $\mathcal{L}$ with domain $D(\mathcal{L}) \subset V$, one can no longer appeal to the Riesz representation theorem directly to define the adjoint. But we do the following.

Theorem-Definition 4.17. *For unbounded operator $\mathcal{L}$ define*

$$D(\mathcal{L}^*) = \{\mathbf{v} \in V \mid |\langle \mathcal{L}\mathbf{w}, \mathbf{v}\rangle| \le M_{\mathbf{v}}\|\mathbf{w}\| \text{ for all } \mathbf{w} \in D(\mathcal{L})\}.$$

For $\mathbf{v} \in D(\mathcal{L}^*)$, $\mathbf{w} \mapsto \langle \mathcal{L}\mathbf{w}, \mathbf{v}\rangle = \langle \mathbf{w}, \mathbf{x}\rangle$ for a unique $\mathbf{x} \in V$. And define

$$\mathcal{L}^*\mathbf{v} = \mathbf{x}.$$

Definition 4.18. *An unbounded operator $\mathcal{L}$ is said to be symmetric if*

$$\langle \mathcal{L}\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, \mathcal{L}\mathbf{w}\rangle,$$

for all $\mathbf{v}, \mathbf{w} \in D(\mathcal{L})$.

Definition 4.19. *An unbounded operator $\mathcal{L}$ is said to be self-adjoint if it is symmetric and $D(\mathcal{L}) = D(\mathcal{L}^*)$.*

Note that for a symmetric operator $D(\mathcal{L}) \subset D(\mathcal{L}^*)$ and therefore, in some sense, making an operator self-adjoint amounts to finding a domain that is sufficiently large. And thus an operator that is not closed will not be self-adjoint.

The following theorem is helpful, see [16].

Theorem 4.20. *Let $\mathcal{L}$ be a symmetric operator on a Hilbert space V. Then the following are equivalent:*

- *$\mathcal{L}$ is self-adjoint.*
- *$\mathcal{L}$ is closed and $N(\mathcal{L}^* \pm \mathrm{i}) = \{\mathbf{0}\}$.*
- *$R(\mathcal{L} \pm \mathrm{i}) = V$.*

See [24] for the following.

Theorem 4.21 (Closed range). *Suppose $\mathcal{L}$ is a closed unbounded operator on a Hilbert space V. Then the following are equivalent:*

- *$R(\mathcal{L})$ is closed.*
- *$R(\mathcal{L}^*)$ is closed.*
- *$R(\mathcal{L}) = N(\mathcal{L}^*)^\perp$.*
- *$R(\mathcal{L}^*) = N(\mathcal{L})^\perp$.*

Corollary 4.22. *Suppose $\mathcal{L}$ is a closed, and unbounded operator on a Hilbert space V. Then $\mathcal{L}$ is boundedly invertible if and only if $\mathcal{L}^*$ is boundedly invertible.*

Proof. Suppose $\mathcal{L}$ is invertible. Therefore by Theorem 4.21, $R(\mathcal{L}^*) = V$ and a candidate for the inverse of $\mathcal{L}^*$ is $(\mathcal{L}^{-1})^*$, which is everywhere defined because $\mathcal{L}^{-1}$ is bounded. We verify for $\mathbf{v} \in V, \mathbf{w} \in D(\mathcal{L}^*)$

$$\langle \mathbf{v}, (\mathcal{L}^{-1})^*\mathcal{L}^*\mathbf{w}\rangle = \langle \mathcal{L}^{-1}\mathbf{v}, \mathcal{L}^*\mathbf{w}\rangle = \langle \mathcal{L}\mathcal{L}^{-1}\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle.$$

And since $\mathbf{v}$ is arbitrary, $(\mathcal{L}^{-1})^*$ is indeed the inverse of $\mathcal{L}^*$. To verify that it is bounded, take unit vectors $\mathbf{v}, \mathbf{w} \in V$

$$\langle \mathbf{v}, (\mathcal{L}^*)^{-1}\mathbf{w}\rangle = \langle \mathcal{L}^{-1}\mathbf{v}, \mathbf{w}\rangle.$$

So,

$$|\langle \mathbf{v}, (\mathcal{L}^*)^{-1}\mathbf{w}\rangle| \leq \|\mathcal{L}^{-1}\|.$$

Taking $\mathbf{v} = (\mathcal{L}^*)^{-1}\mathbf{w}/\|(\mathcal{L}^*)^{-1}\mathbf{w}\|$ establishes the first implication.

Similarly, if $\mathcal{L}^*$ is boundedly invertible, with inverse $(\mathcal{L}^*)^{-1}$ then its adjoint is the candidate inverse for $\mathcal{L}$ because $R(\mathcal{L}) = V$ by Theorem 4.21. For $\mathbf{v} \in D(\mathcal{L}), \mathbf{w} \in V$

$$\langle ((\mathcal{L}^*)^{-1})^*\mathcal{L}\mathbf{v}, \mathbf{w}\rangle = \langle \mathcal{L}\mathbf{v}, (\mathcal{L}^*)^{-1}\mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle.$$

And a bound on the norm follows similarly.

■

We now establish an analogue of Proposition 4.11.

Proposition 4.23. *Suppose $\mathcal{L}$ is a symmetric operator. If $\mathcal{L} - \gamma$ is invertible for some $\gamma \in \mathbb{C} \setminus \mathbb{R}$, then $\mathcal{L}$ is self-adjoint and $\mathcal{L} - z$ is invertible for all $z \in \mathbb{C} \setminus \mathbb{R}$.*

Proof. We first show $\mathcal{L}$ is self-adjoint: Set $\gamma = \alpha + \mathrm{i}\beta$,

$$\mathcal{L} - \gamma = \beta(\beta^{-1}(\mathcal{L} - \alpha) + \mathrm{i}).$$

The operator $\beta^{-1}(\mathcal{L} - \alpha)$ is self-adjoint if and only if $\mathcal{L}$ is. And since it is invertible, Theorem 4.20 implies that it is self-adjoint.

Now, suppose, without loss, that $\operatorname{Im}\gamma > 0$. For $z = \alpha + \mathrm{i}\beta$, $a > 0$, following the calculations in Proposition 4.11, we again have

$$\|(\mathcal{L} - z)\mathbf{v}\| \geq \alpha\|\mathbf{v}\|.$$

Setting $\mathbf{v} = (\mathcal{L} - z)^{-1}\mathbf{w}$, we find

$$\|(\mathcal{L} - z)^{-1}\| \leq \alpha^{-1}.$$

Then through the identity

$$\mathcal{L} - z = \mathcal{L} - \gamma + (\gamma - z) = (\mathcal{L} - \gamma)(\mathrm{id} + (\gamma - z)(\mathcal{L} - \gamma)^{-1}),$$

we see that we can apply Theorem 4.8, for $|\gamma - z|/|\operatorname{Im}\gamma| < 1$ to see that $\mathcal{L} - z$ is invertible. This implies that $\mathcal{L} - z$ is invertible for $\operatorname{Im} z > 0$.

To see that the same claim holds in the lower half-plane, we see that

$$(\mathcal{L} - \gamma)^* = \mathcal{L} - \bar{\gamma}.$$

And Corollary 4.22 implies this is invertible. The same argument can now be repeated in the lower-half plane. ■

4.3 ▪ Compact linear operators

Definition 4.24. *An operator $\mathcal{K} \in L(V, W)$ is said to be compact if the image of bounded sets in V under $\mathcal{K}$ are precompact in W.*

Theorem 4.25.

- *For $\mathcal{K} \in L(V, W)$, suppose there exists a sequence $(\mathcal{K}_n)_{n\geq 1}$ in $L(V, W)$, where $\mathcal{K}_n$ is compact, such that $\|\mathcal{K} - \mathcal{K}_n\|_{V\to W} \to 0$. Then $\mathcal{K}$ is compact and thus the set of compact operators in $L(V, W)$ is a closed linear subspace.*
- *Suppose $\mathcal{K} \in L(V, W)$ is compact. And suppose there exists a sequence $(\mathcal{P}_n)_{n\geq 1}$ in $L(W)$, and $\mathcal{P} \in L(W)$, such that $\|\mathcal{P}\mathbf{u} - \mathcal{P}_n\mathbf{u}\|_W \to 0$ for each fixed $\mathbf{u} \in W$. Then $\|\mathcal{P}\mathcal{K} - \mathcal{P}_n\mathcal{K}\|_{V\to W} \to 0$.*

Proof. For the first claim, it suffices to show that for every sequence in $\Omega := \mathcal{K}\overline{B_1(\mathbf{0})}$, there exists a convergent subsequence. So, let $\{\mathbf{u}_j\}_{j\geq 0}$ be a sequence in Ω. This implies that there exists $\mathbf{v}_j$ such that $\mathbf{u}_j = \mathcal{K}\mathbf{v}_j$. We use a diagonal argument.

Since $\mathcal{K}_n\overline{B_1(\mathbf{0})}$ is compact for every n, we first note that there exists a convergent subsequence of $(\mathcal{K}_1\mathbf{v}_j)_{j\geq 0}$. So let, $(\mathbf{v}_j^{(1)})_{j\geq 1}$ be a subsequence of $(\mathbf{v}_j)_{j\geq 0}$ so that $(\mathcal{K}_1 v_j^{(1)})_{j\geq 0}$ converges. In general, let $(v_j^{(k)})_{j\geq 1}$ be a subsequence of $(\mathbf{v}_j^{(k-1)})_{j\geq 0}$ so that $(\mathcal{K}_k\mathbf{v}_j^{(k)})_{j\geq 0}$ converges. We claim that

$$(\mathcal{K}\mathbf{v}_j^{(j)})_{j\geq 0}$$

converges. Let $\epsilon > 0$. Choose N sufficiently large so that $2\sup_j \|u_j\|\|\mathcal{K} - \mathcal{K}_n\| < \epsilon/2$ for $n \geq N$. Note that this supremum is finite because $\mathcal{K}$ is bounded. And then consider for $j < k$

$$\|\mathcal{K}\mathbf{v}_j^{(j)} - \mathcal{K}\mathbf{v}_k^{(k)}\| \leq \|\mathcal{K}_N(\mathbf{v}_j^{(j)} - \mathbf{v}_k^{(k)})\| + \|(\mathcal{K} - \mathcal{K}_N)(\mathbf{v}_j^{(j)} - \mathbf{v}_k^{(k)})\|.$$

The second term is less that $\epsilon/2$. And for $j, k > N$, $v_j^{(j)}, v_k^{(k)}$ are both in the sequence $(v_\ell^{(N)})_{\ell\geq k}$ and since $\mathcal{K}_N v_\ell^{(N)}$ converges, the sequence is Cauchy, and the first term can be made smaller than $\epsilon/2$. This establishes compactness.

For the second claim, the principle of uniform boundedness implies that $\sup_n \|\mathcal{P}_n\|_W < \infty$. Next consider the function

$$\mathbf{v} \mapsto \|(\mathrm{id} - \mathcal{P}_n)\mathbf{v}\|_W.$$

This is a continuous function on W. And therefore it attains its maximum on the compact set $\Omega = \mathcal{K}\overline{B_1(\mathbf{0})}$, $B_1(\mathbf{0}) \subset V$:

$$\max_{\mathbf{v}\in\Omega} \|(\mathcal{P} - \mathcal{P}_n)\mathbf{v}\|_W = \|(\mathcal{P} - \mathcal{P}_n)\mathbf{v}_n\|_W =: M_n$$

for some $\mathbf{v}_n \in \Omega$. We need to show that $M_n \to 0$. We do this by showing that every subsequence of these positive real numbers has a further subsequence that converges to the same value. So let $(M_{n_k})_{k\geq 1}$ be a subsequence.

Suppose $(\tilde{\mathbf{v}}_k)_{k\geq 1}$ is such that $M_{n_k} = \|(\mathcal{P} - \mathcal{P}_{n_k})\tilde{\mathbf{v}}_k\|_W$. By compactness, this sequence has a subsequence that converges, $(\tilde{\mathbf{v}}_{k_j})_{j\geq 1}$, $\tilde{\mathbf{v}}_{k_j} \to \mathbf{v}$, and there exists $\mathbf{u}_j, \mathbf{v}$ such that

$\mathcal{K}\mathbf{u}_j = \tilde{\mathbf{v}}_{k_j}$, $\mathcal{K}\mathbf{u} = \mathbf{v}$. Then

$$M_{n_k} \leq \|(\mathcal{P} - \mathcal{P}_{n_k})\mathbf{v}\|_W + (\|\mathcal{P}\| + \sup_n \|\mathcal{P}_n\|_W)\|\tilde{\mathbf{v}}_k - \mathbf{v}\|_W,$$

upon sending $k \to \infty$ (really, taking the $\limsup_k$ of both sides), see that the limit of M_{n_k} is zero. ■

The following is immediate because the orthogonal projection $\mathcal{P}_n$ onto the first n orthonormal functions in an orthonormal basis converges to the strongly to the identity, i.e,

$$\|\mathbf{v} - \mathcal{P}_n\mathbf{v}\|_V \to 0,$$

for all fixed $\mathbf{v} \in V$.

Corollary 4.26. *In a (separable) Hilbert space, the set of compact operators is the closure, in operator norm, of the set of all finite-rank operators.*

The following is immediate because bounded operators map bounded sets to bounded sets.

Proposition 4.27. *If $\mathcal{L} \in L(X, Y)$ and $\mathcal{K} \in L(Y, Z)$ and one of $\mathcal{L}, \mathcal{K}$ is compact, then $\mathcal{K}\mathcal{L}$ is compact.*

Lemma 4.28.

- *If* $\dim R(\mathcal{K}) < \infty$ *then* $\dim R(\mathcal{K}^*) < \infty$
- *If $\mathcal{K}$ is compact then so is $\mathcal{K}^*$.*

Lemma 4.29. *Suppose $\mathcal{T} \in L(V)$, is such that* $\dim R(\mathcal{T}) = m$. *Then* $\dim N(z-\mathcal{T}) \leq m$ *for $z \neq 0$.*

Proof. Suppose we can find $\mathbf{v}_1, \dots, \mathbf{v}_{m+1}$ linearly independent vectors in $N(z-\mathcal{T})$. Then the vectors $\mathbf{u}_1, \dots, \mathbf{u}_{m+1}$, $\mathbf{u}_j = \mathcal{T}\mathbf{v}_j$ must be linearly dependent because $\dim R(\mathcal{T}) = m$. That means there exists a vector, $\mathbf{v}$, that is a linear combination of $(\mathbf{v}_j)_j$ such that $\mathbf{v} \in N(\mathcal{T})$. But if $\mathbf{v} \in N(\mathcal{T})$ then $\mathbf{v} \notin N(z - \mathcal{T})$ and we arrive at a contradiction. ■

Lemma 4.30. *Suppose $\mathcal{T} \in L(V)$, is such that* $\dim R(\mathcal{T}) < \infty$ *and that for $z \neq 0$, $z - \mathcal{T}$ is surjective. Then $z - \mathcal{T}$ is injective and hence $z - \mathcal{T}$ has a bounded inverse.*

Proof. Set $m = \dim R(\mathcal{T})$. Suppose $z - \mathcal{T}$ is not injective. Then there exists $\mathbf{u}_1 \neq \mathbf{0}$ such that $(z - \mathcal{T})\mathbf{u}_1 = \mathbf{0}$. But as $z - \mathcal{T}$ is surjective, so is $(z - \mathcal{T})^k$ for any $k > 1$. So, we may find $\mathbf{u}_k$ such that $(z - \mathcal{T})^{k-1}\mathbf{u}_k = \mathbf{u}_1$. The key thing here is that $(z - \mathcal{T})^j\mathbf{u}_j = \mathbf{0}$ for $j \leq k$. Now, we claim that these vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ are linearly independent. By induction, suppose that $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$ are linearly independent. If $\mathbf{u}_k$ were to be a linear combination of these vectors, then $(z - \mathcal{T})^{k-1}\mathbf{u}_k = \mathbf{0} \neq \mathbf{u}_1$, giving a contradiction. So, now for $k > m$, we have shown that

$$\dim N((z - \mathcal{T})^k) \geq k > m.$$

But $(z - \mathcal{T})^k = z^k + \mathcal{T}\mathcal{P}$, for a bounded operator $\mathcal{P}$. Necessarily, $\mathcal{T}\mathcal{P}$ is rank m. From Lemma 4.29, we have a contradiction. Thus $z - \mathcal{T}$ is injective and has a bounded inverse by the open mapping theorem.

■

The following is true generally for compact operators on a Banach space, but we only prove it for Hilbert spaces.

Theorem 4.31 (Fredholm alternative for compact operators in Hilbert space). *Suppose $\mathcal{K} \in L(V)$ is a compact operator on a separable Hilbert space V. For each $z \in \mathbb{C} \setminus \{0\}$ exactly one of the following holds:*

- *$z - \mathcal{K}$ has a bounded inverse.*
- $\dim N(z - \mathcal{K}) > 0$.

Proof. First, suppose $\dim N(z - \mathcal{K}) = 0$. Thus $z - \mathcal{K}$ is injective. We just need to show it is surjective. We first show that its range is closed. Let $(\mathbf{v}_n)_{n\geq 0}$ be a convergent sequence in the range, so that $\mathbf{v}_n = (z - \mathcal{K})\mathbf{u}_n$ for some $\mathbf{u}_n$. We claim this new sequence must be bounded. Suppose it is not, then there exists an unbounded subsequence. For simplicity, well just assume the sequence itself is unbounded. Then set $\mathbf{w}_n = \mathbf{u}_n / \|\mathbf{u}_n\|$. Since the sequence in the range is convergent, we have that

$$\mathbf{v}_n \to \mathbf{v} \quad \Rightarrow \quad z\mathbf{w}_n - \mathcal{K}\mathbf{w}_n \to \mathbf{0}.$$

As $(\mathbf{w}_n)_{n\geq 0}$ is a bounded sequence and $\mathcal{K}$ is compact, we find that there must exist a subsequence that converges $\mathbf{w}_{n_k} \to \mathbf{w}$. But that $\mathbf{w}$ is in the nullspace of $z - \mathcal{K}$, a contradiction. So the sequence $(\mathbf{u}_n)_{n\geq 0}$ is bounded.

We find a subsequence of $(\mathbf{u}_n)_{n\geq 0}$ such that $\mathcal{K}\mathbf{u}_{n_k}$ converges, and therefore $\mathbf{u}_{n_k} \to \mathbf{u}$ itself must converge. And $\mathbf{v} = z\mathbf{u} - \mathcal{K}\mathbf{u}$. So, the range of $z - \mathcal{K}$ is closed. This implies that the operator

$$\mathcal{L} : V \to R(\mathcal{K}), \quad \mathcal{L}\mathbf{v} = z\mathbf{v} - \mathcal{K}\mathbf{v},$$

has a bounded inverse by the open mapping theorem. This implies that there exists $c > 0$ such that for $\mathbf{u} \in R(\mathcal{K})$

$$\|\mathcal{L}^{-1}\mathbf{u}\| \leq c^{-1}\|\mathbf{u}\| \quad \Rightarrow \quad c\|\mathbf{v}\| \leq \|(z - \mathcal{K})\mathbf{v}\|,$$

for all $\mathbf{v} \in V$.

Now let $(\mathcal{K}_n)_{\geq 0}$ be a sequence of finite-rank operators that converge to $\mathcal{K}$ in operator norm:

$$c\|\mathbf{v}\| - \|\mathcal{K}_n\mathbf{v} - \mathcal{K}\mathbf{v}\| \leq \|(z - \mathcal{K}_n)\mathbf{v}\|.$$

Upon choosing n large enough, $n > N_0$, we have

$$\frac{c}{2}\|\mathbf{v}\| \leq \|(z - \mathcal{K}_n)\mathbf{v}\|,$$

for all $\mathbf{v} \in V$. This establishes the injectivity of $z - \mathcal{K}_n$. We now show it is invertible: It follows that the adjoint operator $\bar{z} - \mathcal{K}_n^*$ is surjective. And it is then injective by

Lemmas 4.28, 4.30. This implies that $z - \mathcal{K}_n$ is surjective and hence invertible. The invertibility of $z - \mathcal{K}$ follows from Theorem 4.8. ■

The following is helpful in deducing properties that compact operators cannot have.

Lemma 4.32 (Riesz). *For a normed vector space V, suppose W is proper subspace of V and is not dense in V. Then for every $0 < r < 1$ there exists an element $\mathbf{v} \in V \setminus W$, $\|\mathbf{v}\| = 1$, such that $\|\mathbf{v} - \mathbf{w}\| > r$ for all $\mathbf{w} \in W$.*

Proof. Let $\mathbf{v}$ be an element of V that is not in the closure of W. Then $\inf_{\mathbf{w}\in W} \|\mathbf{w}-\mathbf{v}\| := \rho > 0$. For $\epsilon > 0$, there exists $\mathbf{w}' \in W$ such that $\rho \leq \|\mathbf{v} - \mathbf{w}'\| < \rho + \epsilon$. Now consider the vector

$$\mathbf{u} = \frac{\mathbf{v} - \mathbf{w}'}{\|\mathbf{v} - \mathbf{w}'\|} \in \mathbf{V}.$$

For any $\mathbf{w} \in W$, we compute

$$\|\mathbf{w} - \mathbf{u}\| = \left\| \mathbf{w} - \frac{\mathbf{v}}{\|\mathbf{v} - \mathbf{w}'\|} + \frac{\mathbf{w}'}{\|\mathbf{v} - \mathbf{w}'\|} \right\| \geq \frac{1}{\|\mathbf{v} - \mathbf{w}'\|} \inf_{\mathbf{w}\in W} \|\mathbf{w} - \mathbf{v}\| \geq \frac{\rho}{\rho + \epsilon}$$

where we used that $\mathbf{w} + \mathbf{w}'/\|\mathbf{v} - \mathbf{w}'\| \in W$. And this quantity can be made as close to one as desired by shrinking ϵ. ■

Corollary 4.33. *The closed unit ball in any infinite-dimensional normed vector space is not compact.*

Theorem 4.34. *Suppose $\mathcal{K} \in L(X)$ is compact. Then for any $r > 0$,*

$$\dim \operatorname{span}\{\mathbf{v} \mid \mathcal{K}\mathbf{v} = \lambda \mathbf{v}, \quad |\lambda| > r\} < \infty.$$

Proof. If this were infinite, then $r^{-1}\mathcal{K}\overline{B_1(\mathbf{0})}$ would be an infinite-dimensional superset of the unit ball, implying that $\mathcal{K}$ is not compact, a contradiction. ■

For a compact operator $\mathcal{K}$, and $\mathbf{v}_j \in N(\lambda_j - \mathcal{K})$ for $(\lambda_j)_{1\leq j\leq n}$ all distinct, and nonzero, $\mathbf{v}_k \notin \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$. To see this, apply $\mathcal{K}$

$$\begin{aligned} &\sum_j c_j \mathbf{v}_j = \mathbf{0}, \\ &\sum_j c_j \lambda_j \mathbf{v}_j = \sum_j c_j \mathcal{K} \mathbf{v}_j = \mathbf{0}, \\ &\quad\vdots \\ &\sum_j c_j \lambda_j^{n-1} \mathbf{v}_j = \mathbf{0}. \end{aligned} \tag{4.2}$$

Since the Vandermonde matrix associated to distinct points is invertible, this linear system of equations can be solved uniquely for c_j, giving $c_j = 0$ for all j.

Corollary 4.35. *For a compact operator $\mathcal{K}$ on a Banach space B, the following hold:*

- $\dim N(z-\mathcal{K}) < \infty$ *for* $z \neq 0$.
- *The only possible accumulation point of z such that* $\dim N(z-\mathcal{K}) > 0$, *is at the origin.*

4.3.1 ▪ Relatively compact operators

Another important notion is that of relatively compact operators.

Definition 4.36. *An unbounded operator $\mathcal{K}$ on a Banach space is said to be relatively compact with respect to another unbounded operator $\mathcal{L}$ (i.e., $\mathcal{K}$ is $\mathcal{L}$-compact) if:*

- $D(\mathcal{L}) \subset D(\mathcal{K})$,
- *there exists $z \in \mathbb{C}$ such that $\mathcal{L}-z$ has a bounded inverse, and*
- $\mathcal{K}(\mathcal{L}-z)^{-1}$ *is compact on V.*

We note that by the identity, $z, z' \in \rho(\mathcal{L})$

$$(z-\mathcal{L})^{-1} - (z'-\mathcal{L})^{-1} = (z'-z)(z-\mathcal{L})^{-1}(z'-\mathcal{L})^{-1},$$

and Proposition 4.27, if $\mathcal{K}(\mathcal{L}-z)^{-1}$ is compact for one $z \in \rho(\mathcal{L})$ it is compact for all $z \in \rho(\mathcal{L})$.

4.4 ▪ The spectrum

We are now ready to discuss the structure of the spectrum of both bounded and unbounded operators. The following definitions are slightly non-classical because we are avoiding the discussion of Fredholm operators.

Definition 4.37 (Resolvent and resolvent set). *For an unbounded operator $\mathcal{L}$ on a Banach space define the resolvent set*

$$\rho(\mathcal{L}) = \{z \in \mathbb{C} \mid \mathcal{L}-z \text{ has a bounded inverse}\}.$$

The function $\mathcal{R}(z) = \mathcal{R}(z;\mathcal{L}) = (\mathcal{L}-z)^{-1}$, which is well-defined on $\rho(\mathcal{L})$, is called the resolvent of $\mathcal{L}$.

It follows from the relation

$$\mathcal{L}-\lambda = \mathcal{L}-z+(z-\lambda) = (\mathcal{L}-z)\left(\mathrm{id}+(\mathcal{L}-z)^{-1}(z-\lambda)\right),$$

that the resolvent set is open and the resolvent is analytic on the resolvent set, by Theorem 4.8. We then define the spectrum (which must be closed).

Definition 4.38. *For an unbounded operator $\mathcal{L}$ on a Banach space, define the spectrum by $\sigma(\mathcal{L}) = \rho(\mathcal{L})^c$.*

We further divide up the spectrum.

Definition 4.39. *For an unbounded operator $\mathcal{L}$ on a Banach space define the eigenvalues (of finite multiplicity) and discrete spectrum by*

$$\begin{aligned} \sigma_{\mathrm{ev}}(\mathcal{L}) &= \{z \in \sigma(\mathcal{L}) \mid 0 < \dim N(\mathcal{L} - z) < \infty\}, \\ \sigma_{\mathrm{disc}}(\mathcal{L}) &= \{z \in \sigma_{\mathrm{ev}}(\mathcal{L}) \mid \{z\} \cup \rho(\mathcal{L}) \textit{ contains an open neighborhood of } z\}, \end{aligned}$$

respectively. Lastly, define the essential spectrum

$$\sigma_{\mathrm{ess}}(\mathcal{L}) = \sigma(\mathcal{L}) \setminus \sigma_{\mathrm{disc}}(\mathcal{L}).$$

We note that if all eigenvalues are well-separated from the essential spectrum the set of eigenvalues and the discrete spectrum coincide. To further understand crucial aspects of the essential spectrum, one has to further divide it. It turns out that, in many cases, the essential spectrum does not change under perturbations by relatively compact operators, this is Weyl's famous theorem (in the self-adjoint case). This stability is true more generally, but just for subsets of the essential spectrum.

4.4.1 ▪ The Dunford integral

The Dunford integral is defined using the fact that $(z - \mathcal{L})^{-1}$ is an analytic function on $\rho(\mathcal{L})$. And so, one can define integrals

$$\frac{1}{2\pi \mathrm{i}} \oint_\Gamma (z - \mathcal{L})^{-1} \mathrm{d}z,$$

for $\Gamma \subset \rho(\mathcal{L})$. This is the Dunford integral. We present it in limited generality. See [11] for more detail.

Suppose $\lambda \in \sigma_{\mathrm{disc}}(\mathcal{L})$ and that Γ is a circle that encloses λ but no other elements of the spectrum. Suppose that $\mathcal{L}\mathbf{u} = \lambda \mathbf{u}$ for $\mathbf{u} \in D(\mathcal{L})$. Then for $z \in \Gamma$

$$(z - \mathcal{L})\mathbf{u} = (z - \lambda)\mathbf{u}.$$

And then

$$\frac{\mathbf{u}}{z - \lambda} = (z - \mathcal{L})^{-1}\mathbf{u}.$$

From this, we find

$$\mathbf{u} = \frac{1}{2\pi \mathrm{i}} \oint_\Gamma (z - \mathcal{L})^{-1} \mathbf{u} \, \mathrm{d}z.$$

We obtain the important result

Theorem 4.40. *Suppose λ is an isolated point in $\sigma(\mathcal{L})$. Let Γ_λ be a circular contour that encloses λ but no other point in σ. Then*

$$\mathcal{P}_\lambda = \mathcal{P}_\lambda(\mathcal{L}) := \frac{1}{2\pi\mathrm{i}} \oint_{\Gamma_\lambda} (z-\mathcal{L})^{-1}\mathrm{d}z,$$

is a projection onto the eigenspace associated to $\mathcal{L}$ in the sense that $\mathcal{P}_\lambda \mathbf{u} = \mathbf{u}$ whenever $\mathcal{L}\mathbf{u} = \lambda\mathbf{u}$, $\mathbf{u} \in D(\mathcal{L})$. Furthermore, $\mathcal{P}_\lambda \mathbf{v} \in \mathcal{D}(\mathcal{L})$ for $\mathbf{v} \in V$ and $\mathcal{P}_\lambda$ commutes with $\mathcal{L}$.

Proof. We have already seen that it acts on a projection on the eigenvectors. We now need to see it is a projection more generally. Consider

$$\mathcal{P}_\lambda^2 = \frac{1}{4\pi^2} \oint_{\Gamma_\lambda} \oint_{\Gamma'_\lambda} (z-\mathcal{L})^{-1}(z'-\mathcal{L})^{-1}\mathrm{d}z'\mathrm{d}z$$

where Γ'_λ has a slightly larger radius. Then it follows that (i.e., the first resolvent identity)

$$(z-\mathcal{L})^{-1} - (z'-\mathcal{L})^{-1} = (z-z')(z-\mathcal{L})^{-1}(z,-\mathcal{L})^{-1},$$

giving

$$\mathcal{P}_\lambda^2 = \frac{1}{4\pi^2} \oint_{\Gamma_\lambda} \oint_{\Gamma'_\lambda} \left[\frac{(z-\mathcal{L})^{-1} - (z'-\mathcal{L})^{-1}}{z-z'}\right] \mathrm{d}z'\mathrm{d}z.$$

Then we compute, considering that z is inside Γ'_λ

$$\oint_{\Gamma_\lambda} \oint_{\Gamma'_\lambda} \frac{(z-\mathcal{L})^{-1}}{z-z'}\mathrm{d}z'\mathrm{d}z = 2\pi\mathrm{i} \oint_{\Gamma_\lambda} (z-\mathcal{L})^{-1}\mathrm{d}z.$$

The other contribution vanishes, because z' is outside Γ_λ giving the desired result.

The integral in the definition $\mathcal{P}_\lambda$ is defined as the norm limit of finite linear combinations of $(z_j - \lambda)^{-1}$:

$$\begin{aligned} \mathbf{u}_n &= \sum_{k=1}^n w_j(z_j-\mathcal{L})^{-1}\mathbf{v} \to \mathcal{P}_\lambda\mathbf{v} \text{ in } V, \\ \mathcal{L}\mathbf{u}_n &= \sum_{k=1}^n w_j\mathcal{L}(z_j-\mathcal{L})^{-1}\mathbf{v} = \sum_{k=1}^n w_j\mathbf{v} + \sum_{k=1}^n w_j z_j (z_j-\mathcal{L})^{-1}\mathbf{v} \to \mathbf{w} \in V. \end{aligned}$$

Because $\mathcal{L}$ must be closed, $\mathcal{P}_\lambda\mathbf{v} \in D(\mathcal{L})$. And it follows that $\mathcal{L}$ commutes with finite linear combinations of the resolvent, and therefore it must commute with $\mathcal{P}_\lambda$. ■

We establish one more property.

Proposition 4.41. *Suppose $\lambda \in \sigma(\mathcal{L})$, $\lambda \neq 0$ is isolated for a compact operator $\mathcal{K}$ on a Banach space V. Then* $\dim R(\mathcal{P}_\lambda) < \infty$.

Proof. We show that $\mathcal{P}_\lambda$ is compact and therefore it must have a finite-dimensional range. We have, for $z \in \rho(\mathcal{K})$, $z \neq 0$

$$\begin{aligned} -\mathcal{K} &= z - \mathcal{K} - z, \\ -\mathcal{K}(z-\mathcal{K})^{-1} &= \mathrm{id} - z(z-\mathcal{K})^{-1}, \\ (z-\mathcal{K})^{-1} - z^{-1} &= z^{-1}\mathcal{K}(z-\mathcal{K})^{-1}. \end{aligned}$$

Then Proposition 4.27 implies this operator is compact. And then, because $z = 0$ can be taken outside a small circle around λ

$$\mathcal{P}_\lambda = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_\lambda} z^{-1}\mathcal{K}(z-\mathcal{K})^{-1} \mathrm{d}z.$$

This operator is the limit of compact operators and is therefore compact.

■

4.4.2 ▪ The spectrum of compact operators

Theorem 4.42. *Suppose $\mathcal{K}$ is a compact operator on a Banach space V. Then for every $r > 0$*

$$\sigma(\mathcal{K}) \cap \{z \in \mathbb{C} \mid |z| \geq r\} = \sigma_{\mathrm{ev}}(\mathcal{K}) \cap \{z \in \mathbb{C} \mid |z| \geq r\},$$

is a finite set. Furthermore, $\{0\} = \sigma_{\mathrm{ess}}(\mathcal{K})$ if and only if V is infinite dimensional.

Proof. From Corollary 4.35, we just need to establish the last claim. If V is finite dimensional, then $\sigma(\mathcal{K}) = \sigma_{\mathrm{disc}}(\mathcal{K})$. So, suppose that V is infinite dimensional and that $0 \notin \sigma_{\mathrm{ess}}(\mathcal{K})$. If $0 \in \rho(\mathcal{K})$ then $\mathcal{K}\mathcal{K}^{-1}\overline{B_1(\mathbf{0})}$ is compact, which contradicts that V is infinite-dimensional. So $0 \in \sigma(\mathcal{K})$. And if $0 \in \sigma_{\mathrm{disc}}(\mathcal{K})$ we can use $\mathcal{P}_0$ as follows.

We write $\mathcal{K} = (\mathrm{id} - \mathcal{P}_0)\mathcal{K}(\mathrm{id} - \mathcal{P}_1) + \mathcal{P}_1\mathcal{S}(z)\mathcal{P}_1$. Then using

$$\begin{aligned} (z-\mathcal{K})^{-1}\mathcal{P}_0 &= \frac{1}{2\pi \mathrm{i}} \oint_\Gamma (z-\mathcal{K})^{-1}(z/ - \mathcal{K})^{-1} \mathrm{d}z' \\ &= \frac{1}{2\pi \mathrm{i}} \oint_\Gamma \left[(z-\mathcal{K})^{-1} - (z'-\mathcal{K})^{-1}\right] \frac{\mathrm{d}z'}{z'-z} \\ &= (z-\mathcal{K})^{-1} - \frac{1}{2\pi \mathrm{i}} \oint_\Gamma (z'-\mathcal{K})^{-1} \mathrm{d}z'. \end{aligned}$$

It follows that

$$(z-\mathcal{K})^{-1}(\mathrm{id} - \mathcal{P}_0)$$

has an analytic continuation to a neighborhood of $z = 0$ and therefore $\mathcal{K}$ is boundedly invertible as an operator on $R(\mathrm{id} - \mathcal{P}_1)$. This implies that $\mathrm{dim} R(\mathrm{id} - \mathcal{P}_1) < \infty$, but this contradicts that $\mathrm{dim} R(\mathcal{P}_\lambda) < \infty$. ■

4.4.3 ▪ The spectrum of operators with compact resolvent

Theorem 4.43. *Suppose $\mathcal{L}$ is an unbounded operator on a Hilbert space V. If $z_0 \in \rho(\mathcal{L}) \neq \varnothing$, and $(\mathcal{L} - z_0)^{-1}$ is compact then $\sigma(\mathcal{L}) = \sigma_{\text{disc}}(\mathcal{L})$.*

Proof. Since the theorem is true for $\mathcal{L}$ if and only if it is true for $\mathcal{L} - z_0$, we can suppose that $z_0 = 0$. Then we write, for $z \neq 0$,

$$\begin{aligned} z - \mathcal{L} &= \mathcal{L}(\mathcal{L}^{-1} - z^{-1})z, \\ \mathcal{L}^{-1} - z^{-1} &= \mathcal{L}^{-1}(z - \mathcal{L})z^{-1}. \end{aligned}$$

Setting $w = z^{-1} \neq 0$,

$$\mathcal{L}^{-1} - w = \mathcal{L}^{-1}(w^{-1} - \mathcal{L})w.$$

And therefore $w \in \sigma(\mathcal{L}^{-1})$ if and only if $w^{-1} \in \sigma(\mathcal{L})$. The theorem follows. ■

An important result is the following, which is a restricted version of Weyl's theorem.

Theorem 4.44. *Suppose that $\mathcal{L}$ has a compact resolvent. If $\mathcal{K}$ is $\mathcal{L}$-compact and $\rho(\mathcal{L}) \cap \rho(\mathcal{L} + \mathcal{K}) \neq \varnothing$ then $\mathcal{K} + \mathcal{L}$ has a compact resolvent. And therefore, if there exists $z \in \rho(\mathcal{L})$ such that $\|\mathcal{K}(\mathcal{L} - z)^{-1}\| < 1$ then $\mathcal{K} + \mathcal{L}$ has a compact resolvent.*

Proof. Suppose $z \in \rho(\mathcal{L}) \cap \rho(\mathcal{L} + \mathcal{K})$ and consider

$$\begin{aligned} \mathcal{K} &= \mathcal{L} + \mathcal{K} - z - (\mathcal{L} - z), \\ (\mathcal{L} - z)^{-1} - (\mathcal{L} + \mathcal{K} - z)^{-1} &= (\mathcal{L} + \mathcal{K} - z)^{-1}\mathcal{K}(\mathcal{L} - z)^{-1}. \end{aligned}$$

Because the right-hand side is compact, the first claim follows. For the second we write

$$\mathcal{L} + \mathcal{K} - z = (\text{id} - \mathcal{K}(z - \mathcal{L})^{-1})(\mathcal{L} - z).$$

The last claim then follows from Theorem 4.8. ■

This is true in much greater generality, see [11, Theorem IV.5.35].

4.4.4 ▪ A special case of the spectral theorem for self-adjoint operators

Theorem 4.45. *Suppose $\mathcal{K} \in L(V)$ is self-adjoint and compact on a Hilbert space V. Then there exists an orthonormal basis $\mathbf{u}_j$, $j = 1, 2, 3, \dots$ for V, satisfying*

$$\mathcal{K}\mathbf{u}_j = \lambda_j \mathbf{u}_j,$$

and

$$\mathcal{K}\mathbf{v} = \sum_{j=1}^{\infty} \lambda_j \langle \mathbf{v}, \mathbf{u}_j \rangle \mathbf{u}_j.$$

Theorem 4.46. *Suppose $\mathcal{L}$ is self-adjoint on a Hilbert space V with a compact resolvent. Then there exists an orthonormal basis $\mathbf{u}_j$, $j = 1, 2, 3, \ldots$ for V, satisfying*

$$\mathcal{L}\mathbf{u}_j = \lambda_j \mathbf{u}_j,$$

and

$$\mathcal{L}\mathbf{v} = \sum_{j=1}^{\infty} \lambda_j \langle \mathbf{v}, \mathbf{u}_j \rangle \mathbf{u}_j.$$

Chapter 5

Integral equations

In this section, we consider the numerical solution of integral equations on $\mathbb{I}$ and $\mathbb{T}$. For the case of Fredholm equations, we consider things in some generality. For Volterra equations, we only consider the case of a separable kernel.

5.1 ▪ Fredholm integral equations

A Fredholm integral equation is an operator equation of the form

$$\lambda u(x) - \int_\Omega K(x,y)u(y)\mu(\mathrm{d}y) = f(x), \quad (\lambda - \mathcal{K})u = f.$$

for a suitable kernel function $K(x,y) : \Omega \times \Omega \to \mathbb{C}$ and right-hand side $f : \Omega \to \mathbb{C}$. And before we begin, we need some statements about which Banach spaces one can consider such an integral equation on. To do this, Minkowski's inequality for integrals is helpful, see [6].

Theorem 5.1. *If $1 \leq p \leq \infty$, $f(\diamond, y) \in L^p(\mu)$ for a.e. y, and the function $y \mapsto \|f(\diamond, y)\|_p$ is in $L^1(\nu)$, then $f(x, \diamond) \in L^1(\nu)$ for a.e. x, the function $x \mapsto \int f(x,y)d\nu(y)$ is in $L^p(\mu)$, and*

$$\left\| \int f(\diamond, y)\mu(\mathrm{d}y) \right\|_p \leq \int \|f(\diamond, y)\|_p \mu(\mathrm{d}y).$$

The following holds in much more generality, but we only prove it for $\mathbb{I}$.

Proposition 5.2. *Suppose $K : \mathbb{I} \times \mathbb{I}$ satisfies*

$$\int_{\mathbb{I}\times\mathbb{I}} |K(x,y)|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}y) < \infty.$$

Then $\mathcal{K}$ is a compact, bounded linear operator on $L^2(\mu)$ with norm bounded by

$$\|K\|_{L^2(\mu\times\mu)}.$$

Proof. By Minkowsi's inequality

$$\left(\int_{\mathbb{I}}\left|\int_{\mathbb{I}} K(x,y)u(y)\mu(\mathrm{d}y)\right|^2 \mu(\mathrm{d}x)\right)^{1/2} = \left\|\int_{\mathbb{I}} K(\diamond,y)u(y)\mu(\mathrm{d}y)\right\|_2$$
$$\leq \int_{\mathbb{I}} \|K(\diamond,y)\|_2 |u(y)|\mu(\mathrm{d}y).$$

And our hypotheses imply that

$$\int_{\mathbb{I}} \|K(\diamond,y)\|_2^2 \mu(\mathrm{d}y) < \infty,$$

so that the boundedness claim follows by the Cauchy-Schwarz inequality. It then follows immediately that

$$(p_j(\lhd;\mu)p_k(\rhd;\mu))_{j,k\geq 0},$$

is an orthonormal system for $L^2(\mu\times\mu)$. The fact that it is a basis follows from the Stone-Weierstrass theorem, giving the density of polynomials in $C(\mathbb{I}\times\mathbb{I})$ which implies the density of polynomials in $L^2(\mu\times\mu)$. This all implies that we can approximate $K(x,y)$, to within any desired accuracy in $L^2(\mu\times\mu)$, by a finite-rank kernel. Thus $\mathcal{K}$ is compact on $L^2(\mu)$. ■

Proposition 5.3. *Suppose $K : \mathbb{I}\times\mathbb{I}\to\mathbb{C}$ is such that for $L > 0$, $0 < \gamma \leq 1$ and a non-negative integer k, $\partial_x^j K(x,y)$ exists and satisfies $|\partial_x^j K(x,y)| \leq h(y) \in L^1(\mu)$, $0 \leq j \leq k$ and*

$$\int_{\mathbb{I}} |\partial_x^k K(x,y) - \partial_x^k K(x',y)|\mu(\mathrm{d}y) \leq L|x-x'|^\gamma, \quad x,y \in \mathbb{I}.$$

Then $\mathcal{K}$ is a bounded linear operator from $C(\mathbb{I})$ to $C^{k,\gamma}(\mathbb{I})$, with operator norm bounded by $(k+1)\|h\|_{L^1} + L$, and is therefore a compact, bounded linear operator on $C(\mathbb{I})$.

Proof. Set $f = \mathcal{K}u$, and first consider the case $k = 0$. We have

$$|f(x)| \leq \int_{\mathbb{I}} |K(x,y)||u(y)|\mu(\mathrm{d}y) \leq \|u\|_\infty \int_{\mathbb{I}} h(y)\mu(\mathrm{d}y) < \infty.$$

Therefore $\|f\|_\infty \leq M\|u\|_\infty$, $M := \int_{\mathbb{I}} h(y)\mu(\mathrm{d}y)$. This shows that $\mathcal{K}$ is bounded on $C(\mathbb{I})$. Now, consider

$$|f(x) - f(x')| = \left|\int_{\mathbb{I}} [K(x,y) - K(x',y)]u(y)\mu(\mathrm{d}y)\right|$$
$$\leq \|u\|_\infty \int_{\mathbb{I}} |K(x,y) - K(x',y)|\mu(\mathrm{d}y) \leq L\|u\|_\infty |x-x'|^\gamma.$$

Therefore

$$\|f\|_{C^{0,\gamma}} \leq (L+M)\|u\|_\infty.$$

This implies that $\mathcal{K}$ is bounded from $C(\mathbb{I})$ to $C^{0,\gamma}(\mathbb{I})$. But, then we note that we have shown

$$\|f - \mathcal{I}_N^{\mathrm{T}} f\|_\infty \leq C\|f\|_{C^{0,\gamma}} N^{-\gamma}\log N.$$

Thus the identity operator mapping $C^{0,\gamma}(\mathbb{I})$ to $C(\mathbb{I})$ is compact and Proposition 4.27 implies that so is $\mathcal{K}$.

The generalized dominated convergence theorem (Theorem 1.26, see specifically, [6, Theorem 2.27]) can then be used to show that, for $0 < j \leq k$,

$$f^{(j)}(x) = \int_{\mathbb{I}} \partial_x^j K(x,y)u(y)\mu(\mathrm{d}y).$$

The claim then follows by repeating the above estimates. ■

Also, the metric $|x-x'|$ could be replaced with another metric, say $d(x,x')$ for $C(\mathbb{T})$ and the theorem would remain true.

5.1.1 ▪ The Nyström (collocation) method

The Nyström method for solving a Fredholm integral equation is based around simply replacing the integral with a quadrature rule, and then enforcing that the resulting equation should hold at a set of nodes. This last part is called collocation. So, Nyström technically refers to the method of discretizing the integral and collocation refers to the method of "closing" the system. To illustrate this, we consider

$$\lambda u(x) - \int_{\mathbb{I}} K(x,y)u(y)\mu(\mathrm{d}y) = f(x).$$

We suppose that, at least, $f \in C(\mathbb{I})$. We use the N-point Gauss-Legendre quadrature rule for μ with nodes x_j and weights w_j. The weights are normalized so that they satisfy $\sum_{j=1}^N w_j = 1$. Then the Nyström approximation becomes

$$\lambda u_j - \sum_{k=1}^N K(x_j,x_k)w_k u_k = f(x_j), \quad j = 1,2,\ldots,N,$$

where we hope that $u_j \approx u(x_j)$.

We understand the convergence of such an approximation using projections, namely $\mathcal{I}_N^\mu = \mathcal{I}_N^\mu$ where μ is Lebesgue measure on $[-1,1]$. And to do this, we consider a pure collocation method. To a vector $\mathbf{v}$ of function values v_j at the nodes, we abuse notation and use $\mathcal{I}_N^\mu \mathbf{v}$ to denote the polynomial interpolant. Then the pure collocation method reads

$$\lambda \tilde{u}_j - \int_{\mathbb{I}} K(x_j,y)\mathcal{I}_N^\mu \tilde{\mathbf{u}}(y)\mu(\mathrm{d}y) = f(x_j), \quad j = 1,2,\ldots,N,$$

This system is equivalent to

$$v_N - \mathcal{I}_N^\mu \int_{\mathbb{I}} K(\diamond,y)v_N(y)\mu(\mathrm{d}y) = (\lambda - \mathcal{I}_N^\mu \mathcal{K})v_N = \mathcal{I}_N^\mu f, \quad v_N \in R(\mathcal{I}_N^\mu). \tag{5.1}$$

Thus, our method of proving the convergence of the Nyström method is to first prove that this method converges. Note that without strong assumptions on $K(x,y)$ this method is not implementable as is.

We begin with an important observation that if we can solve $(\lambda - \mathcal{I}_N^\mu \mathcal{K})v = \mathcal{I}_N^\mu f$ then we are guaranteed that $v \in R(\mathcal{I}_N^\mu)$. And furthermore, if the finite-dimensional system (5.1) failed to have a solution, it would imply a non-trivial nullspace for $\lambda - \mathcal{I}_N^\mu \mathcal{K}$ restricted to $R(\mathcal{I}_N^\mu)$. All of this is to say, that we can consider solving $(\lambda - \mathcal{I}_N^\mu \mathcal{K})v_N = \mathcal{I}_N^\mu f$ for $v_N \in C(\mathbb{I})$. We now establish the following.

Theorem 5.4. *Suppose $\mathcal{K}_n \in L(V)$ is such that $\mathcal{K}_n \to \mathcal{K}$ in operator norm. Then if $\lambda \in \rho(\mathcal{K})$, then, for sufficiently large n the following hold:*

- *$\lambda - \mathcal{K}_n$ is invertible.*
- *If $\mathbf{w} \in V_n$, $R(\mathcal{K}_n) \subset V_n$ for V_n a subspace of V, then $(\lambda - \mathcal{K}_n)^{-1}\mathbf{w} \in V_n$.*
- *There exists a constant $C > 0$ such that*

$$\|(\lambda - \mathcal{K}_n)^{-1}\mathbf{w} - (\lambda - \mathcal{K})^{-1}\mathbf{w}'\|_V \leq C\left(\|\mathcal{K} - \mathcal{K}_n\|_V \|\mathbf{w}\|_V + \|\mathbf{w} - \mathbf{w}'\|_V\right).$$

Furthermore, if $V_n = R(\mathcal{P}_n)$ for a bounded projection $\mathcal{P}_n$ and $\mathcal{K}_n = \mathcal{P}_n\mathcal{K}$, $\mathbf{w} = \mathcal{P}_n\mathbf{w}'$, then

$$c\|\mathbf{v}' - \mathcal{P}_n\mathbf{v}'\|_V \leq \|(\lambda - \mathcal{K}_n)^{-1}\mathbf{w} - (\lambda - \mathcal{K})^{-1}\mathbf{w}'\|_V \leq c^{-1}\|\mathbf{v}' - \mathcal{P}_n\mathbf{v}'\|_V,$$

$\mathbf{v}' = (\lambda - \mathcal{K})^{-1}\mathbf{w}'$, for some $0 < c < 1$.

Proof. The first claim follows from Theorem 4.8. The second claim follows from the relation

$$\mathbf{v} = (\lambda - \mathcal{K}_n)^{-1}\mathbf{w} \;\Rightarrow\; \lambda\mathbf{v} = \mathcal{K}_n\mathbf{v} + \mathbf{w}.$$

And the last claim follows from Theorem 4.8 in the following way. Let

$$\rho_n = \|\mathcal{K} - \mathcal{K}_n\|_V \|(\lambda - \mathcal{K})^{-1}\|_V.$$

And we write

$$\begin{aligned}&(\lambda - \mathcal{K}_n)^{-1}\mathbf{w} - (\lambda - \mathcal{K})^{-1}\mathbf{w}' \\ &= (\lambda - \mathcal{K}_n)^{-1}\mathbf{w} - (\lambda - \mathcal{K})^{-1}\mathbf{w} + (\lambda - \mathcal{K})^{-1}\mathbf{w} - (\lambda - \mathcal{K})^{-1}\mathbf{w}'.\end{aligned}$$

A triangle inequality gives the claim.

It remains to establish the lower bound which requires more careful rearrangement of the equations. We have

$$\begin{aligned}&(\lambda - \mathcal{K})\mathbf{v}' = \mathbf{w}', \\ &(\lambda\mathcal{P}_n - \mathcal{P}_n\mathcal{K})\mathbf{v}' = \mathcal{P}_n\mathbf{w}', \\ &(\lambda - \mathcal{P}_n\mathcal{K})\mathbf{v}' = \lambda(\mathbf{v}' - \mathcal{P}_n\mathbf{v}') + \mathcal{P}_n\mathbf{w}', \\ &(\lambda - \mathcal{P}_n\mathcal{K})\mathbf{v}' = \lambda(\mathbf{v}' - \mathcal{P}_n\mathbf{v}') + (\lambda - \mathcal{P}_n\mathcal{K})\mathbf{v}, \\ &(\lambda - \mathcal{P}_n\mathcal{K})(\mathbf{v}' - \mathbf{v}) = \lambda(\mathbf{v}' - \mathcal{P}_n\mathbf{v}'),\end{aligned} \tag{5.2}$$

and then the desired bounds follow from Theorem 4.8. ■

Theorem 5.5. *Consider the operator*

$$\mathcal{K}u(x) = \int_{\mathbb{I}} K(x,y)u(y)\mu(\mathrm{d}y),$$

where $\mu(\mathrm{d}x) = w_{\alpha,\beta}(x)\mathrm{d}x$ is a Jacobi measure (3.7) with parameters α, β. Assume $k + \gamma > \max\{\alpha + \frac{1}{2}, \beta + \frac{1}{2}, 0\}$ and assume the hypotheses of Proposition 5.3 hold with these parameters. Further, assume that for each x

$$K(x,\diamond) \in C^{k,\gamma}(\mathbb{I}), \quad |\partial_y^k K(x,y) - \partial_y^k K(x,y')| \leq L|y-y'|^\gamma,$$

where L is independent of x. Then if $f \in C^{k,\gamma}(\mathbb{I})$ and $\lambda \in \rho(\mathcal{K})$, as an operator on $C(\mathbb{I})$, the Nyström method applied to $(\lambda - \mathcal{K})u = f$ converges, and the approximant $\mathcal{I}_N^\mu \mathbf{u}(x)$ to $u(x)$ satisfies

$$\|\mathcal{I}_N^\mu \mathbf{u} - u\|_\infty = O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}).$$

Before we prove this theorem, we pause to recall in this theorem that if $f \in C^{k,\gamma}(\mathbb{I})$ then

$$\|f - \mathcal{I}_N^\mu f\|_\infty = O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}),$$

using Jackson's fourth theorem (Theorem 3.4) and (3.11).

Proof. In light of Theorem 5.4, we define an operator $\mathcal{K}_N$ by

$$\mathcal{K}_N u(x) = \sum_{j=1}^N K(x,x_j)w_j u(x_j),$$

and it suffices to show that $\|\mathcal{K} - \mathcal{I}_N^\mu \mathcal{K}_N\|_\infty$ but because of the sampling of u at the nodes, this is a challenge when the operator is posed on $C(\mathbb{I})$. But, importantly, based on the assumptions, we have

$$\|\mathcal{I}_N^\mu f - f\|_\infty = O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}).$$

But we first show that $\|\mathcal{K} - \mathcal{I}_N^\mu \mathcal{K}\|_\infty = O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\})$. which then establishes the claim for the pure collocation method (5.1). So, consider

$$\begin{aligned}
\|\mathcal{K} - \mathcal{I}_N^\mu \mathcal{K}\|_\infty &\leq \|\operatorname{id} - \mathcal{I}_N^\mu\|_{C^{k,\gamma} \to C(\mathbb{I})} \|\mathcal{K}\|_{C(\mathbb{I}) \to C^{k,\alpha}}, \\
\|\operatorname{id} - \mathcal{I}_N^\mu\|_{C^{k,\gamma} \to C(\mathbb{I})} &= O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}), \\
\|\mathcal{K}\|_{C(\mathbb{I}) \to C^{k,\alpha}} &< \infty.
\end{aligned}$$

Therefore, if v_N is the solution to (5.1),

$$\|v_N - u\|_\infty = O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}).$$

We now consider $\lambda - \mathcal{I}_N^\mu \mathcal{K}$ and $\lambda - \mathcal{I}_N^\mu \mathcal{K}_N$ as operators on $V_N = R(\mathcal{I}_N^\mu)$ with the $\|\diamond\|_\infty$ norm. We have shown that $\mathcal{I}_N^\mu \mathcal{K} \to \mathcal{K}$ in norm, so that $\|(\lambda - \mathcal{I}_N^\mu \mathcal{K})^{-1}\|_\infty < 2\|(\lambda - \mathcal{K})^{-1}\|_\infty$ for N sufficiently large. And

$$\|(\lambda - \mathcal{I}_N^\mu \mathcal{K})^{-1}\|_{V_N} \leq \|(\lambda - \mathcal{I}_N^\mu \mathcal{K})^{-1}\|_\infty.$$

It will suffice to show that

$$\|\mathcal{I}_N^\mu \mathcal{K}_N - \mathcal{I}_N^\mu \mathcal{K}\|_{V_N} = O(N^{-k-\gamma} \max\{N^{\alpha+\frac{1}{2}}, N^{\beta+\frac{1}{2}}, \log N\}).$$

For $u \in V_N$:

$$\begin{aligned} &\int K(x,y)u(y)\mu(\mathrm{d}y) - \int K(x,y)u(y)\mu_N(\mathrm{d}y) \\ &= \int [K(x,y) - p_x(y)]u(y)\mu(\mathrm{d}y) - \int [K(x,y) - p_x(y)]u(y)\mu_N(\mathrm{d}y), \end{aligned}$$

for any $p_x \in V_N$, by the exactness of the Gaussian quadrature rule. From the assumption that

$$K(x,\diamond) \in C^{k,\gamma}(\mathbb{I}), \quad |\partial_y^k K(x,y) - \partial_y^k K(x,y')| \leq L|y-y'|^\gamma,$$

by Jackson's fourth theorem, we can choose $p_x(y)$ so that

$$|K(x,y) - p_x(y)| \leq CLN^{-k-\gamma},$$

and this claim follows, again using the norm bound on $\mathcal{I}_N^\mu$.

■

5.1.2 ▪ The Galerkin projection method

The Galerkin projection method for solving

$$\lambda u(x) - \int_{\mathbb{I}} K(x,y)u(y)\mu(\mathrm{d}y) = f(x).$$

Is similar to (5.1), but the system is not closed by evaluating at a set of points, but rather by enforcing that inner products should vanish. Namely, set $u_N(x) = \sum_{j=0}^{N-1} c_j p_j(x;\mu)$, and enforce

$$\left\langle \lambda u_N - \int_{\mathbb{I}} K(\diamond,y)u_N(y)\mu(\mathrm{d}y) - f, p_k \right\rangle_\mu = 0, \quad k = 0,1,2,\ldots N-1.$$

This gives a linear system for the unknown coefficients c_j. But note that this, in general, suffers from the same issues as the pure collocation method described above — one has to resort to the approximation of integrals to compute the entries in this linear system. So, we then consider a special case where the Galerkin projection method becomes what is called a finite-section method.

We know that $K(x,y) = \sum_{\ell,j=0}^\infty k_{\ell j}\, p_\ell(x;\mu)p_j(y;\mu)$. We know that the approximation of a general kernel by one of this form is possible, but we have not made claims about convergence of such an approximation as yet, beyond that of convergence in $L^2(\mu \times \mu)$. Then we compute

$$\int_{\mathbb{I}} K(x,y)p_j(y;\mu)\mu(\mathrm{d}y) = \sum_{\ell=0}^\infty k_{\ell j} p_\ell(x;\mu).$$

So, supposing that $\lambda \in \rho(\mathcal{L})$, as an operator on $L^2(\mu)$, we know that there exists a solution $u \in L^2(\mu)$, $u = \sum_j c_j p_j$, the coefficients $\mathbf{c} = (c_j)_{j \geq 0}$ satisfy

$$\lambda \mathbf{c} - \begin{bmatrix} k_{00} & k_{01} & k_{02} & \cdots \\ k_{10} & k_{11} & k_{12} & \cdots \\ k_{20} & k_{21} & k_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \mathbf{c} = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \end{bmatrix}, \quad f_j = \langle f, p_j \rangle_\mu.$$

Taking the upper-left principal $N \times N$ sublock of the matrix and solving

$$\lambda \mathbf{c}_N - \begin{bmatrix} k_{00} & k_{01} & k_{02} & \cdots & k_{0,N-1} \\ k_{10} & k_{11} & k_{12} & \cdots & k_{1,N-1} \\ k_{20} & k_{21} & k_{22} & \cdots & k_{2,N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{N-1,0} & k_{N-1,1} & k_{N-1,2} & \cdots & k_{N-1,N-1} \end{bmatrix} \mathbf{c}_N = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-1} \end{bmatrix}, \tag{5.3}$$

Is precisely the same thing as applying the Galerkin method when $K(x, y)$ is replaced with $K_N(x, y) = \sum_{\ell,j=0}^{N} k_{\ell j}\, p_\ell(x; \mu) p_j(y; \mu)$ and f is replaced with $f_N(x) = \sum_{j=0}^{N-1} f_j p_j(x; \mu)$. In this case we can see that $\mathbf{c}_N$, provided this linear system is nonsingular, is the true solution of

$$\lambda u(x) - \int_{\mathbb{I}} K_N(x, y) u(y) \mu(\mathrm{d}y) = \lambda u(x) - \mathcal{K}_N u(x) = f(x).$$

And so, the analysis of the implementation of such a method proceeds in two ways, if the coefficients $k_{\ell j}$ are known, then one has to understand the convergence of $K_N \to K$ since,

$$\|\mathcal{K} - \mathcal{K}_N\|_{L^2(\mu)} \leq \|K - K_N\|_{L^2(\mu \times \mu)},$$

and the convergence of $\sum_{j=0}^{N-1} \langle f, p_j \rangle_\mu p_j$ to f. If the coefficients are not known, then we consider

$$\tilde{K}_N(x, y) = \sum_{\ell,j=1}^{N} \tilde{k}_{\ell j} p_\ell(x; \mu) p_j(y; \mu),$$

where $\tilde{k}_{\ell j} \approx k_{\ell j}$ are obtained via some approximate means (e.g., Gaussian quadrature). And replace, say, $\langle f, p_j \rangle_\mu$ is replaced by $\langle f, p_j \rangle_{\mu,N}$.

In either case we can apply the following theorem.

Theorem 5.6. *Suppose $K, K_N \in L^2(\mu \times \mu)$ are such that $\|K - K_N\|_{L^2(\mu \times \mu)} \to 0$ where*

$$K_N(x, y) = \sum_{\ell,j=1}^{N} k_{\ell j} p_\ell(x; \mu) p_j(y; \mu).$$

Denote the associated Fredholm integral operators by $\mathcal{K}, \mathcal{K}_N$. Then if $f \in L^2(\mu)$ and $\lambda \in \rho(\mathcal{K})$, for sufficiently large N the linear system is uniquely solvable for $\mathbf{c}_N = (c_j)_{j=0}^{N-1}$, for N sufficiently large, and there exists $C > 0$ such that $u_N(x) = \sum_{j=0}^{N-1} c_j p_j(x; \mu)$ satisfies

$$\|u_N - u\|_{L^2(\mu)} \leq C \left[\|K - K_N\|_{L^2(\mu \times \mu)} + \|f - f_N\|_{L^2(\mu)} \right],$$

where $f_N(x) = \sum_{j=0}^{N-1} f_j p_j(x;\mu)$

We point out in this theorem that f_j need not be given by $\langle f, p_j\rangle_\mu$ — it can be replaced by an approximation.

Exercise 5.1. *Show that the finite-section method and the Nyström method are equivalent when we use*

$$f_j = \langle f, p_j\rangle_{\mu,N}, \quad k_{\ell j} = \langle h_\ell, p_j\rangle_{\mu,N}, \quad h_\ell(x) = \langle K(x,\diamond), p_\ell\rangle_{\mu,N}.$$

5.2 ▪ Application: A boundary integral equation

Let $\Omega \subset \mathbb{R}^2$ be a bounded, simply connected open set with a smooth boundary $\partial\Omega$. By smooth boundary, we mean that there exists a parameterization

$$\partial\Omega = \{(\alpha(\theta), \beta(\theta)) : 0 \le \theta < 2\pi\},$$

where α, β are infinitely differentiable and $\alpha'(\theta)^2 + \beta'(\theta)^2 \neq 0$. We aim to solve the Dirichlet problem on Ω

$$\begin{cases} \Delta u(x,y) := u_{xx}(x,y) + u_{yy}(x,y) = 0 & (x,y) \in \Omega, \\ u(x,y) = g(x,y) & (x,y) \in \partial\Omega. \end{cases}$$

Consider the distribution $G(\xi,\eta) = G(\xi,\eta;x,y)$ for fixed $(x,y) \in \Omega$ that solves

$$\begin{cases} \Delta G(\xi,\eta;x,y) = \delta(\xi - x, \eta - y), & (\xi,\eta) \in \Omega, \\ G(\xi,\eta;x,y) = 0 & (\xi,\eta) \in \partial\Omega. \end{cases}$$

Here the Laplacian is with respect to the variables (ξ,η). Recall Green's second identity

$$\int_\Omega [\psi(\xi,\eta)\Delta\phi(\xi,\eta) - \phi(\xi,\eta)\Delta\psi(\xi,\eta)]\, \mathrm{d}x\mathrm{d}y = \oint_{\partial\Omega} [\psi(\xi,\eta)\nabla\phi(\xi,\eta) - \phi(\xi,\eta)\nabla\psi(\xi,\eta)] \cdot \mathrm{d}S$$

We choose $\psi(\xi,\eta) = u(\xi,\eta)$ and $\phi = G$ to find

$$u(x,y) = \oint_{\partial\Omega} u(\xi,\eta)\nabla G(\xi,\eta) \cdot \mathrm{d}S = \oint_{\partial\Omega} g(\xi,\eta)\nabla G(\xi,\eta) \cdot \mathrm{d}S.$$

This gives a solution expression! Since we need to, in the end, be very precise about things

$$\oint_{\partial\Omega} g(x,y)\nabla\psi(x,y) \cdot \mathrm{d}S = \int_0^{2\pi} g(\alpha(\theta),\beta(\theta)) \left[\psi_x(\alpha(\theta),\beta(\theta))\beta'(\theta) - \psi_y(\alpha(\theta),\beta(\theta))\alpha'(\theta)\right] \mathrm{d}t.$$

Example 5.7. Suppose $\Omega = \{(x,y) : x^2 + y^2 < 1\}$. Then

$$\tilde{G}(\xi,\eta;x,y) = \frac{1}{2\pi} \log\sqrt{(x-\xi)^2 + (y-\eta)^2},$$

is what is called the *fundamental solution* and it satisfies $\Delta\tilde{G} = \delta(x - \xi, y - \eta)$. But it fails to satisfy the boundary condition. But, by reflecting the point (x,y) outside the disk via

$$(x',y') = (x,y)/(\sqrt{x^2+y^2}),$$

we find

$$G(\xi, \eta; x, y) = \tilde{G}(\xi, \eta; x, y) - \tilde{G}(\xi, \eta; x', y').$$

This construction will fail for a general domain. So, we seek a solution of the form

$$u(x, y) = \oint_{\partial\Omega} \mu(\xi, \eta) \nabla \tilde{G}(\xi, \eta) \cdot \mathrm{d}S.$$

Then, it turns out that for $(x, y) \in \partial\Omega$

$$\lim_{(x',y') \to (x,y)} \oint_{\partial\Omega} \mu(\xi, \eta) \nabla \tilde{G}(\xi, \eta; x', y') \cdot \mathrm{d}S = \frac{\mu(x, y)}{2} + \oint_{\partial\Omega} \mu(\xi, \eta) \nabla \tilde{G}(\xi, \eta; x, y) \cdot \mathrm{d}S.$$

We now find a second-kind boundary integral equation that $\mu(x, y)$ should solve

$$\frac{\mu(x, y)}{2} + \oint_{\partial\Omega} \mu(\xi, \eta) \nabla \tilde{G}(\xi, \eta; x, y) \cdot \mathrm{d}S = g(x, y), \quad (x, y) \in \partial\Omega.$$

Thus, we have converted a fundamentally two-dimensional problem to a one-dimensional problem!

5.2.1 ▪ Discretization of the second-kind integral equation

We first compute

$$\nabla \tilde{G}(\xi, \eta; x, y) = \left(\frac{1}{2\pi} \frac{\xi - x}{(\xi - x)^2 + (\eta - y)^2}, \frac{1}{2\pi} \frac{\eta - y}{(\xi - x)^2 + (\eta - y)^2} \right).$$

Then, setting

$$\begin{aligned} \xi &= \alpha(\phi), \quad \eta = \beta(\phi), \\ x &= \alpha(\theta), \quad y = \beta(\theta), \\ k(\theta, \phi) &= \nabla \tilde{G}(\xi, \eta; x, y) \cdot (\beta'(\theta), -\alpha'(\theta)), \\ \mu(\theta) &= \mu(x, y), \\ g(\theta) &= g(x, y), \end{aligned}$$

we have the parameterized integral equation

$$\frac{1}{2}\mu(\theta) + \int_0^{2\pi} k(\theta, \phi) \mu(\phi) \mathrm{d}\phi = g(\theta).$$

We should be concerned about the smoothness of $k(\theta, \phi)$ because it looks that the denominator could vanish but this is not of concern:

$$k(\theta, \phi) = \frac{1}{2\pi} \frac{\beta'(\phi)(\alpha(\phi) - \alpha(\theta)) - \alpha'(\phi)(\beta(\phi) - \beta(\theta))}{(\alpha(\phi) - \alpha(\theta))^2 + (\beta(\phi) - \beta(\theta))^2},$$

Therefore using $\alpha(\phi) = \alpha(\theta) + \alpha'(\theta)(\phi - \theta) + \alpha''(\theta)(\phi - \theta)^2/2 \cdots$ and similarly for β as $t \to s$

$$\begin{aligned} k(\theta, \phi) &\sim \frac{1}{2\pi} \frac{\beta'(\phi)\alpha'(\phi)(\phi - \theta) - \alpha'(\phi)\beta'(\phi)(\phi - \theta)}{(\alpha'(\phi)(\theta - \phi))^2 + (\beta'(\phi)(\theta - \phi))^2} \\ &\quad + \frac{1}{4\pi} \frac{\beta'(\phi)\alpha''(\phi) - \alpha'(\phi)\beta''(\phi)}{\alpha'(\phi)^2 + \beta'(\phi)^2} + \cdots, \\ k(\theta, \theta) &= \frac{1}{4\pi} \frac{\beta'(\theta)\alpha''(\theta) - \alpha'(\theta)\beta''(\theta)}{\alpha'(\theta)^2 + \beta'(\theta)^2}, \end{aligned}$$

so that if $\partial\Omega$ is smooth then so is $k(\theta, \phi)$.

We consider a collocation approach. Let

$$\check{\theta}_\ell = L\frac{\ell - 1}{N}, \quad \ell = 1, 2, \dots, N,$$

be the collocation nodes. Set $\mu_\ell \approx \mu(\check{\theta}_\ell)$ and then

$$\frac{1}{2}\mu_\ell + \frac{L}{N}\sum_{j=1}^{N} k(\check{\theta}_\ell, \check{\theta}_j)\mu_j = g(\check{\theta}_\ell), \quad \ell = 1, 2, \dots, N.$$

In matrix-vector notation

$$\frac{1}{2}\begin{bmatrix}\mu_\ell\end{bmatrix}_{\ell=1}^{N} + \frac{L}{N}\underset{\ell=1 \qquad\qquad\qquad\qquad N}{\begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & k(\check{\theta}_\ell, \check{\theta}_j) & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix}_{j=1}^{N}} \begin{bmatrix}\mu_\ell\end{bmatrix}_{\ell=1}^{N} = \begin{bmatrix}g(\check{\theta}_\ell)\end{bmatrix}_{\ell=1}^{N}.$$

5.2.2 ▪ The solution for star-shaped domains

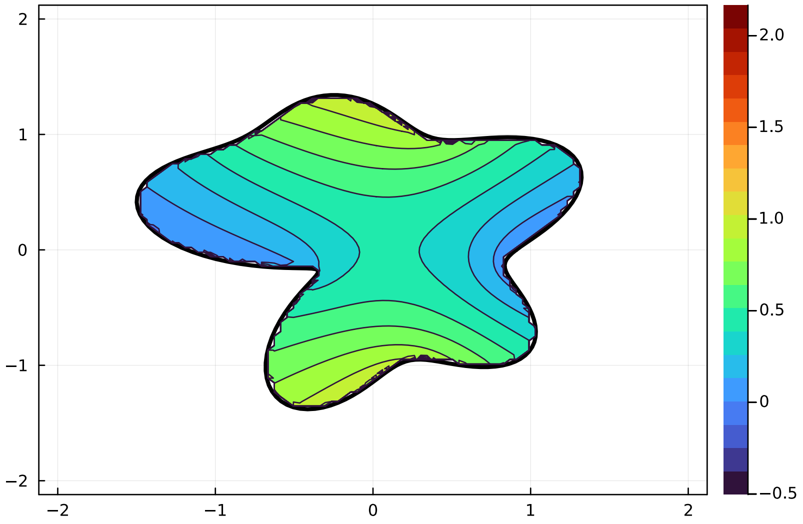
A domain $\Omega \subset \mathbb{R}^2$ is said to be star shaped if its boundary can be given by a smooth parameterization

$$\partial\Omega = \{(r(\theta)\cos(\theta), r(\theta)\sin(\theta)) : 0 \le \theta \le 2\pi\}, \quad r(\theta) > 0.$$

Here we impose that the radius is given as a function of the angle t. So, we just consider when

$$r(\theta) = 1 + \sum_{j=-M}^{M} c_j\, \mathrm{e}^{\mathrm{i}j\theta}, \quad c_j = \bar{c}_{-j},$$

is positive.



Chapter 6

Periodic differential operators

In this chapter, we consider the numerical solution of two problems related to operators of the form

$$\mathcal{L}u(\theta) = \underbrace{\sum_{j=q}^{k} c_j \frac{\mathrm{d}^j u}{\mathrm{d}\theta^j}(\theta)}_{\mathcal{L}_0 u} + \underbrace{\sum_{j=0}^{p} a_j(\theta) \frac{\mathrm{d}^j u}{\mathrm{d}\theta^j}(\theta)}_{\mathcal{L}_1 u}, \quad \begin{array}{l} a_j \in C^\infty(\mathbb{T}), \quad j = 0, 1, \ldots, p, \\ c_j \in \mathbb{C}, \quad j = p+1, \ldots, k, \quad c_k \neq 0. \end{array} \tag{6.1}$$

Specifically, we consider the

- approximation of solutions of $\mathcal{L}u = f$, and
- the approximation of the eigenvalues of $\mathcal{L}$.

6.1 ▪ Theoretical foundations

By the nature of Fourier series, we can express operators $\mathcal{L}$ as bi-infinite matrices with entries

$$\mathbf{L} = (\ell_{jk})_{-\infty < j,k < \infty}, \quad \ell_{jk} = \frac{1}{2\pi} \langle \mathcal{L}\, \mathrm{e}^{\mathrm{i}k\diamond}, \mathrm{e}^{\mathrm{i}j\diamond} \rangle,$$

provided the Fourier basis functions are in $\mathcal{D}(\mathcal{L})$. We keep the convention that second index, k in this case, is the column index and should increase from left-to-right, and first index, j in this case, is the row index and should increase from top to bottom. Then, we need to understand the multiplication operator $\mathcal{M}(g)f = gf$. Suppose $g \in C^\infty(\mathbb{T})$ and consider

$$\mathbf{M}(g) = \left(\frac{1}{2\pi} \int_0^{2\pi} g(\theta)\, \mathrm{e}^{\mathrm{i}(k-j)\theta}\, \mathrm{d}\theta \right)_{-\infty < j,k < \infty} = \frac{1}{\sqrt{2\pi}} \left(c_{j-k}(g) \right)_{-\infty < j,k < \infty}.$$

This implies that $\mathbf{M}(g)$ is a Toeplitz operator — the entries are constant down each diagonal:

$$\mathbf{M}(g)\mathbf{u} := \left[\begin{array}{ccc|ccc} \ddots & \ddots & \ddots & & & \iddots \\ \ddots & g_0 & g_{-1} & g_{-2} & & \\ \ddots & g_1 & g_0 & g_{-1} & g_{-2} & \\ \hline & g_2 & g_1 & g_0 & g_{-1} & \ddots \\ & & g_2 & g_1 & g_0 & \ddots \\ \iddots & & & \ddots & \ddots & \ddots \end{array} \right] \left[\begin{array}{c} \vdots \\ u_{-2} \\ u_{-1} \\ \hline u_0 \\ u_1 \\ \vdots \end{array} \right], \quad g_j = \frac{1}{\sqrt{2\pi}} c_j(g).$$

This matrix is blocked so that when the row just below the horizontal line multiplies the vector it gives the coefficient of the zeroth mode. And in this chapter we will use calligraphic and boldface fonts interchangeably for operators where the boldface emphasizes its matrix representation in this standard orthonormal basis.

Proposition 6.1. *Suppose and $g \in H^{\max\{1/2+\epsilon,k\}}(\mathbb{T})$ for any $\epsilon > 0$. Then $\mathcal{M}(g)$ is a bounded linear operator on $H^k(\mathbb{T})$, $k = 0, 1, 2, \dots$.*

Proof. Suppose $k = 0$. Then we have the g is continuous, and therefore bounded. The claim follows. Now, for $k \geq 1$, we see that we get a weighted sum of the columns of $\mathbf{M}(g)$:

$$\left\langle gf, \frac{\mathrm{e}^{\mathrm{i}j\diamond}}{\sqrt{2\pi}} \right\rangle = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} c_k(f) c_{j-k}(g).$$

Wwe then note that

$$(1 + |j|)^{2k} \leq (1 + |j - \ell| + 1 + |\ell|)^{2k} \leq c(1 + |j - \ell|)^{2k} + c(1 + |\ell|)^{2k},$$

where c is a constant such that $(|x| + |y|)^{2k} \leq c|x|^{2k} + c|y|^{2k}$. Then

$$\begin{aligned} \|gf\|_{H^k}^2 &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} (1 + |j|)2k \left| \sum_{\ell=-\infty}^{\infty} c_\ell(f) c_{j-\ell}(g) \right|^2 \\ &\leq \frac{c}{2\pi} \sum_{j=-\infty}^{\infty} \left| \sum_{\ell=-\infty}^{\infty} (1 + |\ell|)^{2\ell} c_\ell(f) c_{j-\ell}(g) \right|^2 \\ &+ \frac{c}{2\pi} \sum_{j=-\infty}^{\infty} \left| \sum_{\ell=-\infty}^{\infty} (1 + |\ell - j|)^{2\ell} c_\ell(f) c_{j-\ell}(g) \right|^2 \end{aligned}$$

Then

$$\|g\tilde{f}\|_{L^2}^2 = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \left| \sum_{\ell=-\infty}^{\infty} (1 + |\ell - j|)^{2\ell} c_\ell(f) c_{j-\ell}(g) \right|^2$$

, where $\tilde{f}$ has Fourier coefficients $c_j(\tilde{f}) = (1 + |j|)^k c_j(f)$, and $\tilde{f} \in L^2$. And then because $\|g\|_\infty \leq C_k \|g\|_{H^k}$, we have

$$\|g\tilde{f}\|_{L^2} \leq C_k \|g\|_{H^k} \|\tilde{f}\|_{L^2} = \|g\|_{H^k} \|f\|_{H^k}.$$

By interchanging f, g the bound on the other term follows and the proposition also follows. ■

We arrive at an important theorem, the proof of which is helpful in the sequel.

Theorem 6.2. *Suppose $\mathcal{L}$ is given by (6.1) with domain $H^k(\mathbb{T})$. Then $\mathcal{L}$ is closed and has purely discrete spectrum with no accumulation points in the finite complex plane.*

Proof. It follows that $\mathcal{L}_0' = c_k \frac{\mathrm{d}^k}{\mathrm{d}\theta^k}$ is closed with $D(\mathcal{L}_0') = H^k(\mathbb{T})$. Then it follows that if we decompose, $\mathcal{L} = \mathcal{L}_0' + \mathcal{L}_1'$, then $\mathcal{L}_1'$ is a bounded operator from $H^k(\mathbb{T})$ to $H^1(\mathbb{T})$. Thus

$$\mathcal{L}_1'(\mathcal{L}_0' - z)^{-1}$$

is compact. By inspection $\mathcal{L}_0'$ has purely discrete spectrum $c_k(\mathrm{i}j)^k$, $j = 0, \pm 1, \pm 2, \ldots$. And then for any $\rho > 0$, $z_\rho = c_k \mathrm{i}^{k+1}\rho \in \rho(\mathcal{L}_0')$:

$$|c_k(\mathrm{i}j)^k - c_k\mathrm{i}^{k+1}\rho|^2 = |c_k|^2|j^k - \mathrm{i}\rho|^2 = |c_k|^2(|j|^{2k} + \rho^2).$$

Then we consider

$$\mathcal{L}_1'(z_\rho - \mathcal{L}_0')^{-1}.$$

Let $\mathcal{N} : H^k(\mathbb{T}) \to H^{k-1}(\mathbb{T})$ be defined by

$$\sum_j c_j\, \mathrm{e}^{\mathrm{i}j\theta} \to \sum_j (1 + |j|) c_j\, \mathrm{e}^{\mathrm{i}j\theta}\,.$$

Then we note that $\mathcal{L}_1'\mathcal{N}$ is bounded from $H^k(\mathbb{T})$ to $L^2(\mathbb{T})$. And then

$$\mathcal{N}^{-1}(z_\rho - \mathcal{L}_0')^{-1}$$

is bounded from $L^2(\mathbb{T})$ to $H^k(\mathbb{T})$ with norm bound

$$\begin{aligned}\|\mathcal{N}^{-1}(z_\rho - \mathcal{L}_0')^{-1}\|_{L^2\to H^k} &= \max_{j\in\mathbb{Z}} \frac{1}{|c_k|}\frac{(1+|j|)^{k-1}}{\sqrt{|j|^{2k}+\rho^2}} \\ &\leq \max\left\{\max_{j\neq 0} \frac{2}{|c_k|}\frac{1}{\sqrt{|j|^2 + \rho^2/|j|^{2k-2}}}, \frac{1}{|c_k|\rho}\right\}.\end{aligned}$$

For every $\epsilon > 0$, $\rho > 0$, there exists $j_0 = j_0(\epsilon)$ such that

$$\frac{2}{|c_k|}\frac{2}{\sqrt{|j|^2 + \rho^2/|j|^{2k-2}}} < \epsilon,$$

for all $|j| > j_0$. And then there exists $\rho_0 = \rho_0(j_0)$ such that

$$\frac{2}{|c_k|}\frac{2}{\sqrt{|j|^2 + \rho^2/|j|^{2k-2}}} < \epsilon,$$

for all $|j| < j_0$ if $\rho > \rho_0$. So, for $\rho > \rho_0$

$$\|\mathcal{N}^{-1}(z_\rho - \mathcal{L}_0')^{-1}\|_{L^2\to H^k} < \epsilon.$$

We can then apply Theorem 4.44 to establish the theorem. ■

6.2 ▪ The finite-section method

We now present the finite-section method, which is a type of Galerkin method (more precisely, a Bubnov–Galerkin method[4]). We call it the finite-section method because rather than working out the matrix entries as one does in a Galerkin method, using inner products, it is easier to write the whole system as a bi-infinite linear system and then take principal subblocks of it as our approximation.

Define the matrix representation of the derivative operator

$$\mathbf{D} = (\mathrm{i} j \delta_{jk})_{-\infty < j,k < \infty}.$$

Proposition 6.3. *The operator* $\mathbf{D}^\ell$ *is bounded from* $H^k(\mathbb{T})$ *to* $H^{k-\ell}(\mathbb{T})$*.*

We arrive at the representation of (6.1) in Fourier space,

$$\mathbf{L} = \underbrace{\sum_{j=q}^{k} c_j \mathbf{D}^j}_{\mathbf{L}_o} + \underbrace{\sum_{j=0}^{p} \mathbf{M}(a_j)\mathbf{D}^j}_{\mathbf{L}_r}$$

where we suppose that we can compute the exact Fourier coefficients for a_j. In practice, one has to resort to an approximation, and we know the DFT is the right tool for this. We look to understand how to approximate solutions of

$$\mathcal{L}u = f \quad \Leftrightarrow \quad \mathbf{L}\mathbf{u} = \mathbf{f},$$

where $\mathbf{u} = (c_j(u))_{j=-\infty}^{\infty}$, $\mathbf{f} = (c_j(f))_{j=-\infty}^{\infty}$ for a given right-hand side function $f \in L^2(\mathbb{T})$.

Recall the Fourier projection operator $\mathcal{P}_N$. Here we use $\mathbf{P}_N$ to denote its matrix representation:

$$\mathbf{P}_N = \left[\begin{array}{ccc|ccc} \ddots & \ddots & \ddots & & & \iddots \\ \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} & & \\ \ddots & \mathbf{0} & \mathbf{I}_- & \mathbf{0} & \mathbf{0} & \\ \hline & \mathbf{0} & \mathbf{0} & \mathbf{I}_+ & \mathbf{0} & \ddots \\ & & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots \\ \iddots & & & \ddots & \ddots & \ddots \end{array}\right],$$

where $\mathbf{I}_+, \mathbf{I}_-$ are identity matrices of sizes $N_+ + 1$ and N_-, respectively.

And now, we consider truncation, setting $V_N = R(\mathbf{P}_N) \subset \ell^2(\mathbb{Z})$ and solving

$$\mathbf{P}_N \mathbf{L} \mathbf{u}_N = \mathbf{P}_N \mathbf{f}, \quad \mathbf{u}_N \in V_N.$$

If the row index j lies in $(-\infty, -N_-) \cup (N_+, \infty)$, then the equation is trivially satisfied. Furthermore, since $\mathbf{u}_N \in V_N$, only columns k for $k = -N_-, \ldots, N_+$ contribute. So, define

$$\tilde{\mathbf{P}}_N = \left[\begin{array}{cccc|cccc} \cdots & \mathbf{0} & \mathbf{0} & \mathbf{I}_- & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \hline \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_+ & \mathbf{0} & \mathbf{0} & \cdots \end{array}\right],$$

[4] For a Bubnov–Galerkin method, the same set of orthogonal basis functions are used for both the domain and range.

and the finite-section system is

$$\mathbf{P}_N \mathbf{L} \mathbf{u}_N = \mathbf{P}_N \mathbf{f}, \quad \mathbf{u}_N \in V_N \tag{6.2}$$

if and only if

$$\mathbf{L}_N \tilde{\mathbf{u}}_N = \tilde{\mathbf{P}}_N \mathbf{f}, \quad \mathbf{L}_N = \tilde{\mathbf{P}}_N \mathbf{L} \tilde{\mathbf{P}}_N^T, \quad \mathbf{u}_N = \tilde{\mathbf{P}}_N^T \tilde{\mathbf{u}}_N. \tag{6.3}$$

Our goal is to then establish that $\mathbf{u}_N \to \mathbf{u}$, and obtain estimates on the rate of convergence.

The first observation is similar to one that was made in the analysis of the Nyström collocation method for Fredholm integral equations. If $\tilde{\mathbf{u}}_N \in V_N$ is a solution of $\mathbf{L}_N \tilde{\mathbf{u}}_N = \tilde{\mathbf{P}}_N \mathbf{f}$ then $\mathbf{u}_N = \tilde{\mathbf{P}}_N^T \tilde{\mathbf{u}}_N$ is a solution of $\mathbf{P}_N \mathbf{L} \mathbf{u}_N = \mathbf{P}_N \mathbf{f}$. And then $\mathbf{u}_N$ is a solution of

$$\mathbf{L}_o \mathbf{u}_N + \mathbf{P}_N \mathbf{L}_r \mathbf{u}_N = (\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r) \mathbf{u}_N = \mathbf{P}_N \mathbf{f}. \tag{6.4}$$

This is because $\mathbf{P}_N \mathbf{L}_o = \mathbf{L}_o \mathbf{P}_N$. We refer to

- (6.2) as the finite-section system,
- (6.3) as the reduced finite-section system (this is the system actually constructed in a numerical method), and
- (6.4) as the extended finite-section system as it is an operator acting on the entire domain of $\mathbf{L}$.

Similarly, if $\mathbf{u}_N$ is a solution of (**??**), then we write

$$\begin{aligned} \mathbf{L}_o \mathbf{u}_N &= \mathbf{P}_N \mathbf{f} - \mathbf{P}_N \mathbf{L}_r \mathbf{u}_N, \\ \mathbf{P}_N \mathbf{L}_o \mathbf{u}_N + \mathbf{L}_o (\mathrm{id} - \mathbf{P}_N) \mathbf{u}_N &= \mathbf{P}_N \mathbf{f} - \mathbf{P}_N \mathbf{L}_r \mathbf{u}_N, \\ \mathbf{L}_o (\mathrm{id} - \mathbf{P}_N) \mathbf{u}_N &= \mathbf{0}, \end{aligned}$$

where we applied $\mathrm{id} - \mathbf{P}_N$ on both sides to obtain the last relation. But because $N(\mathbf{L}_o)$ is a subset of V_N, for N sufficiently large, we find $(\mathrm{id} - \mathbf{P}_N)\mathbf{u}_N \in V_N$ and therefore $(\mathrm{id} - \mathbf{P}_N)\mathbf{u}_N = \mathbf{0}$, and $\mathbf{u}_N \in V_N$. Thus $\mathbf{u}_N$ is a solution of $\mathbf{P}_N \mathbf{L} \mathbf{u}_N = \mathbf{P}_N \mathbf{f}$ and $\tilde{\mathbf{u}}_N = \tilde{\mathbf{P}}_N^T \mathbf{u}_N$ is a solution of $\mathbf{L}_N \tilde{\mathbf{u}}_N = \tilde{\mathbf{P}}_N \mathbf{f}$. Thus, for N sufficiently large, one of these linear systems is uniquely solvable if and only if they all are.

We need to show that $\mathbf{L}_N$ is non-singular, and obtain a bound on the inverse norm. We *cannot* hope that

$$\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r \to \mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r,$$

in norm because (1) the operators are unbounded so norms are not finite, and (2) truly unbounded operators are not compact so we cannot hope to approximate them with finite-rank operators. So, we need to look for a way to apply the existing theory differently.

Take $z \in \rho(\mathbf{L}_o)$ and if we consider the same finite-section method applied to the operator

$$(\mathbf{L}_o - z)^{-1} (\mathbf{L}_o + \mathbf{L}_r) = \mathrm{id} + (\mathbf{L}_0 - z)^{-1} (z + \mathbf{L}_r),$$

this is indeed of form $\mathrm{id} + \mathbf{K}$ where $\mathbf{K}$ is compact on $H^k(\mathbb{T})$. And we consider what our finite-section, reduced finite-section and extended finite-section systems look like, respectively:

$$\begin{aligned} \mathbf{P}_N (\mathrm{id} + (\mathbf{L}_0 - z)^{-1} (z + \mathbf{L}_r)) \mathbf{u}_N &= \mathbf{P}_N (\mathbf{L}_0 - z)^{-1} \mathbf{f}, \quad \mathbf{u}_N \in V_N, \\ \tilde{\mathbf{P}}_N (\mathrm{id} + (\mathbf{L}_0 - z)^{-1} (z + \mathbf{L}_r)) \tilde{\mathbf{u}}_N &= \tilde{\mathbf{P}}_N (\mathbf{L}_0 - z)^{-1} \mathbf{f}, \\ (\mathrm{id} + \mathbf{P}_N (\mathbf{L}_0 - z)^{-1} (z + \mathbf{L}_r)) \mathbf{u}_N &= \mathbf{P}_N (\mathbf{L}_0 - z)^{-1} \mathbf{f}. \end{aligned}$$

And we have the same equivalency between these systems.

But next, we note that $\mathbf{P}_N = \tilde{\mathbf{P}}_N^T \tilde{\mathbf{P}}_N$ commutes with the diagonal operator $(\mathbf{L}_0 - z)^{-1}$ and, further, we can write

$$\tilde{\mathbf{P}}_N(\mathbf{L}_0 - z)^{-1} = \tilde{\mathbf{P}}_N \mathbf{P}_N(\mathbf{L}_0 - z)^{-1} = \tilde{\mathbf{P}}_N(\mathbf{L}_0 - z)^{-1}\mathbf{P}_N = \tilde{\mathbf{P}}_N(\mathbf{L}_0 - z)^{-1}\tilde{\mathbf{P}}_N^T\tilde{\mathbf{P}}_N.$$

The three systems become, respectively,

$$\begin{aligned}(\operatorname{id} + (\mathbf{L}_0 - z)^{-1}\mathbf{P}_N(z + \mathbf{L}_r))\mathbf{u}_N &= (\mathbf{L}_0 - z)^{-1}\mathbf{P}_N\mathbf{f}, \quad \mathbf{u}_N \in V_N,\\ (\operatorname{id} + \tilde{\mathbf{P}}_N(\mathbf{L}_0 - z)^{-1}\tilde{\mathbf{P}}_N^T\tilde{\mathbf{P}}_N(z + \mathbf{L}_r))\tilde{\mathbf{u}}_N &= \tilde{\mathbf{P}}_N(\mathbf{L}_0 - z)^{-1}\tilde{\mathbf{P}}_N^T\tilde{\mathbf{P}}_N\mathbf{f},\\ (\operatorname{id} + (\mathbf{L}_0 - z)^{-1}\mathbf{P}_N(z + \mathbf{L}_r))\mathbf{u}_N &= (\mathbf{L}_0 - z)^{-1}\mathbf{P}_N\mathbf{f}.\end{aligned}$$

And from the last, we see this is equivalent to (6.4). So, we use the last equation to analyze the convergence of $\mathbf{u}_N$. And we use (6.3) to solve, numerically, for $\mathbf{u}_N$.

6.2.1 ▪ Convergence

To demonstrate convergence we abstract the system as

$$\begin{aligned}(\operatorname{id} + (\mathbf{L}_0 - z)^{-1}\mathbf{P}_N(z + \mathbf{L}_r))\mathbf{u}_N &= (\mathbf{L}_0 - z)^{-1}\mathbf{P}_N\mathbf{f},\\ (\operatorname{id} + \mathbf{P}_N\mathbf{K}(z))\mathbf{u}_N &= \mathbf{P}_N\mathbf{g}(z).\end{aligned}$$

We first need to show that $\operatorname{id} + \mathbf{P}_N\mathbf{K}(z) \to \operatorname{id} + \mathbf{K}(z)$ in operator norm on $H^k(\mathbb{T})$.

Proposition 6.4. *Suppose $\mathbf{L}$ is invertible on $L^2(\mathbb{T})$. Then there exists a constant $C_{k,p}$, depending only on k, p, such that if N satisfies*

$$N^{k-p} \geq 2C_{k,p}\|\mathbf{L}_0 - z\|_{H^k \to L^2}\|\mathbf{L}^{-1}\|_{L^2 \to H^k}\|(\mathbf{L}_0 - z)^{-1}\|_{L^2 \to H^k}\|z + \mathbf{L}_r\|_{H^k \to H^{k-p}},$$

then the linear systems for $\mathbf{u}_N, \tilde{\mathbf{u}}_N$ are uniquely solvable and $\mathbf{u}_N$ satisfies

$$C^{-1}\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{H^k} \leq \|\mathbf{u} - \mathbf{u}_N\|_{H^k} \leq C\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{H^k},$$

and $C \leq 2\|\mathbf{L}_0 - z\|_{H^k \to L^2}\|\mathbf{L}^{-1}\|_{L^2 \to H^k}$.

Proof. We have that

$$\operatorname{id} + \mathbf{K}(z) - (\operatorname{id} + \mathbf{P}_N\mathbf{K}(z)) = (\operatorname{id} - \mathbf{P}_N)\mathbf{K}(z).$$

Breaking apart $\mathbf{K}(z)$ we have

$$(\mathbf{L}_0 - z)^{-1}(z + \mathbf{L}_r).$$

We then see that $\mathbf{L}_0 - z$ is an injective, surjective operator from $H^k(\mathbb{T})$ to $L^2(\mathbb{T})$ and the open mapping theorem implies that the inverse is bounded from $L^2(\mathbb{T})$ to $H^k(\mathbb{T})$. Further, $z + \mathbf{L}_r$ is bounded from $H^k(\mathbb{T})$ to $H^{k-p}(\mathbb{T})$. Then we write

$$(\mathbf{L}_0 - z)^{-1}(z + \mathbf{L}_r) - \mathbf{P}_N(\mathbf{L}_0 - z)^{-1}(z + \mathbf{L}_r) = (\mathbf{L}_0 - z)^{-1}(\operatorname{id} - \mathbf{P}_N)(z + \mathbf{L}_r),$$

and this gives

$$\|(\operatorname{id} - \mathbf{P}_N)\mathbf{K}(z)\|_{H^k \to H^k} \leq \|(\mathbf{L}_0 - z)^{-1}\|_{L^2 \to H^k}\|\operatorname{id} - \mathbf{P}_N\|_{H^{k-p} \to L^2}\|z + \mathbf{L}_r\|_{H^k \to H^{k-p}}.$$

From Theorem 2.3, we have that

$$\|\operatorname{id} - \mathbf{P}_N\|_{H^{k-p}\to L^2} \le C_{k,p} N^{-k+p}.$$

Theorem 4.8 implies that $\operatorname{id} + \mathbf{P}_N\mathbf{K}(z)$ is invertible if N is such that

$$\|(\operatorname{id} - \mathbf{P}_N)\mathbf{K}(z)\|_{H^k\to H^k} \|(\operatorname{id} + \mathbf{K}(z))^{-1}\|_{H^k} < 1/2.$$

For this, it suffices that $N > N_0$ where N_0 satisfies

$$N_0^{k-p} \ge 2C_{k,p}\|\mathbf{L}_0 - z\|_{H^k\to L^2}\|\mathbf{L}^{-1}\|_{L^2\to H^k}\|(\mathbf{L}_0 - z)^{-1}\|_{L^2\to H^k}\|z + \mathbf{L}_r\|_{H^k\to H^{k-p}}.$$

And then

$$\|(\operatorname{id} + \mathbf{P}_N\mathbf{K}(z))^{-1}\|_{H^k} \le 2\|\mathbf{L}_0 - z\|_{H^k\to L^2}\|\mathbf{L}^{-1}\|_{L^2\to H^k}, \quad N > N_0.$$

This implies all of our linear systems for $\mathbf{u}_n, \tilde{\mathbf{u}}_N$ have a unique solution.

Then, following (5.2), we have

$$(\operatorname{id} + \mathbf{P}_N\mathbf{K}(z))(\mathbf{u} - \mathbf{u}_N) = (\mathbf{u} - \mathbf{P}_N\mathbf{u}).$$

and therefore there exists a constant $C \ge 1$ such that

$$C^{-1}\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{H^k} \le \|\mathbf{u} - \mathbf{u}_N\|_{H^k} \le C\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{H^k}$$

■

To obtain estimates directly on $\|\mathbf{u} - \mathbf{u}_N\|_{L^2}$ it is convenient to introduce Sobolev spaces of negative index.

Definition 6.5. *The Sobolev space $H^s(\mathbb{T})$, $s < 0$, is given by*

$$H^s(\mathbb{T}) := \left\{ \mathbf{f} = (f_j)_{-\infty<j<\infty} \mid \sum_{j=-\infty}^{\infty} |f_j|^2(1+|j|)^{2s} < \infty \right\},$$

and the norm is given by

$$\|\mathbf{f}\|_{H^s}^2 := \sum_{j=-\infty}^{\infty} |f_j|^2(1+|j|)^{2s}.$$

The following is then immediate.

Proposition 6.6. *Suppose $\mathbf{L}$ is invertible on $H^{-k}(\mathbb{T})$. Then there exists a constant $C_{k,p}$, depending only on k,p, such that if*

$$N^{k-p} \ge 2C_{k,p}\|\mathbf{L}_0 - z\|_{L^2\to H^{-k}}\|\mathbf{L}^{-1}\|_{H^{-k}\to L^2}\|(\mathbf{L}_0 - z)^{-1}\|_{H^{-k}\to L^2}\|z + \mathbf{L}_r\|_{L^2\to H^{p-k}},$$

then the linear systems for $\mathbf{u}_N, \tilde{\mathbf{u}}_N$ are uniquely solvable and $\mathbf{u}_N$ satisfies

$$C^{-1}\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{L^2} \le \|\mathbf{u} - \mathbf{u}_N\|_{L^2} \le C\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{L^2},$$

and $C \leq 2\|\mathbf{L}_0 - z\|_{L^2\to H^{-k}}\|\mathbf{L}^{-1}\|_{H^{-k}\to L^2}$.

This is useful now because since $\mathbf{L}$ is invertible, we know that if $\mathbf{f} \in L^2(\mathbb{T})$ then $\mathbf{u} = \mathbf{L}^{-1}\mathbf{f} \in H^k(\mathbb{T})$ and therefore

$$\|\mathbf{u} - \mathbf{P}_N\mathbf{u}\|_{L^2} = O(N^{-k}).$$

Furthermore, if $f \in H^\ell(\mathbb{T})$, then the relation

$$\mathbf{u} = -(\mathbf{L}_0 - z)^{-1}\left[(z + \mathbf{L}_r)\mathbf{u} + \mathbf{f}\right], \tag{6.5}$$

implies that $\mathbf{u} \in H^{\ell+k-p}(\mathbb{T})$ and — the smoother the data, the faster the convergence.

6.3 ▪ Spectrum approximation

This section is devoted to understand in what sense the eigenvalues of the matrix $\mathbf{L}_N$ approximate those of $\mathbf{L}$. This question complicated for many reasons. First, as N increases, $\mathbf{L}_N$ will gain eigenvalues (counted according to multiplicity) and, not only that, the existing eigenvalues will change. Furthermore, the spectrum of $\mathbf{L}$ is unbounded. So, we cannot hope to approximate all the eigenvalues of $\mathbf{L}$ in finite time!

Definition 6.7. *Define the* spectral supremum *of a sequence of operators* $(\mathcal{L}_N)_{N\geq 1}$ *as*

$$\operatorname*{spec\,sup}_{N\to\infty} \mathcal{L}_N = \bigcap_{M\geq 1} \overline{\bigcup_{N\geq M} \sigma(\mathcal{L}_N)}$$

Lemma 6.8. $\lambda \in \operatorname{spec\,sup}_{N\to\infty}\mathcal{L}_N$ *if and only if there exists a sequence* $(\lambda_N)_{N\geq 1}$, $\lambda_N \in \sigma(\mathcal{L}_N)$ *that has a convergent subsequence that converges to* λ.

Proof. Suppose $(\lambda_N)_{N\geq 1}$ is such a sequence and λ is the subsequential limit. Then

$$\lambda \in \overline{\bigcup_{N\geq M} \sigma(\mathcal{L}_N)},$$

for every M and $\lambda \in \operatorname{spec\,sup}_{N\to\infty}\mathcal{L}_N$. Conversely, suppose $\lambda \in \operatorname{spec\,sup}_{N\to\infty}\mathcal{L}_N$. Then since $\lambda \in \overline{\bigcup_{N\geq M}\sigma(\mathcal{L}_N)}$, for every M, we can find λ_{N_k} in this set such that $|\lambda - \lambda_{N_k}| < 1/k$, $\lambda_{N_k} \in \sigma(\mathcal{L}_{N_k})$, $N_k > N_{k+1}$, and extending λ_{N_k} to a sequence Λ_N is trivial as we have found the subsequence. ■

Definition 6.9. *Define the* spectral infimum *of a sequence of operators* $(\mathcal{L}_N)_{N\geq 1}$ *as*

$$\operatorname*{spec\,inf}_{N\to\infty} \mathcal{L}_N = \bigcap_{\substack{S\subset\mathbb{N}\\ |S|=\infty}} \overline{\bigcup_{N\in S} \sigma(\mathcal{L}_N)}.$$

Here the intersection is over all infinite subsets of $\mathbb{N}$.

Lemma 6.10. *$\lambda \in \text{spec inf}_{N\to\infty} \mathcal{L}_N$ if and only if there exists a sequence $(\lambda_N)_{N\geq 1}$ such that $\lambda_N \in \sigma(\mathcal{L}_N)$ and $\lambda_N \to \lambda$.*

Proof. Let Ω be the set of λ such that there exists a sequence $(\lambda_N)_{N\geq 1}$ such that $\lambda_N \in \sigma(\mathcal{L}_N)$ and $\lambda_N \to \lambda$.

Suppose there does exist a sequence of λ_N's with λ being the limit, $\lambda \in \Omega^c$. Then there exists $\epsilon > 0$ that for every N, there exists $N' > N$ such that $\text{dist}(\lambda, \sigma(\mathcal{L}_{N'})) > \epsilon$. So, we find an infinite set $S \subset \mathbb{N}$ such that

$$\lambda \notin \overline{\bigcup_{N\in S} \sigma(\mathcal{L}_N)}.$$

Thus Ω^c is a subset of the complement of the spectral infimum:

$$\underset{N\to\infty}{\text{spec inf}}\, \mathcal{L}_N \subset \Omega.$$

Now suppose $\lambda \in \Omega$. For $\epsilon > 0$, every infinite subset $S \subset \mathbb{N}$ has $N \in S$ such that $\text{dist}(\lambda, \sigma(\mathcal{L}_N)) < \epsilon$ and then λ must be in the closure

$$\lambda \in \overline{\bigcup_{N\in S} \sigma(\mathcal{L}_N)}.$$

The claim follows. ■

It is immediate that

$$\underset{N\to\infty}{\text{spec inf}}\, \mathcal{L}_N \subset \underset{N\to\infty}{\text{spec sup}}\, \mathcal{L}_N.$$

Definition 6.11. *We say that the spectrum of $\mathcal{L}_N$ converges to that of $\mathcal{L}$ if*

$$\underset{N\to\infty}{\text{spec inf}}\, \mathcal{L}_N = \underset{N\to\infty}{\text{spec sup}}\, \mathcal{L}_N = \sigma(\mathcal{L}).$$

Theorem 6.12. *Let $\mathbf{L}$ and $\mathbf{L}_N$ be as in the previous section. Then the spectrum of $\mathbf{L}_N$ converges to that of $\mathbf{L}$.*

We claim that this is a direct corollary of the following lemma that we prove below.

Lemma 6.13. *Let $\mathbf{L}$, $\mathbf{L}_o$ and $\mathbf{L}_r$ be as in the previous section. Then the $L^2(\mathbb{T})$ spectrum of $\mathbf{L}_0 + \mathbf{P}_N\mathbf{L}_r$ converges to the $L^2(\mathbb{T})$ spectrum of $\mathbf{L}$.*

Proof of Theorem 6.12. It suffices to show that within any bounded region of $\mathbb{C}$, the eigenvalues of the two operators eventually coincide, as $N \to \infty$. Indeed, let $\mathbf{v} \neq \mathbf{0}$ be such that $\mathbf{L}_N\mathbf{v} = \lambda\mathbf{v}$. Then for

$$\mathbf{w} = \tilde{\mathbf{P}}_N^T \mathbf{v},$$

we have, using $\mathbf{P}_N = \tilde{\mathbf{P}}_N^T \tilde{\mathbf{P}}_N$,

$$(\mathbf{L}_0 + \mathbf{P}_N \mathbf{L}_r)\tilde{\mathbf{P}}_N^T \mathbf{v} = \mathbf{P}_N \mathbf{L} \tilde{\mathbf{P}}_N^T \mathbf{v} = \tilde{\mathbf{P}}_N^T \underbrace{\tilde{\mathbf{P}}_N \mathbf{L} \tilde{\mathbf{P}}_N^T}_{\mathbf{L}_N} \mathbf{v} = \lambda \mathbf{w}.$$

Therefore $\sigma(\mathbf{L}_N) \subset \sigma(\mathbf{L}_0 + \mathbf{P}_N \mathbf{L}_r)$. Next, suppose that

$$(\mathbf{L}_0 + \mathbf{P}_N \mathbf{L}_r)\mathbf{w} = \lambda \mathbf{w}, \quad \mathbf{w} \neq \mathbf{0}.$$

If it was the case that $\mathbf{w} \in R(\mathbf{P}_N)$ then $\lambda \in \sigma(\mathbf{L}_N)$. Write

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_1 = \mathbf{P}_N \mathbf{w} + (\operatorname{id} - \mathbf{P}_N)\mathbf{w}.$$

We obtain the two relations

$$\begin{aligned} \mathbf{L}_0 \mathbf{w}_0 + \mathbf{P}_N \mathbf{L}_r \mathbf{w} &= \lambda \mathbf{w}_0, \\ \mathbf{L}_0 \mathbf{w}_1 &= \lambda \mathbf{w}_1 \end{aligned}$$

From this we see that if $\mathbf{w}_1 \neq 0$, $\mathbf{w}_1$ is an eigenfunction of $\mathbf{L}_0$, that lies in the range of $\operatorname{id} - \mathbf{P}_N$. But since $\mathbf{L}_0$ is diagonal with entries that grow in modulus in both directions along the diagonal. So, $\lambda = O(N^k)$ is necessary for $\mathbf{w}_1$ to be such an eigenfunction. So if $|\lambda| \leq CN^{k-1}$, for N sufficiently large, we have that $\mathbf{w}_1 = \mathbf{0}$ and $\lambda \in \sigma(\mathbf{L}_N)$. ■

So, the real hard work goes into proving Lemma 6.13. And a useful object is the ϵ-pseudospectrum.

Definition 6.14. *Suppose $\mathcal{L} \in L(V, W)$ such that $V \subset W$ and V is continuously embedded in W. Then the ϵ-pseudospectrum is defined by*

$$\sigma_\epsilon(\mathcal{L}) = \{z \in \mathbb{C} \mid \|(z - \mathcal{L})^{-1}\|_{W \to V} \geq \epsilon^{-1}\},$$

with the convention that $\|(z - \mathcal{L})^{-1}\|_{W \to V} = \infty$ if $z - \mathcal{L}$ is not invertible.

Proof of Lemma 6.13. Suppose that $z \notin \sigma_\epsilon(\mathbf{L})$ where, just to be clear,

$$\mathbf{L} \in L(H^k(\mathbb{T}), L^2(\mathbb{T})).$$

We consider

$$\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z = \mathbf{L} - z - (\operatorname{id} - \mathbf{P}_N)\mathbf{L}_r = (\mathbf{L} - z)(\operatorname{id} - (\mathbf{L} - z)^{-1}(\operatorname{id} - \mathbf{P}_N)\mathbf{L}_r). \tag{6.6}$$

We estimate

$$\begin{aligned} &\|(\mathbf{L} - z)^{-1}(\operatorname{id} - \mathbf{P}_N)\mathbf{L}_r)\|_{H^k \to H^k} \\ &\quad \leq \|(\mathbf{L} - z)^{-1}\|_{L^2 \to H^k} \|\operatorname{id} - \mathbf{P}_N\|_{H^{k-p} \to L^2} \|\mathbf{L}_r\|_{H^k \to H^{k-p}} \\ &\quad \leq \epsilon^{-1} \|\operatorname{id} - \mathbf{P}_N\|_{H^{k-p} \to L^2} \|\mathbf{L}_r\|_{H^k \to H^{k-p}} \leq C\epsilon^{-1} N^{p-k}, \end{aligned} \tag{6.7}$$

for some z-independent constant C. We see that if $N^{p-k} < C^{-1}\epsilon$ then $\sigma(\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r) \subset \sigma_\epsilon(\mathbf{L})$. This establishes the second requirement in the definition of spectrum convergence because: If λ was such an accumulation point but not in the spectrum, then it is outside the ϵ-pseudospectrum for some ϵ. But only a finite number of eigenvalues can be outside ϵ-pseudospectrum this choice of ϵ.

It remains to show that every eigenvalue λ of $\mathbf{L}$ is approximated. That is, we need to show that for any $\delta > 0$, there exists $N_0(\delta) > 0$ such that for $N > N_0$, $\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r$ has an eigenvalue λ_N satisfying $|\lambda_N - \lambda| \leq \delta$.

So let λ be an eigenvalue of $\mathbf{L}$ and let $\mathbf{u}$, $\|\mathbf{u}\|_{H^k} = 1$, be an associated eigenfunction. Then we know that

$$\mathbf{u} = \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma_\delta} (z - \mathbf{L})^{-1} \mathbf{u} \mathrm{d}z,$$

where Γ_δ is a circle centered at λ with radius δ. Then consider

$$\mathbf{u}_N := \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma_\delta} (z - \mathbf{L}_o - \mathbf{P}_N \mathbf{L}_r)^{-1} \mathbf{u} \mathrm{d}z.$$

There could be an eigenvalue of $\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r$ on Γ_δ, but then we are done. So we assume this is not the case. And if we show that $\mathbf{u}_N \neq \mathbf{0}$ then there must be an eigenvalue λ_N satisfying $|\lambda_N - \lambda| < \delta$.

We have

$$\|\mathbf{u} - \mathbf{u}_N\|_{H^k} \leq \delta \max_{z \in \Gamma_\delta} \|(z - \mathbf{L})^{-1} \mathbf{u} - (z - \mathbf{L}_o - \mathbf{P}_N \mathbf{L}_r)^{-1} \mathbf{u}\|_{H^k}.$$

We work the use the resolvent identity

$$\begin{aligned} (\mathrm{id} - \mathbf{P}_N)\mathbf{L}_r &= \mathbf{L} - z - (\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z), \\ (\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z)^{-1} (\mathrm{id} - \mathbf{P}_N)(\mathbf{L} - z)^{-1} &= (\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z)^{-1} - (\mathbf{L} - z)^{-1}. \end{aligned}$$

This gives

$$(z - \mathbf{L})^{-1} \mathbf{u} - (z - \mathbf{L}_o - \mathbf{P}_N \mathbf{L}_r)^{-1} \mathbf{u} = \frac{1}{\lambda - z} (\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z)^{-1} (\mathrm{id} - \mathbf{P}_N) \mathbf{L}_r \mathbf{u}.$$

It will suffice to get a uniform upper bound on the norm of $(\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z)^{-1}$. Again, consider (6.6) and the bound (6.7). It follows that

$$z \mapsto \|(\mathbf{L} - z)^{-1}\|_{L^2 \to H^k}$$

is a continuous function of z on the resolvent set for $\mathbf{L}$. Therefore, this is bounded by a constant C_δ on the compact set Γ_δ. Then using Theorem 4.8, for N sufficiently large, $N > N_0(\delta)$, we have that

$$\|(\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z)^{-1}\|_{L^2 \to H^k} \leq 2C_\delta.$$

To complete the argument, for $N > N_0(\delta)$

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_N\|_{H^k} &\leq \max_{z \in \Gamma_\delta} \|(\mathbf{L}_o + \mathbf{P}_N \mathbf{L}_r - z)^{-1} (\mathrm{id} - \mathbf{P}_N) \mathbf{L}_r \mathbf{u}\|_{H_k} \\ &\leq 2C_\delta \|\mathrm{id} - \mathbf{P}_N\|_{H^{k-p} \to L^2} \|\mathbf{L}_r\|_{H^k \to H^{k-p}} \leq C'_\delta N^{-k+p}. \end{aligned}$$

And as soon as the right-hand side is less than 1, we can conclude that $\mathbf{u}_N \neq \mathbf{0}$ and the proof is complete. ■

Many other questions could be asked:

- What about rates of convergence?

- What about convergence of eigenfunctions?

- What improves for self-adjoint operators?

For a discussion of these extensions, see [22].

6.4 ▪ Difficulties in spectrum approximation

We had great success in the previous section approximating the eigenvalues of rather large class of differential operators. We now give three examples to show that things are not always so straightforward and such successes are to be celebrated.

Example 6.15. Consider the operator $\mathcal{M}(\mathrm{e}^{\mathrm{i}\diamond})$ on $L^2(\mathbb{T})$. It follows that

$$\sigma(\mathcal{M}(\mathrm{e}^{\mathrm{i}\diamond})) = \{\mathrm{e}^{\mathrm{i}\theta} : \theta \in \mathbb{T}\}.$$

Furthermore, it is clear that this operator is unitary with inverse $\mathcal{M}(\mathrm{e}^{-\mathrm{i}\diamond})$. If we apply the finite-section method to this, we first find its bi-infinite matrix representation

$$\mathbf{M}(\mathrm{e}^{\mathrm{i}\diamond}) = \left[\begin{array}{ccc|ccc} \ddots & \ddots & \ddots & & & \iddots \\ \ddots & 0 & 0 & 0 & & \\ \ddots & 1 & 0 & 0 & 0 & \\ \hline & 0 & 1 & 0 & 0 & \ddots \\ & & 0 & 1 & 0 & \ddots \\ \iddots & & & \ddots & \ddots & \ddots \end{array}\right].$$

So, any finite-section truncation of this will result in a lower-triangular nilpotent matrix with zeros on the diagonal — $\sigma(\tilde{\mathbf{P}}_N \mathbf{M}(\mathrm{e}^{\mathrm{i}\diamond})\tilde{\mathbf{P}}_N^T) = \{0\}$. And this is a complete and utter failure of the finite-section method.

We also point out that, interestingly, the spectrum of $\mathcal{M}(\mathrm{e}^{\mathrm{i}\diamond})$ has no discrete components. So, it requires more work to understand in what sense we could approximate this set with a discrete set. And if one instead takes $\tilde{\mathbf{P}}_N \mathbf{M}(\mathrm{e}^{\mathrm{i}\diamond})\tilde{\mathbf{P}}_N^T$ and adds a 1 to the upper-right corner of the matrix, obtaining a circulant approximation, the eigenvalues are roots of unity and clearly do a much better job of approximating, in some sense, the true spectrum.

Example 6.16. In the previous case, we might think the difficulty comes from the fact that the spectrum is a continuous set — the resolvent is not compact. So, if we introduce an unbounded operator, maybe things will be resolved. So consider

$$\mathbf{M}(\mathrm{e}^{\mathrm{i}\diamond})\mathbf{D}.$$

But, again finite-sections of this matrix result in strictly lower-triangular matrices with $\sigma(\tilde{\mathbf{P}}_N \mathbf{M}(\mathrm{e}^{\mathrm{i}\diamond})\mathbf{D}\tilde{\mathbf{P}}_N^T) = \{0\}$.

Exercise 6.1. *What is the true spectrum of* $\mathbf{M}(\mathrm{e}^{\mathrm{i}\diamond})\mathbf{D}$*? Does the "circulant trick" from the previous example help resolve the issues?*

Example 6.17. Consider the measure,

$$\mu(\mathrm{d}x) = \frac{1}{2}(\mathbb{1}_{[-2,-1]}(x) + \mathbb{1}_{[1,2]}(x))\mathrm{d}x.$$

And we consider the associated Jacobi matrix

$$\mathcal{J}(\mu)$$

as an operator on $\ell^2(\mathbb{N})$. If $x \in (-2,-1) \cup (1,2)$ then it follows that the orthogonal polynomial sequence $(p_n(x))_{n\geq 0}$ is bounded in n. Thus, we have a bounded solution of

$$\mathcal{J}(\mu)\mathbf{v} = x\mathbf{v}.$$

This implies that $x \in \sigma(\mathcal{J}(\mu))$. And since the spectrum is closed $\operatorname{supp}(\mu) \subset \sigma(\mathcal{J}(\mu))$. More is true, it actually follows that $\sigma(\mathcal{J}(\mu)) = \operatorname{supp}(\mu)$, see [5], for example. So, let us see how our finite-section ideas do on this operator. Since the measure is even in the sense that

$$\int x^{2j-1}\mu(\mathrm{d}x) = 0, \quad j = 1, 2 \ldots,$$

it follows that $p_{2j-1}(0) = 0$. This implies that $\mathbf{J}_{2j-1}$ has an eigenvalue at the origin for all j. Thus, the finite-section approximation (1) fails to be invertible even when the true operator is invertible and (2) gives spurious eigenvalues. This latter phenomenon is often called *spectral polution.*

Chapter 7

Boundary-value problems

In this chapter, we present a slightly non-classical take on the classical Chebyshev pseudospectral method for solving boundary-value problems for ODEs. The classical approach first treats two-point problems for second-order differential equations using the so-called Chebyshev differentiation matrix which precisely the matrix representation of the operator

$$\mathbf{f} \mapsto \left[\frac{\mathrm{d}}{\mathrm{d}x} \mathcal{I}_N^{\mathrm{T}} \mathbf{f}(\check{x}_j) \right].$$

The entries in this matrix can be computed explicitly, and [20] is a particularly good reference for this. But, because we have built up some significant orthogonal polynomial theory, we take a different approach and perform the same task using the ultraspherical (Gegenbauer) polynomials. We also sacrifice convergence theory for simple implementations.

This takes us a significant way to describing the sparse ultraspherical method of Olver and Townsend [15]. So, we finish the chapter with a discussion of this method.

7.1 ▪ The classical Chebyshev approach though the ultraspherical lens

7.1.1 ▪ Normalization constants

Before we begin, we have to decide on a normalization for our polynomials. The classical ultraspherical polynomials, denoted by $C_j^{(\lambda)}(x)$ which are orthogonal with respect to $w_\lambda(x) = (1-x^2)^{\lambda - \frac{1}{2}}$, are not orthonormal [13]. For convenience, define

$$p_j(x;\lambda) = p_j(x;\lambda - 1/2, \lambda - 1/2), \quad \pi_j(x;\lambda) = \pi_j(x;\lambda - 1/2, \lambda - 1/2).$$

Here we recall that the p_j's are orthonormal with respect to the normalized weight function, normalized so that it is a probability measure.

So, consider some quantities

$$\int_{-1}^{1} w_\lambda(x)\mathrm{d}x = \sqrt{\pi}\frac{\Gamma(\lambda+\frac{1}{2})}{\Gamma(\lambda+1)} =: Z_\lambda,$$
$$\tilde{w}_\lambda(x) = Z_\lambda^{-1} w_\lambda(x),$$
$$k_j = k_j(\lambda) := \frac{2^j(\lambda)_j}{j!},$$
$$h_j = h_j(\lambda) := \frac{2^{1-2\lambda}\pi\Gamma(j+2\lambda)}{(j+\lambda)\Gamma(\lambda)^2 j!}.$$

Then define

$$c_j(\lambda) = \frac{\Gamma(\lambda+1)}{\sqrt{\pi}\Gamma(\lambda+\frac{1}{2})}\frac{h_j(\lambda)}{k_j(\lambda)^2},$$

so that

$$p_j(x;\lambda) = \frac{1}{\sqrt{c_j(\lambda)}}\pi_j(x;\lambda).$$

Recall that the monic ultraspherical polynomials satisfy

$$\pi_j'(x;\lambda) = j\pi_{j-1}(x;\lambda+1).$$

and therefore

$$p_j'(x;\lambda) = j\sqrt{\frac{c_{j-1}(\lambda+1)}{c_j(\lambda)}}p_{j-1}(x;\lambda+1).$$

This can be further simplified:

$$j\sqrt{\frac{c_{j-1}(\lambda+1)}{c_j(\lambda)}} = j\sqrt{\frac{2(\lambda+1)(j+2\lambda)}{2j\lambda+j}} =: d_j(\lambda).$$

7.1.2 ▪ Differentiation and evaluation

Suppose

$$u(x) = \sum_{j=0}^{N-1} c_j p_j(x;0) = c_0 + \sum_{j=0}^{N-1} c_j\sqrt{2}T_j(x).$$

Then

$$u'(x) = \sum_{j=1}^{N} c_j d_j(0) p_{j-1}(x;1) = \sum_{j=0}^{N-1} c_{j+1}d_{j+1}(0)p_j(x;1) = \sum_{j=0}^{N-1} c_{j+1}d_{j+1}(0)U_j(x).$$

But we can continue

$$u''(x) = \sum_{j=2}^{N} c_j d_j(0) d_{j-1}(1) p_{j-2}(x;2) = \sum_{j=0}^{N-2} c_{j+2}d_{j+2}(0)d_{j+1}(1)p_j(x;2).$$

The leads us to define

$$\mathbf{D}_{\lambda\to\lambda+1} = \begin{bmatrix} 0 & d_1(\lambda) & & & \\ & 0 & d_2(\lambda) & & \\ & & 0 & d_3(\lambda) & \\ & & & 0 & \ddots \\ & & & & \ddots \end{bmatrix},$$

and

$$\mathbf{D}_k = \mathbf{D}_{k-1\to k}\cdots\mathbf{D}_{1\to 2}\mathbf{D}_{0\to 1}, \quad \mathbf{D}_0 = \mathrm{id}\,.$$

Thus, if $\mathbf{c} = (c_j)_{j\geq 0}$ are such that (formally)

$$u(x) = \sum_j c_j p_j(x;0),$$

then for $\mathbf{d} = \mathbf{D}_k\mathbf{c} = (d_j)_{j\geq 0}$

$$u^{(k)}(x) = \sum_j d_j p_j(x;k). \tag{7.1}$$

This gives a sparse way to compute the new coefficients. We also know how to use the three-term recurrence to evaluate the series. When the coefficients are known, Clenshaw is the best way to evaluate the series, but if the coefficients are unknown — they are the solution of a linear system — we use forward recurrence to compute the coefficients.

Specifically, let $P = (x_1,\ldots,x_m)$ be a grid on $[-1,1]$. Then define the evaluation matrix

$$\mathbf{P}_{\lambda\to P} = \begin{bmatrix} p_0(x_1;\lambda) & p_1(x_1;\lambda) & p_2(x_1;\lambda) & \cdots \\ p_0(x_2;\lambda) & p_1(x_2;\lambda) & p_2(x_2;\lambda) & \cdots \\ \vdots & \vdots & \vdots & \\ p_0(x_m;\lambda) & p_1(x_m;\lambda) & p_2(x_m;\lambda) & \cdots \end{bmatrix}$$

And then we have the relation

$$\begin{bmatrix} \\ u^{(k)}(x_j) \\ \\ \end{bmatrix} = \mathbf{P}_{k\to P}\mathbf{D}_k\mathbf{c}.$$

And to construct $\mathbf{P}_{\lambda\to P}$ we use that the columns $\mathbf{p}_j$ satisfy the three-term recurrence:

$$\mathbf{p}_{j+1} = \frac{1}{b_j}\left[\mathbf{x}\cdot\mathbf{p}_j - a_j\mathbf{p}_j - b_{j-1}\mathbf{p}_j\right], \quad \mathbf{p}_{-1} = \mathbf{0}, \quad \mathbf{p}_0 = \mathbf{1}.$$

where $\mathbf{x} = (x_j)_{j=1}^m$ and $\cdot$ denotes the entrywise product.

7.1.3 ▪ Numerical solution of boundary-value problems

We now are in position to discuss how to solve

$$\sum_{j=0}^k a_j(x)\frac{\mathrm{d}^j u}{\mathrm{d}x^j}(x) = f(x), \quad x\in(-1,1),$$

$$\mathbf{S}\begin{bmatrix} u(-1) \\ u'(-1) \\ \vdots \\ u^{(k-1)}(-1) \end{bmatrix} + \mathbf{T}\begin{bmatrix} u(1) \\ u'(1) \\ \vdots \\ u^{(k-1)}(1) \end{bmatrix} = \mathbf{b}, \quad \mathbf{S},\mathbf{T}\in\mathbb{C}^{k\times k}, \mathbf{b}\in\mathbb{C}^k.$$

Here N will denote the degree of the approximation, i.e.,

$$u(x) \approx \sum_{j=0}^{N-1} c_j p_j(x; 0),$$

and we set up a linear system to compute the coefficients c_j. Since we have, in effect, k constraints from the boundary conditions, we need $N-k$ equations. To construct these equations, we use collocation and enforce that the differential equation should hold at a set of nodes $\mathbf{x}_N = (x_j)_{j=1}^{N-k}$. A good choice for these nodes is the roots of Chebyshev polynomials T_{N-k} or U_{N-k}.

We then discretize

$$a_j(x)\frac{\mathrm{d}^j u}{\mathrm{d}x^j}(x) \to \operatorname{diag}(a_j(\mathbf{x}_N))\mathbf{P}_{j\to\mathbf{x}_N}\mathbf{D}_j\mathbf{c},$$

where

$$a_j(\mathbf{x}_N) = (a_j(x_1), \ldots, a_j(x_{N-k}))^T,$$

and set

$$\mathbf{A} = \sum_{j=0}^{k} \operatorname{diag}(a_j(\mathbf{x}_N))\mathbf{P}_{j\to\mathbf{x}_N}\mathbf{D}_j.$$

To encode the boundary conditions, we use the discretization

$$u^{(j)}(a) \to \mathbf{P}_{j\to\{a\}}\mathbf{D}_j\mathbf{c},$$

so that all the boundary conditions are discretized as

$$\underbrace{\left[\mathbf{S}\begin{bmatrix}\mathbf{P}_{0\to\{-1\}}\mathbf{D}_0\\ \mathbf{P}_{1\to\{-1\}}\mathbf{D}_1\\ \vdots\\ \mathbf{P}_{k-1\to\{-1\}}\mathbf{D}_{k-1}\end{bmatrix} + \mathbf{T}\begin{bmatrix}\mathbf{P}_{0\to\{1\}}\mathbf{D}_0\\ \mathbf{P}_{1\to\{1\}}\mathbf{D}_1\\ \vdots\\ \mathbf{P}_{k-1\to\{1\}}\mathbf{D}_{k-1}\end{bmatrix}\right]}_{\mathbf{B}}\mathbf{c} = \mathbf{b}.$$

The full discretization of the boundary-value problem is then given by

$$\begin{bmatrix}\mathbf{B}\\ \mathbf{A}\end{bmatrix}\mathbf{c} = \begin{bmatrix}\mathbf{b}\\ f(\mathbf{x}_N)\end{bmatrix}.$$

7.1.4 ▪ Eigenvalue problems

We discuss how to approximate eigenpairs (u, λ) such that

$$\sum_{j=0}^{k} a_j(x)\frac{\mathrm{d}^j u}{\mathrm{d}x^j}(x) = \lambda u(x), \quad x \in (-1, 1),$$

$$\mathbf{S}\begin{bmatrix}u(-1)\\ u'(-1)\\ \vdots\\ u^{(k-1)}(-1)\end{bmatrix} + \mathbf{T}\begin{bmatrix}u(1)\\ u'(1)\\ \vdots\\ u^{(k-1)}(1)\end{bmatrix} = \mathbf{0}, \quad \mathbf{S}, \mathbf{T} \in \mathbb{C}^{k\times k}.$$

One immediate option is to rewrite it as

$$\begin{bmatrix}\mathbf{B}\\ \mathbf{A}\end{bmatrix}\mathbf{c} = \lambda\begin{bmatrix}\mathbf{0}\\ \mathbf{P}_{0\to\mathbf{x}_N}\end{bmatrix}\mathbf{c}.$$

This is now a generalized eigenvalue problem. Numerical methods do exist in standard packages for such eigenvalue problems.

Another option is to do a “basis recombination” step to only work with the nullspace of $\mathbf{B}$. As $\mathbf{B} \in \mathbb{C}^{k\times N}$, we suppose it has a nullspace of dimension $N-k$. And we want to construct an orthonormal basis for this nullspace — we want to find orthonormal vectors that are orthogonal to the rows of the complex conjugate of $\mathbf{B}$. So, compute a full QR decomposition:

$$\mathbf{B}^* = \mathbf{QR}, \quad \mathbf{Q} \in \mathbb{C}^{N\times N}, \quad \mathbf{R} \in \mathbb{C}^{N\times k}.$$

It follows that the columns of $\mathbf{U} := \mathbf{Q}_{1:N,k+1:N} = [\mathbf{q}_1, \ldots \mathbf{q}_{N-k}]$, $\mathbf{U} \in \mathbb{C}^{N\times N-k}$, form an orthonormal basis for $N(\mathbf{B})$. Computing this basis requires the use of k Householder reflectors and can be constructed in $O(N^2)$ FLOPs (taking k to be a constant). So, set $\mathbf{c} = \mathbf{Ud}$ and the eigenvalue problem becomes

$$\mathbf{AUd} = \lambda\mathbf{P}_{0\to\mathbf{x}_N}\mathbf{Ud}.$$

This is still a generalized eigenvalue problem. But now $\mathbf{C} := \mathbf{P}_{0\to\mathbf{x}_N}\mathbf{U} \in \mathbb{C}^{N-k\times N-k}$ and we can (likely) invert it to obtain

$$[\mathbf{C}^{-1}\mathbf{AU}]\mathbf{d} = \lambda\mathbf{d}.$$

Because of a lack of sparsity, there is an unavoidable complexity of $O(N^3)$ FLOPs.

7.1.5 ▪ Linear, time-dependent problems

Next, we discuss the numerical discretization of

$$\frac{\partial u}{\partial t}(x,t) - \sum_{j=0}^{k} a_j(x)\frac{\partial^j u}{\partial x^j}(x,t) = f(x,t), \quad x \in (-1,1), \quad t > 0,$$

$$u(x,0) = u_0(x),$$

$$\mathbf{S}\begin{bmatrix} u(-1,t)\\ u'(-1,t)\\ \vdots\\ u^{(k-1)}(-1,t)\end{bmatrix} + \mathbf{T}\begin{bmatrix} u(1,t)\\ u'(1,t)\\ \vdots\\ u^{(k-1)}(1,t)\end{bmatrix} = \mathbf{b}(t), \quad \mathbf{S},\mathbf{T} \in \mathbb{C}^{k\times k},\ \mathbf{b}(t) \in \mathbb{C}^k.$$

Here we do not discuss, at length, numerical methods for the time integration.

If $f = 0, \mathbf{b} = 0$ then the problem can be reduced as above to

$$\frac{\mathrm{d}\mathbf{d}}{\mathrm{d}t}(t) = [\mathbf{C}^{-1}\mathbf{AU}]\mathbf{d}(t), \quad \mathbf{d}(t) = \mathrm{e}^{-\mathbf{C}^{-1}\mathbf{AU}t}\mathbf{d}(0),$$

where $\mathbf{d}(0)$ is determined by the initial condition $u_0(x)$.

If $f \neq 0$ but $\mathbf{b} = 0$, then some type of integration is required. To see this, write

$$\mathbf{P}_{0\to\mathbf{x}_N}\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t}(t) - \mathbf{Ac}(t) = f(\mathbf{x}_N, t).$$

We impose, again, the boundary conditions by setting $\mathbf{c}(t) = \mathbf{U}\mathbf{d}(t)$ and writing

$$\mathbf{P}_{0\to\mathbf{x}_N}\mathbf{U}\frac{\mathrm{d}\mathbf{d}}{\mathrm{d}t}(t) - \mathbf{A}\mathbf{U}\mathbf{d}(t) = f(\mathbf{x}_N, t).$$

This is really a differential-algebraic equation (DAE). By inverting the matrix coefficient, if desired, it can be turned into a standard ordinary differential equation.

The story is a bit more complicated if $\mathbf{b} \neq 0$ because there is no way in which $\mathbf{c}(t) = \mathbf{U}\mathbf{d}(t)$ can satisfy non-zero boundary data. And a possible approach is to return to the $\mathbf{Q}$ matrix above and partition it as $\mathbf{Q} = \begin{bmatrix}\mathbf{V} & \mathbf{U}\end{bmatrix}$. We set

$$\mathbf{c}(t) = \mathbf{U}\mathbf{d}(t) + \mathbf{V}\mathbf{e}(t) = \mathbf{Q}\begin{bmatrix}\mathbf{e}(t)\\ \mathbf{d}(t)\end{bmatrix}.$$

The boundary conditions give

$$\mathbf{B}\mathbf{c}(t) = \mathbf{B}\mathbf{Q}\begin{bmatrix}\mathbf{e}(t)\\ \mathbf{d}(t)\end{bmatrix} = \mathbf{R}^*\begin{bmatrix}\mathbf{e}(t)\\ \mathbf{d}(t)\end{bmatrix},$$

since $\mathbf{B} = \mathbf{R}^*\mathbf{Q}^*$. Note that the last $N-k$ columns of $\mathbf{R}^*$ are zero. Thus we can solve directly for $\mathbf{e}(t) = \tilde{\mathbf{R}}^{-1}\mathbf{b}(t)$, where $\mathbf{R}^* = \begin{bmatrix}\tilde{\mathbf{R}} & \mathbf{0}\end{bmatrix}$. Again, we have

$$\begin{aligned}\mathbf{P}_{0\to\mathbf{x}_N}\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t}(t) - \mathbf{A}\mathbf{c}(t) &= f(\mathbf{x}_N, t),\\ \mathbf{P}_{0\to\mathbf{x}_N}\mathbf{U}\frac{\mathrm{d}\mathbf{d}}{\mathrm{d}t}(t) + \mathbf{P}_{0\to\mathbf{x}_N}\mathbf{V}\frac{\mathrm{d}\mathbf{e}}{\mathrm{d}t}(t) - \mathbf{A}\mathbf{U}\mathbf{d}(t) - \mathbf{A}\mathbf{V}\mathbf{e}(t) &= f(\mathbf{x}_N, t).\end{aligned}$$

Upon rearranging,

$$\mathbf{P}_{0\to\mathbf{x}_N}\mathbf{U}\frac{\mathrm{d}\mathbf{d}}{\mathrm{d}t}(t) - \mathbf{A}\mathbf{U}\mathbf{d}(t) = f(\mathbf{x}_N, t) + \mathbf{A}\mathbf{V}\mathbf{e}(t) - \mathbf{P}_{0\to\mathbf{x}_N}\mathbf{V}\frac{\mathrm{d}\mathbf{e}}{\mathrm{d}t}(t).$$

The general approach

If $\mathbf{S}, \mathbf{T}$ depend on t, or if using the QR decomposition above becomes too taxing, there is a general approach using DAEs. For this, we just attack the equations for $\mathbf{c}$ directly:

$$\begin{aligned}\mathbf{P}_{0\to\mathbf{x}_N}\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t}(t) &= \mathbf{A}\mathbf{c}(t) + f(\mathbf{x}_N, t),\\ \mathbf{0} &= \mathbf{B}\mathbf{c} - \mathbf{b}(t),\\ \begin{bmatrix}\mathbf{0}\\ \mathbf{P}_{0\to\mathbf{x}_N}\end{bmatrix}\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t}(t) &= \begin{bmatrix}\mathbf{B}\mathbf{c} - \mathbf{b}(t)\\ \mathbf{A}\mathbf{c}(t) + f(\mathbf{x}_N, t).\end{bmatrix}\end{aligned}$$

7.2 ▪ The sparse ultraspherical method

In the previous chapter, a collocation approach was used. The evaluation operator $\mathbf{P}_{\lambda\to\mathbf{x}}$ was used to enforce that the differential equation should hold at a set of points. This is convenient and led to the coefficients of the differential equation entering as diagonal matrices. The issue with this approach is that it, in general, will lead to dense matrices. This chapter is devoted to the so-called sparse ultraspherical spectral method, a Petrov–Galerkin method[5], results in sparse matrices. From a linear-algebraic perspective, the only complication is that these sparse systems will result in fill-in if Gaussian elimination is used naively. So, one employs either the adaptive QR algorithm or iterative methods with preconditioning.

[5] A Petrov–Galerkin method uses different orthogonal basis functions, and likely a different inner product, for the domain and the range.

7.2.1 ▪ Connection coefficients

To derive the important aspects of the ultraspherical method we consider the discretization of following differential operator,

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}}{\mathrm{d}x} + q(x).$$

We have seen that the derivative operator $\mathbf{D}_{\lambda\to\lambda+1}$ is sparse. So, we may want to consider

$$\mathbf{D}_{1\to 2}\mathbf{D}_{0\to 1} + \mathbf{D}_{0\to 1},$$

but unfortunately, this is meaningless because the range of the two operators are the coefficients in the expansion in different bases — $p_j(x;2)$ versus $p_j(x;1)$. So, to fix this issue we need to convert an expansion in $p_j(x;1)$ to one in $p_j(x;2)$, and we, of course, use connection coefficients for this purpose. So, we want to write

$$p_k(x;\lambda) = \sum_{j=0}^{k} c_{k,j} p_j(x;\lambda+1).$$

So, set

$$c_{k,j} = \int_{-1}^{1} p_j(x;\lambda) p_k(x;\lambda+1)\tilde{w}_{\lambda+1}(x)\mathrm{d}x$$

It follows that this vanishes for $j < k$, by orthogonality of $p_k(x;\lambda+1)$. Furthermore, for $k > j+2$, the orthogonality of $p_k(x;\lambda)$ and $(1-x^2)p_j(x;\lambda+1)$ implies this vanishes. So, it remains to compute, for $k > 0$:

$$\begin{aligned}\int_{-1}^{1} p_k(x;\lambda+1)p_k(x;\lambda)\tilde{w}_{\lambda+1}(x)\mathrm{d}x &= \int_{-1}^{1} p_k(x;\lambda+1)\frac{x^k}{\sqrt{c_k(\lambda)}}\tilde{w}_{\lambda+1}(x)\mathrm{d}x\\ &= \int_{-1}^{1} p_k(x;\lambda+1)p_k(x;\lambda+1)\frac{\sqrt{c_k(\lambda+1)}}{\sqrt{c_k(\lambda)}}\tilde{w}_{\lambda+1}(x)\mathrm{d}x\\ &= \frac{\sqrt{c_k(\lambda+1)}}{\sqrt{c_k(\lambda)}},\end{aligned}$$

$$\int_{-1}^{1} p_{k-1}(x;\lambda+1)p_k(x;\lambda)\tilde{w}_{\lambda+1}(x)\mathrm{d}x = 0,$$

$$\begin{aligned}\int_{-1}^{1} p_{k-2}(x;\lambda+1)p_k(x;\lambda)\tilde{w}_{\lambda+1}(x)\mathrm{d}x &= \int_{-1}^{1} p_{k-2}(x;\lambda+1)p_k(x;\lambda)(1-x^2)\frac{Z_\lambda}{Z_{\lambda+1}}\tilde{w}_{\lambda}(x)\mathrm{d}x\\ &= -\frac{Z_\lambda}{Z_{\lambda+1}}\int_{-1}^{1} p_k(x;\lambda)\frac{x^k}{\sqrt{c_{k-2}(\lambda+1)}}\tilde{w}_{\lambda}(x)\mathrm{d}x\\ &= -\frac{Z_\lambda}{Z_{\lambda+1}}\frac{\sqrt{c_k(\lambda)}}{\sqrt{c_{k-2}(\lambda+1)}}.\end{aligned}$$

We then obtain the simplified relations

$$\begin{aligned}\frac{\sqrt{c_k(\lambda+1)}}{\sqrt{c_k(\lambda)}} &= \sqrt{\frac{(\lambda+1)(k+2\lambda)(k+2\lambda+1)}{2(2\lambda+1)(k+\lambda)(k+\lambda+1)}},\\ \frac{Z_\lambda}{Z_{\lambda+1}}\frac{\sqrt{c_k(\lambda)}}{\sqrt{c_{k-2}(\lambda+1)}} &= \sqrt{\frac{(k-1)k(\lambda+1)}{2(2\lambda+1)(k+\lambda-1)(k+\lambda)}}.\end{aligned}$$

So, define

$$s_k(\lambda) := \begin{cases} 1 & k = 0, \\ \sqrt{\frac{(\lambda+1)(k+2\lambda)(k+2\lambda+1)}{2(2\lambda+1)(k+\lambda)(k+\lambda+1)}} & \text{otherwise,} \end{cases}$$
$$t_k(\lambda) := \sqrt{\frac{(k-1)k(\lambda+1)}{2(2\lambda+1)(k+\lambda-1)(k+\lambda)}}$$

And we have

$$\begin{aligned} p_0(x;\lambda) &= s_0(\lambda)p_0(x;\lambda+1), \\ p_1(x;\lambda) &= s_1(\lambda)p_1(x;\lambda+1), \\ p_k(x;\lambda) &= s_k(\lambda)p_k(x;\lambda+1) - t_k(\lambda)p_{k-2}(x;\lambda+1), \quad k \geq 2. \end{aligned}$$

And we then define

$$\mathbf{C}_{\lambda\to\lambda+1} = \begin{bmatrix} s_0(\lambda) & 0 & -t_2(\lambda) & & & \\ & s_1(\lambda) & 0 & -t_3(\lambda) & & \\ & & s_2(\lambda) & 0 & -t_4(\lambda) & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}$$

Therefore if $\mathbf{d} = \mathbf{C}_{\lambda\to\lambda+1}\mathbf{c}$ then

$$\sum_j d_j p_j(x;\lambda+1) = \sum_j c_j p_j(x,\lambda).$$

Amazingly, this is sparse! The conversion going the other direction is not. And, for convenience, define

$$\mathbf{C}_{j\to k} = \prod_{l=0}^{k-j-1} \mathbf{C}_{j+\ell\to j+\ell+1}$$

Then the discretization of

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}}{\mathrm{d}x},$$

is given by the sparse matrix

$$\mathbf{D}_{1\to 2}\mathbf{D}_{0\to 1} + \mathbf{C}_{1\to 2}\mathbf{D}_{0\to 1} = \mathbf{D}_1\mathbf{C}_{1\to 2}\mathbf{D}_1$$

We then need to handle multiplication by functions in this basis. That is done in the following section.

7.2.2 ▪ Function multiplication

To handle multiplication, in general, we suppose that our input coefficients have rapidly converging orthogonal polynomial expansions. Specifically, assume

$$q(x) = \sum_{j=0}^{N-1} \alpha_j p_j(x;0).$$

An expansion in a different orthogonal polynomial basis can be assumed, and the derivation below generalizes straightforwardly by replacing the recurrence coefficients in (7.2) appropriately.

In general, q may not be a polynomial, but we can replace it with one at the cost of a small error will need to be understood. let $\mathbf{J}_\lambda$ be the Jacobi operator associated to $\tilde{w}_\lambda$. Then

$$u(x) = \sum_j u_j p_j(x;\lambda), \quad \mathbf{v} = \mathbf{J}_\lambda \mathbf{u} \Rightarrow xu(x) = \sum_j v_j p_j(x;\lambda),$$

and therefore

$$q(x)u(x) = \sum_j w_j p_j(x;\lambda), \quad \mathbf{w} = q(\mathbf{J}_\lambda)\mathbf{u}.$$

We need to develop (stable) methods to evaluate $q(\mathbf{J}_\lambda)\mathbf{u}$ or $q(\mathbf{J}_\lambda)$. To evaluate the latter, we will be able to replace $\mathbf{u}$ with an identity matrix. The following gives the recurrence

$$\begin{aligned}
\mathbf{p}_0 &= \mathbf{u}, \\
\mathbf{p}_1 &= \sqrt{2}\mathbf{J}_\lambda \mathbf{p}_0, \\
\mathbf{p}_2 &= 2\mathbf{J}_\lambda \mathbf{p}_1 - \sqrt{2}\mathbf{p}_0, \\
\mathbf{p}_j &= 2\mathbf{J}_\lambda \mathbf{p}_{j-1} - \mathbf{p}_{j-2}, \quad j \geq 3,
\end{aligned} \tag{7.2}$$

which is run simultaneously with the iterates

$$\begin{aligned}
\mathbf{q}_{-1} &= \mathbf{0}, \\
\mathbf{q}_j &= \mathbf{q}_{j-1} + \alpha_j \mathbf{p}_j, \quad 0 \leq j \leq \mathrm{degree}(q).
\end{aligned}$$

We denote by $\mathbf{M}_\lambda(q)$ the resulting operator. Thus, our discretization

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}}{\mathrm{d}x} + q(x),$$

can now be completed:

$$\mathbf{D}_{1\to 2}\mathbf{D}_{0\to 1} + \mathbf{C}_{1\to 2}\mathbf{D}_{0\to 1} + \mathbf{M}_2(q)\mathbf{C}_{1\to 2}\mathbf{C}_{0\to 1}.$$

And the discretization of the general operator

$$\mathcal{L} = \sum_{j=0}^{k} a_j(x)\frac{\mathrm{d}^k}{\mathrm{d}x^k} \to \sum_{j=0}^{k} \mathbf{M}_k(a_j)\mathbf{C}_{j\to k}\mathbf{D}_j =: \mathbf{L}$$

7.2.3 ▪ Adding boundary conditions

To account for boundary conditions, at say ±1, the process is essentially the same as above. We find the evaluation matrices $\mathbf{P}_{k\to\{\pm 1\}}\mathbf{D}_k$, which, in principle, have an infinite number of columns.

As an example, we consider solving

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x) - xu(x) = 0, \quad u(\pm 1) = \mathrm{Ai}(\pm 1).$$

Then the full, infinite-dimensional version of the boundary-value problem becomes

$$\begin{bmatrix} \mathbf{P}_{0\to\{1\}} \\ \mathbf{P}_{0\to\{-1\}} \\ \mathbf{D}_2 - \mathbf{M}_2(\diamond)\mathbf{C}_{0\to 2} \end{bmatrix} \mathbf{c} = \begin{bmatrix} \mathrm{Ai}(1) \\ \mathrm{Ai}(-1) \\ \mathbf{0} \end{bmatrix}.$$

And, in general, a kth-order boundary-value problem will have k dense rows at the top, corresponding to enforcing boundary conditions, along with a sparse block below it. The sparsity of this block depends on the degree of the polynomial coefficients. Each increase in a degree by one, increases the bandwidth of this block by one.

7.2.4 ▪ Solving the linear system

There are many approaches one may use to solve the resulting linear system. The naive approach is to use finite sections and Gaussian elimination. This is fine, and will work. But one can do better. Due to the nature of the first few rows being dense, there is a huge risk of fill in giving an $O(N^2)$ solver for an $N \times N$ finite section. But hey, at least it is not $O(N^3)$.

There actually exists a direct and adaptive method that requires $O(b^2N)$ FLOPS where b is the bandwidth of the lower block. This is the *adaptive QR algorithm*. See [15] for a description of this. It requires a clever use of the dense blocks and Givens rotations to track when the residual will drop below a specified tolerance before a backward substitution solve step is initiated.

Yet another approach is to use (preconditioned) iterative methods. Since the matrix is sparse, this is a good idea as long as a decent preconditioner can be guessed. A good choice, in general, for a preconditioner is to take the “purely banded” part of the matrix because solving a linear system involving this matrix will require $O(b^2N)$ operations using Gaussian elimination. And then something like GMRES can be applied. While this approach might not seem adaptive, it can all be implemented in a matrix-free setting and adaptivity can be incorporated at this level. In the simplest setting, if one is not satisfied with the result of one round of GMRES, the approximate solution can be augmented with zeros and used as an initial guess for a larger system. Even more simply, a diagonal scaling, following [15], can be used, and this is convenient to establish the convergence of the method.

7.3 ▪ Convergence

7.4 ▪ Examples

Bibliography

[1] M J Ablowitz and A S Fokas. *Complex Variables: Introduction and Applications.* Cambridge University Press, second edition, 2003.

[2] N I Achieser. *Theory of Approximation.* Dover Publications, 1992.

[3] K Atkinson and W Han. *Theoretical Numerical Analysis.* Springer, New York, NY, 2009.

[4] W Cheney and W Light. *A Course in Approximation Theory*, volume 101 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, Rhode Island, 1 2009.

[5] P Deift. *Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert Approach.* Amer. Math. Soc., Providence, RI, 2000.

[6] G B Folland. *Real analysis.* John Wiley and Sons Inc., New York, 1999.

[7] W Gautschi. *Orthogonal Polynomials: Applications and Computation.* Oxford University Press, 2004.

[8] G H Hardy. Note on Lebesgue's Constants in the Theory of Fourier Series. *Journal of the London Mathematical Society*, s1-17(1):4–13, 1 1942.

[9] G H Hardy, J E Littlewood, and G Pólya. Inequalities, 1934.

[10] N J Higham. The numerical stability of barycentric Lagrange interpolation. *IMA Journal of Numerical Analysis*, 24(4):547–556, 10 2004.

[11] T Kato. *Perturbation theory for linear operators.* Springer, New York, NY, 1995.

[12] R Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences.* Springer New York, New York, NY, 2014.

[13] F W J Olver, D W Lozier, R F Boisvert, and C W Clark. *NIST Handbook of Mathematical Functions.* Cambridge University Press, 2010.

[14] S Olver, R M Slevinsky, and A Townsend. Fast algorithms using orthogonal polynomials. *Acta Numerica*, 29:573–699, 5 2020.

[15] S Olver and A Townsend. A Fast and Well-Conditioned Spectral Method. *SIAM Review*, 55(3):462–489, 1 2013.

[16] M Reed and B Simon. *IV: Analysis of Operators.* Academic Press, 1978.

[17] T J Rivlin. *An Introduction to the Approximation of Functions.* Dover Publications, 1969.

[18] W Rudin. *Principles of mathematical analysis.* McGraw-Hill, Inc., New York, NY, 3rd edition, 1964.

[19] G Szegő. *Orthogonal Polynomials*, volume 23 of *Colloquium Publications*. American Mathematical Society, Providence, Rhode Island, 12 1939.

[20] L N Trefethen. *Spectral methods in MATLAB*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[21] L N Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1 2019.

[22] T Trogdon. On the convergence of spectral methods involving non-compact operators. *arXiv preprint 2304.14319*, 2023.

[23] G N Watson. THE CONSTANTS OF LANDAU AND LEBESGUE. *The Quarterly Journal of Mathematics*, os-1(1):310–318, 1930.

[24] K Yoshida. *Functional analysis*. Springer Berlin / Heidelberg, 6th editio edition, 1980.