Numerical Analysis - Part II

Anders C. Hansen

Lecture 12

Spectral Methods

Large matrices versus small matrices

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 algebraic 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. We will get to know some of these algorithms in Section 4.

However, an enticing alternative to this strategy are methods that produce small matrices in the first place. Although, these matrices will usually not be sparse anymore, the much smaller the size of the matrices renders its solution affordable. The key point for such approximations are better convergence properties requiring much smaller number of parameters.

General idea of spectral methods

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

$$\mathcal{L}u = f \tag{1}$$

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 $\psi_1, \ldots, \psi_N$. 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):

$$\sum_{n=1}^{N} c_n \mathcal{L}\psi_n = f. \tag{2}$$

General idea of spectral methods

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 (2) approximately. Assume that the $(\psi_n)_{1\leq n\leq N}$ 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 (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

$$\sum_{n=1}^{N} c_n \langle \mathcal{L}\psi_n, \psi_m\rangle = \langle f, \psi_m\rangle \qquad \forall m = 1,\dots,N. \tag{3}$$

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 $\tilde{f}_m = \langle f, \psi_m\rangle$.

Fourier approximation of functions

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.

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

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

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

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

Fourier approximation of functions

We consider the *truncated Fourier approximation* of a function f on the interval $[-1, 1]$:

$$f(x) \approx \phi_N(x) = \sum_{n=-N/2+1}^{N/2} \widehat{f}_n e^{i\pi n x}, \quad x \in [-1, 1], \tag{4}$$

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 (4).

Theorem 1 (The de la Valleé Poussin theorem)

If the function f is Riemann integrable and $\widehat{f}_n = \mathcal{O}(n^{-1})$ for $|n| \gg 1$, then $\phi_N(x) = f(x) + \mathcal{O}(N^{-1})$ as $N \to \infty$ for every point $x \in (-1, 1)$ where f is Lipschitz.

Carlesson's Theorem

Let f be an L^2 periodic function with Fourier coefficients $\widehat{f}(n)$. Then

$$\lim_{N\to\infty} \sum_{|n|\leq N} \widehat{f}(n)e^{inx} = f(x)$$

for almost every x.

There exists a L^1 periodic function where the Fourier series diverges everywhere (Kolmogorov), however, the above result can be extended to L^p functions for $p > 1$.

Remark 2 (The Gibbs effect at the end points)

Note that if f is smoothly differentiable then, integrating by parts,

$$\widehat{f_n} = \frac{(-1)^{n+1}}{2\pi i n}[f(1) - f(-1)] + \frac{1}{\pi i n}\widehat{f_n'} = \mathcal{O}(n^{-1}) \text{ for } |n| \gg 1.$$

Since such an f is Lipschitz on $(-1, 1)$, we deduce from Theorem 1 that ϕ_N converges to f there with speed $\mathcal{O}(N^{-1})$. However, convergence with speed $\mathcal{O}(N^{-1})$ is very slow and moreover, we cannot guarantee convergence at the endpoints -1 and 1. In fact, it is possible to show that

$$\phi_N(\pm 1) \to \frac{1}{2}[f(-1) + f(1)] \text{ as } n \to \infty$$

and hence, unless f is periodic we fail to converge.

The Gibbs phenomenon

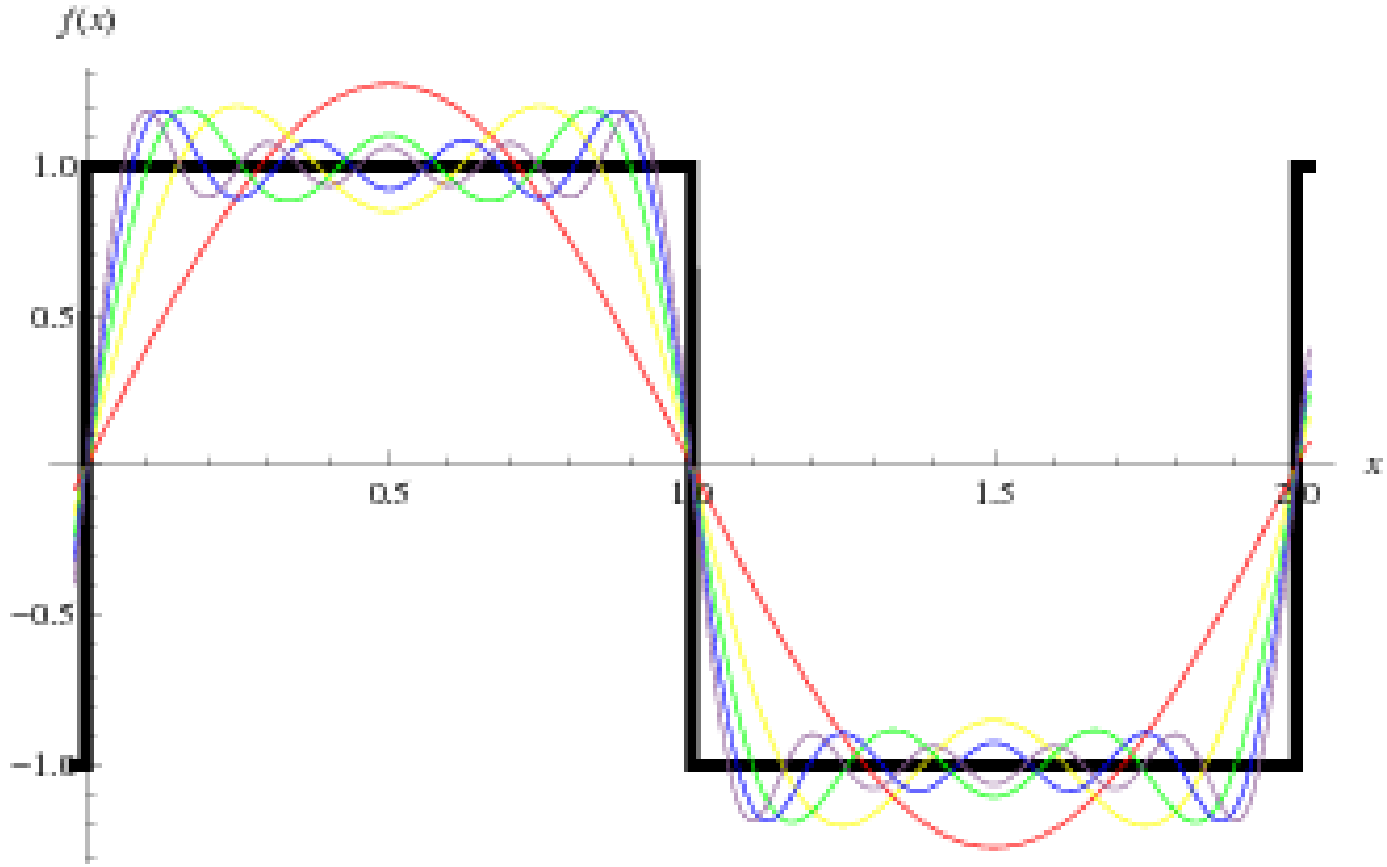


Figure: Convergence of the Fourier series

Fourier approximation for periodic functions

Suppose f is an analytic function in $[-1, 1]$, that can be extended analytically to a closed complex domain Ω. In addition let f be periodic with period 2. In particular, $f^{(m)}(-1) = f^{(m)}(1)$ for all $m \in \mathbb{Z}_+$. Then, by multiple integration by parts, we get

$$\widehat{f_n} = \frac{1}{\pi i n}\widehat{f_n'} = \frac{1}{(\pi i n)^2}\widehat{f_n''} = \frac{1}{(\pi i n)^3}\widehat{f_n'''} = \dots.$$

Thus, we have

$$\widehat{f_n} = \frac{1}{(\pi i n)^m}\widehat{f_n^{(m)}}, \quad m = 0, 1, \dots. \tag{5}$$

But, how large is $\left|\widehat{f_n^{(m)}}\right|$?

Fourier approximation for periodic functions

To answer this question we use Cauchy's theorem of complex analysis, which states that

$$f^{(m)}(x) = \frac{m!}{2\pi i} \int_\gamma \frac{f(z)\, dz}{(z-x)^{m+1}}, \quad x \in [-1,1],$$

where γ is the positively oriented boundary of Ω. Therefore, with $\alpha^{-1} > 0$ being the minimal distance between γ and $[-1,1]$ and $M = \max\{|f(z)| : \ z \in \gamma\} < \infty$, it follows that

$$|f^{(m)}(x)| \leq \frac{m!}{2\pi} \int_\gamma \frac{|f(z)|\ |dz|}{|z-x|^{m+1}} \leq \frac{M \text{ length } \gamma}{2\pi} m!\, \alpha^{m+1},$$

and hence, we can bound $\left|\widehat{f_n^{(m)}}\right| \leq c\, m!\, \alpha^{m+1}$ for some $c > 0$.

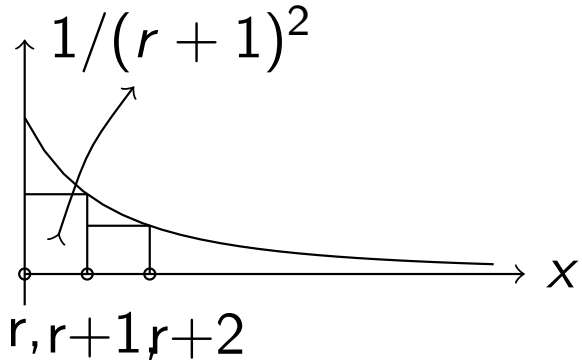
Now, using (5) and the above upper bound,

$$
\begin{aligned}
|\phi_N(x) - f(x)| &= \Big| \sum_{n=-N/2+1}^{N/2} \widehat{f}_n e^{i\pi nx} - \sum_{n=-\infty}^{\infty} \widehat{f}_n e^{i\pi nx} \Big| \\
&\leq \sum_{|n|\geq N/2} |\widehat{f}_n| = \sum_{|n|\geq N/2} \frac{|\widehat{f_n^{(m)}}|}{|\pi n|^m} \leq \frac{cm!\alpha^{m+1}}{\pi^m} \sum_{n=N/2}^{+\infty} \frac{1}{n^m} .
\end{aligned}
$$

Fourier approximation for periodic functions

Using, that for any $r \in \mathbb{N}$, and $m > 1$

$$\sum_{n=r+1}^{+\infty} \frac{1}{n^m} \leq \int_r^{\infty} \frac{dt}{t^m} = \frac{1}{m-1} r^{-m+1},$$



we deduce that

$$|\phi_N(x) - f(x)| \leq c' m! \left(\frac{\alpha}{\pi N}\right)^{m-1}, \quad m \geq 2.$$

Fourier approximation for periodic functions

Finally, we have a competition between $(\alpha/(\pi N))^{m-1}$ and $m!$ for large m. Because of Stirling's formula

$$m! \approx \sqrt{2\pi}\, m^{m+1/2} e^{-m}$$

we have

$$m!\left(\frac{\alpha}{\pi N}\right)^{m-1} \approx \sqrt{2\pi m}\,\frac{m}{e}\left(\frac{\alpha m}{\pi e N}\right)^{m-1}$$

which becomes very small for large N. Hence, $|\phi_N - f| = \mathcal{O}(N^{-p})$ for any $p \in \mathbb{N}$ and we deduce that the Fourier approximation of an analytic periodic function is of infinite order.

Definition 3 (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 $p = 1, 2, \ldots$.

Remark 4

It is possible to prove that there exist constants $c_1, w > 0$ such that $\|\phi_N - f\| \leq c_1 e^{-wN}$ for all $N \in \mathbb{N}$ uniformly in $[-1, 1]$. Thus, convergence is at least at an exponential rate.

The algebra of Fourier expansions

Let $\mathcal{A}$ be the set of all functions $f : [-1, 1] \to C$, which are analytic in $[-1, 1]$, periodic with period 2, and that can be extended analytically into the complex plane. Then $\mathcal{A}$ is a linear space, i.e., $f, g \in \mathcal{A}$ and $\alpha \in \mathbb{C}$ then $f + g \in \mathcal{A}$ and $af \in \mathcal{A}$. In particular, with f and g expressed in its Fourier series, i.e.,

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

we have

$$f(x) + g(x) = \sum_{n=-\infty}^{\infty} (\widehat{f}_n + \widehat{g}_n) e^{i\pi nx}, \quad \alpha f(x) = \sum_{n=-\infty}^{\infty} \alpha \widehat{f}_n e^{i\pi nx}. \quad (6)$$

The algebra of Fourier expansions

Moreover,

$$f(x) \cdot g(x) = \sum_{n=-\infty}^{\infty} \left(\sum_{m=-\infty}^{\infty} \widehat{f}_{n-m} \widehat{g}_m \right) e^{i\pi nx} = \sum_{n=-\infty}^{\infty} \left(\widehat{f} * \widehat{g} \right)_n e^{i\pi nx}, \tag{7}$$

where $*$ denotes the convolution operator (recall `https://en.wikipedia.org/wiki/Convolution`), hence $\widehat{(f \cdot g)}_n = (\widehat{f} * \widehat{g})_n$.
Moreover, if $f \in \mathcal{A}$ then $f' \in \mathcal{A}$ and

$$f'(x) = i\pi \sum_{n=-\infty}^{\infty} n \cdot \widehat{f}_n e^{i\pi nx}. \tag{8}$$

Since $\{\widehat{f}_n\}$ decays faster than $\mathcal{O}(n^{-p})$ for any $p \in \mathbb{N}$, this provides that all derivatives of f have rapidly convergent Fourier expansions.

Application to differential equations

Consider the two-point boundary value problem: $y = y(x)$, $-1 \leq x \leq 1$, solves

$$y'' + a(x)y' + b(x)y = f(x), \quad y(-1) = y(1), \tag{9}$$

where $a, b, f \in \mathcal{A}$ and we seek a *periodic solution* $y \in \mathcal{A}$ for (9). Substituting y, a, b and f by their Fourier series and using (6)-(8) we obtain an infinite dimensional system of linear equations for the Fourier coefficients $\widehat{y}_n$:

$$-\pi^2 n^2 \widehat{y}_n + i\pi \sum_{m=-\infty}^{\infty} m\widehat{a}_{n-m}\widehat{y}_m + \sum_{m=-\infty}^{\infty} \widehat{b}_{n-m}\widehat{y}_m = \widehat{f}_n, \quad n \in \mathbb{Z}. \tag{10}$$

Application to differential equations

Since $a, b, f \in \mathcal{A}$, their Fourier coefficients decrease rapidly, like $\mathcal{O}(n^{-p})$ for every $p \in \mathbb{N}$. Hence, we can truncate (10) into the N-dimensional system

$$-\pi^2 n^2 \widehat{y}_n + i\pi \sum_{m=-N/2+1}^{N/2} m\widehat{a}_{n-m}\widehat{y}_m + \sum_{m=-N/2+1}^{N/2} \widehat{b}_{n-m}\widehat{y}_m = \widehat{f}_n, \quad (11)$$

where $n = -N/2+1, \ldots, N/2$.

Remark 5

The matrix of (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 $a(x) = f(x) = \cos \pi x$, $b(x) = \sin 2\pi x$ (which incidentally even leads to a sparse matrix) we get

$N = 16$	error of size 10^{-10}
$N = 22$	error of size 10^{-15} (which is already hitting ϵ_{Mach}).

The discrete Fourier transform (DFT)

Recall the DFT:

`https://en.wikipedia.org/wiki/DFT_matrix`)

Exercise: Prove that the DFT is a unitary matrix.

Computation of Fourier coefficients (DFT)

We have to compute

$$\widehat{f_n} = \frac{1}{2}\int_{-1}^{1} f(t)e^{-i\pi nt}\,dt, \quad n \in \mathbb{Z}. \tag{12}$$

For this, suppose we wish to compute the integral on $[-1,1]$ of a function $h \in \mathcal{A}$ by means of the Riemann sums on the uniform partition

$$\int_{-1}^{1} h(t)\,dt \approx \frac{2}{N}\sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right). \tag{13}$$

This is known as a *rectangle rule*. We want to know how good this approximation is. As in the definition of the DFT, let $\omega_N = e^{2\pi i/N}$. Then we have

$$\begin{aligned}\frac{2}{N}\sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right) &= \frac{2}{N}\sum_{k=-N/2+1}^{N/2}\sum_{n=-\infty}^{\infty}\widehat{h}_n e^{2\pi ink/N} \\ &= \frac{2}{N}\sum_{n=-\infty}^{\infty}\widehat{h}_n \sum_{k=-N/2+1}^{N/2}\omega_N^{nk}.\end{aligned} \tag{14}$$

Computation of Fourier coefficients (DFT)

Since $\omega_N^N = 1$ we have

$$\sum_{k=-N/2+1}^{N/2} \omega_N^{nk} = \omega_N^{-n(N/2-1)} \sum_{k=0}^{N-1} \omega_N^{nk} = \begin{cases} N, & n \equiv 0 \pmod N, \\ 0, & n \not\equiv 0 \pmod N, \end{cases}$$

`https://en.wikipedia.org/wiki/DFT_matrix`)
and we deduce that

$$\frac{2}{N} \sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right) = 2 \sum_{r=-\infty}^{\infty} \widehat{h}_{Nr} \,.$$

Computation of Fourier coefficients (DFT)

Hence, the error committed by the Riemann approximation is

$$e_N(h) := \frac{2}{N} \sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right) - \int_{-1}^{1} h(t)\,dt = 2 \sum_{r=-\infty}^{\infty} \widehat{h}_{Nr} - 2\widehat{h}_0$$

$$= 2 \sum_{r=1}^{\infty} \left(\widehat{h}_{Nr} + \widehat{h}_{-Nr}\right).$$

Since $h \in \mathcal{A}$, its Fourier coefficients decay at spectral rate, namely $\widehat{h}_{Nr} = \mathcal{O}((Nr)^{-p})$, and hence the error of the Riemann sums approximation (13) decays spectrally as a function of N,

$$e_N(h) = \mathcal{O}(N^{-p}) \quad \forall p \in \mathbb{N}.$$

Computation of Fourier coefficients (DFT)

Going back to the computation of the Fourier coefficients (12), we see that we may compute the integral of $h(x) = \frac{1}{2} f(x) e^{-i\pi n x}$ by means of the Riemann sums, and this gives a spectral method for calculating the Fourier coefficients of f:

$$\widehat{f}_n \approx \frac{1}{N} \sum_{k=-N/2+1}^{N/2} f\left(\frac{2k}{N}\right) \omega_N^{-nk}, \qquad n = -N/2+1, \ldots, N/2\,. \tag{15}$$

Remark 6

One can recognise that formula (15) is the *discrete Fourier transform (DFT)* of the sequence $(y_k) = \left(f\left(\frac{2k}{N}\right)\right)$, see Definition **??**, hence not only have we a spectral rate of convergence, but also a fast algorithm (FFT) of computing the Fourier coefficients.

The fast Fourier transform (FFT)

The *fast Fourier transform (FFT)* is a computational algorithm, which computes the leading N Fourier coefficients of a function in just $\mathcal{O}(N \log_2 N)$ operations. We assume that N is a power of 2, i.e. $N = 2m = 2^p$, and for $\boldsymbol{y} \in \Pi_{2m}$, denote by

$$\boldsymbol{y}^{(\mathrm{E})} = \{y_{2j}\}_{j\in\mathbb{Z}} \qquad \text{and} \qquad \boldsymbol{y}^{(\mathrm{O})} = \{y_{2j+1}\}_{j\in\mathbb{Z}}$$

the even and odd portions of $\boldsymbol{y}$, respectively. Note that $\boldsymbol{y}^{(\mathrm{E})}, \boldsymbol{y}^{(\mathrm{O})} \in \Pi_m$.

The fast Fourier transform (FFT)

To execute FFT, we start from vectors of unit length and in each s-th stage, $s = 1...p$, assemble 2^{p-s} vectors of length 2^s from vectors of length 2^{s-1} with

$$x_\ell = x_\ell^{(\mathrm{E})} + \omega_{2^s}^\ell x_\ell^{(\mathrm{O})}, \qquad \ell = 0, \ldots, 2^{s-1} - 1. \tag{16}$$

Therefore, it costs just s products to evaluate the first half of $\boldsymbol{x}$, provided that $\boldsymbol{x}^{(\mathrm{E})}$ and $\boldsymbol{x}^{(\mathrm{O})}$ are known. It actually costs nothing to evaluate the second half, since

$$x_{2^{s-1}+\ell} = x_\ell^{(\mathrm{E})} - \omega_{2^s}^\ell x_\ell^{(\mathrm{O})}, \qquad \ell = 0, \ldots, 2^{s-1} - 1.$$

Altogether, the cost of FFT is $p2^{p-1} = \frac{1}{2} N \log_2 N$ products.