Dr H. Fawzi

Mathematical Tripos Part II: Michaelmas Term 2023

Numerical Analysis – Lecture 11

3 Spectral Methods

Finite difference schemes rest upon the replacement of derivatives by a linear combination of function values. This leads to the solution of a system of linear equations, which on the one hand tends to be large (due to the slow convergence properties of the approximation) but on the other hand is highly structured and sparse, leading itself to effective algorithms for its solution. In this chapter we look at spectral methods, a different way to discretize PDEs.

The basic idea of spectral methods The basic idea of spectral methods is simple. Consider a PDE of the form

$$\mathcal{L}u = f \tag{3.1}$$

where $\mathcal{L}$ is a differential operator (e.g., $\mathcal{L} = \frac{\partial^2}{\partial x^2}$, or $\mathcal{L} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, etc.) and f is a right-hand side function. We consider a finite-dimensional subspace of functions V spanned by a basis $\psi_1, \dots, \psi_N$. A typical choice for V is a space of (trigonometric) polynomials of finite degree. We seek an approximate solution to the PDE by a linear combination of the ψ_n, i.e., $u_N(x) = \sum_{n=1}^N c_n \psi_n(x)$. Plugging $u_N(x)$ in the PDE we get the following linear equation in the unknowns (c_n):

$$\sum_{n=1}^N c_n \mathcal{L}\psi_n = f. \tag{3.2}$$

In general the equation will not have a solution, as there is no reason to expect that the original PDE has a solution in the subspace V. However, we can seek to satisfy equation (3.2) approximately. Assume that the $(\psi_n)_{1\le n\le N}$ are an orthonormal family of functions, with respect to some inner product $\langle \cdot, \cdot \rangle$. Then instead of looking for (c_n) that satisfy (3.2), we will require only that the projection of $\mathcal{L}u_N - f$ on the subspace V is zero. This is the same as requiring that

$$\sum_{n=1}^N c_n \langle \mathcal{L}\psi_n, \psi_m \rangle = \langle f, \psi_m \rangle \qquad \forall m = 1, \dots, N. \tag{3.3}$$

If we call A the matrix $A_{m,n} = \langle \mathcal{L}\psi_n, \psi_m \rangle$, we end up with a $N \times N$ linear system $Ac = \tilde{f}$, where $\tilde{f}_m = \langle f, \psi_m \rangle$.

Remark 3.1 *The equations* (3.3) *are known as the* Galerkin equations*. Another approach to converting* (3.2) *into a finite set of equations, is to require that equality holds exactly at some specific points $(x_i)_{1\le i\le N}$. These lead to so-called* collocation methods*.*

In this chapter we will focus on two of the most common choices of basis functions (ψ_n); namely the Fourier basis, and the basis of Chebyshev polynomials.

3.1 Fourier approximation of functions

We focus on one-dimensional problems on the domain $[-1, 1]$. The basis of functions we consider here is

$$\psi_n(x) = e^{i\pi n x}, \qquad n \in \mathbb{Z}.$$

These functions are orthonormal with respect to the normalized L^2 inner product on $[-1,1]$, i.e.,

$$\langle \psi_n, \psi_m \rangle = \frac{1}{2}\int_{-1}^{1} \psi_n(x)\overline{\psi_m(x)} = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{else.} \end{cases}$$

Given a function $f : [-1,1] \to \mathbb{R}$, its *truncated Fourier approximation* is

$$f(x) \approx \phi_N(x) = \sum_{n=-N/2}^{N/2} \widehat{f}_n e^{i\pi n x}, \quad x \in [-1,1], \tag{3.4}$$

where here and elsewhere in this section $N \geq 2$ is an even integer and

$$\widehat{f}_n = \langle f, \psi_n \rangle = \frac{1}{2}\int_{-1}^{1} f(t)e^{-i\pi n t}\, dt, \quad n \in \mathbb{Z}$$

are the (Fourier) coefficients of this approximation. We want to analyse the approximation properties of (3.4). Observe that the basis functions (ψ_n) are all 2-periodic, and so if we hope to have convergence we should require f to be 2-periodic. We now recall a basic result from Fourier analysis giving simple sufficient conditions for the convergence of ϕ_N to f.

Theorem 3.2 *Assume $f : \mathbb{R} \to \mathbb{R}$ is 2-periodic and Lipschitz continuous. Then $\phi_N(x) \to f(x)$ for all $x \in \mathbb{R}$.*

If f is assumed smooth enough, then one can show that the Fourier series converges exponentially fast. This is the object of the next theorem.

Theorem 3.3 *Assume $f : \mathbb{R} \to \mathbb{R}$ is 2-periodic, and has an analytic continuation into the complex strip $\{z \in \mathbb{C} : -a < \operatorname{Im} z < a\}$ on which we further assume $|f(z)| \leq M$. Then, the following holds, with $c = e^{-a\pi} \in (0,1)$:*

- *$|\hat{f}_n| \leq Mc^{|n|}$ for all $n \in \mathbb{Z}$; and*
- *$\max_{x\in[-1,1]} |f(x) - \phi_N(x)| \leq \frac{2Mc}{1-c}c^{N/2}$.*

Proof. We start by proving the first bullet point. We know that $\hat{f}_n = \frac{1}{2}\int_{-1}^{1} f(x)e^{-\mathrm{i}\pi n x}dx$. The key part of the proof, is to show that $\hat{f}_n$ has the following alternative representation, as an integral in the complex plane:

$$\hat{f}_n = \tfrac{1}{2}\int_{-1}^{1} f(x + \mathrm{i}a')e^{-\mathrm{i}\pi n(x+\mathrm{i}a')}dx \tag{3.5}$$

for any $0 < a' < a$. To prove (3.5), note that since $F(z) = f(z)e^{-\mathrm{i}\pi n z}$ is analytic on the rectangle $[-1,1] \times [-a', a'] \subset \mathbb{C}$, by Cauchy's theorem we have $\int_\gamma F = 0$ where γ is the contour around this rectangle. Furthermore, since F is 2-periodic, we have $\int_{[1,1+\mathrm{i}a']} F = -\int_{[-1+\mathrm{i}a',-1]} F$. It thus follows that $\int_{[-1,1]} F = \int_{[-1+\mathrm{i}a',1+\mathrm{i}a']} F$, which proves (3.5). This immediately gives $|\hat{f}_n| \leq Me^{\pi n a'}$, which proves the desired inequality for $n \leq 0$, by letting $a' \to a$. To prove the inequality for $n \geq 0$ we use $x - \mathrm{i}a'$ instead of $x + \mathrm{i}a'$ in (3.5).

The second statement in the theorem is an immediate corollary of Theorem 3.2 and the first point. Indeed, for any $x \in [-1,1]$ we can write

$$|f(x) - \phi_N(x)| = |\sum_{|n|>N/2} \hat{f}_n e^{\mathrm{i}\pi n x}| \leq \sum_{|n|>N/2} |\hat{f}_n| \leq M \sum_{|n|>N/2} c^{|n|} = \frac{2Mc}{1-c}c^{N/2}.$$

□

For nonsmooth functions the convergence of the Fourier series can be much slower, see Figures 1 and 2. The general rule is that smoothness of a function controls the decay rate of $\hat{f}_n$, and thus the convergence rate of ϕ_N to f. For functions that are only assumed C^k for some integer k, one obtains an algebraic decay rate $\mathcal{O}(N^{-k'})$ (where k' is related to k, typically $k' = k+1$) instead of an exponential rate. The following definition will be convenient:

Definition 3.4 (Convergence at spectral speed) An N-term approximation ϕ_N of a function f converges to f at *spectral* speed if $\|\phi_N - f\|$ decays faster than $\mathcal{O}(N^{-p})$ for any $p = 1, 2, \ldots$.

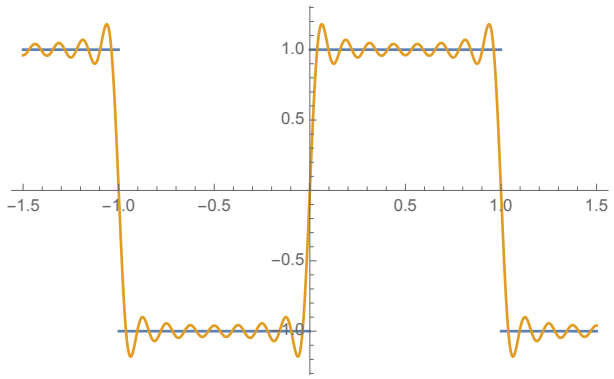

Figure 1: Fourier series approximations of the "square wave" with $N = 30$. The pronounced oscillations at the discontinuity points are known as the *Gibbs effect*.

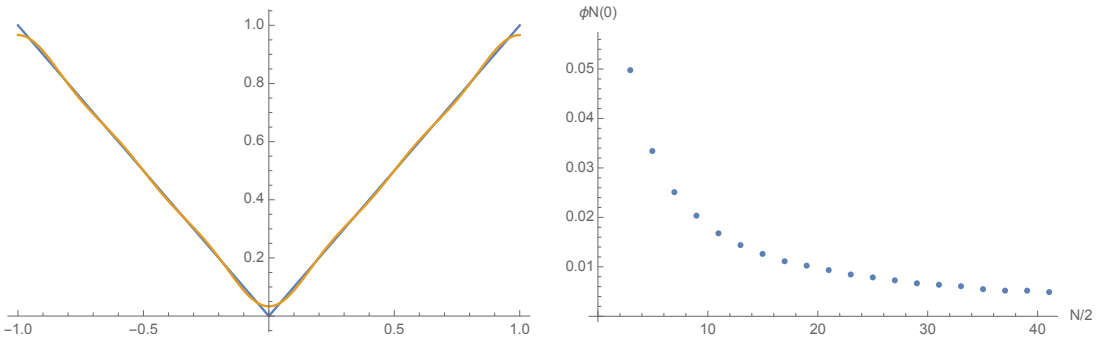

Figure 2: Left: Fourier series approximation of the absolute value function on $[-1, 1]$ with $N = 10$. We see that convergence is quite slow at the singularity point. Right: Convergence of $\phi_N(0)$ to $f(0) = 0$.

The algebra of Fourier expansions Assume f and g are two functions that can be expressed in their Fourier series, i.e.,

$$f(x) = \sum_{n=-\infty}^{\infty} \widehat{f}_n e^{i\pi nx}, \quad g(x) = \sum_{n=-\infty}^{\infty} \widehat{g}_n e^{i\pi nx}.$$

To apply spectral methods, we need to understand how Fourier expansion behaves under pointwise addition and multiplication of functions, and differentiation. It is easy to verify the following formulas:

$$f(x) + g(x) = \sum_{n=-\infty}^{\infty} (\widehat{f}_n + \widehat{g}_n) e^{i\pi nx}, \quad \alpha f(x) = \sum_{n=-\infty}^{\infty} \alpha \widehat{f}_n e^{i\pi nx} \tag{3.6}$$

and

$$f(x) \cdot g(x) = \sum_{n=-\infty}^{\infty} \left(\sum_{m=-\infty}^{\infty} \widehat{f}_{n-m} \widehat{g}_m \right) e^{i\pi nx} = \sum_{n=-\infty}^{\infty} \left(\widehat{f} * \widehat{g}\right)_n e^{i\pi nx}, \tag{3.7}$$

where $*$ denotes the convolution operator, hence $\widehat{(f \cdot g)}_n = (\widehat{f} * \widehat{g})_n$. Moreover, the derivative f' of f has the following Fourier expansion

$$f'(x) = i\pi \sum_{n=-\infty}^{\infty} n \cdot \widehat{f}_n e^{i\pi nx}. \tag{3.8}$$

Example 3.5 (Application to differential equations) Consider the two-point boundary value problem: $y = y(x)$, $-1 \le x \le 1$, solves

$$y'' + a(x)y' + b(x)y = f(x), \quad y(-1) = y(1), \tag{3.9}$$

where a, b, f are 2-periodic and we seek a *periodic solution* y for (3.9). Substituting y, a, b and f by their Fourier series and using (3.6)-(3.8) we obtain an infinite dimensional system of linear equations for the Fourier coefficients $\widehat{y}_n$:

$$-\pi^2 n^2 \widehat{y}_n + i\pi \sum_{m=-\infty}^{\infty} m\widehat{a}_{n-m}\widehat{y}_m + \sum_{m=-\infty}^{\infty} \widehat{b}_{n-m}\widehat{y}_m = \widehat{f}_n, \quad n \in \mathbb{Z}. \tag{3.10}$$

We now truncate the infinite linear system. First, we assume that $\widehat{y}_m = 0$ for $|m| > N/2$. This means that the two summation terms above can be restricted to $|m| \leq N/2$. Furthermore, we only impose equality in (3.10) for $|n| \leq N/2$. This leads to the finite-dimensional linear system of size $(N+1) \times (N+1)$ in the variables $(\widehat{y}_m)_{|m| \leq N/2}$:

$$-\pi^2 n^2 \widehat{y}_n + \sum_{|m| \leq N/2} (i\pi m\widehat{a}_{n-m} + \widehat{b}_{n-m})\widehat{y}_m = \widehat{f}_n, \qquad n = -N/2, \ldots, N/2. \tag{3.11}$$

Remark 3.6 The matrix of (3.11) is in general dense, but our theory predicts that fairly small values of N, hence very small matrices, are sufficient for high accuracy. For instance: choosing $a(x) = f(x) = \cos \pi x$, $b(x) = \sin 2\pi x$ we get

$N = 16$	error of size $\approx 10^{-10}$
$N = 22$	error of size $\approx 10^{-15}$ (which is already hitting the accuracy of computer arithmetic)