Chapter 1
Introduction to Spectral Methods

1.1 Basic Concepts

Our starting point is a basic question. Suppose we have an equation for some vector function $u(x)$, $x \in \Omega \subseteq \mathbb{R}^n$

$$\mathscr{L}u = f\,, \tag{1.1}$$

with boundary conditions

$$\mathscr{B}u = 0, \qquad x \in \partial\Omega\,, \tag{1.2}$$

where $\mathscr{L}$ and $\mathscr{B}$ are some linear operator. How can we find the best approximation of the unknown function u? One of possible methods is based on the wide class of discretization schemes known as *method of weighted residuals (MWR)*. The idea of the method is to approximate the unknown function $u(x)$ by a sum of so-called *trial* or *basis* functions $\phi_n(x)$

$$\widetilde{u}(x) = \sum_{n=0}^{N} a_n \phi_n(x), \tag{1.3}$$

where a_n are unknown coefficients to be determined and the tilde denotes an approximate solution of (1.1). If one substitute the approximation (1.3) into Eq. (1.1), *the residual* R can be calculated as

$$R = \mathscr{L}\widetilde{u} - f \tag{1.4}$$

Due to the fact that $\widetilde{u}$ is different from the exact solution u, the residual R does not vanish for all $x \in \Omega$. The next step is to determine unknown coefficients a_n so that the chosen function approximates the exact solution in the best way. To this end, *test* or *weighting* functions $\chi_n(x)$, $n = 0, \ldots, N$ are selected so that the residual function R is minimized, e.g., the weighted average of the residual over the domain of interest is set to zero,

$$\int_\Omega \chi_n(x) R\, dx = 0\,, \qquad n = 0, \ldots, N\,. \tag{1.5}$$

The various methods differ mainly in the choise of trial and test functions and their minimization strategies [2, 1].

1.1.0.1 Various numerical methods

The choice of the trial functions $\phi_n(x)$ is one of the key difference between finite-element and finite-difference methods on the one hand and spectral methods on the other hand. In the case of finite-element methods the domain Ω is divided into small finite intervals and $\phi_n(x)$ are typically choosen to be a *local* polynomial of fixed degree, defined on these sub-intervals only. The finite-difference methods have a local charachter as well. Generally, the unknown function $u(x)$ is approximated by a sequence of overlapping polynomials of low order, interpolating the solution at a given set of discretization points and the rersult is represented in the form of weighted sum of values of $u(x)$ at the interpolation points. In contrast, the trial functions for spectral methods are *global* smooth functions, e.g., Fourier or Chebyshev series. The particular choice of the trial functions is usually connected to the geometry of the problem in question. For instance, on periodic intervals, the sines and cosines of a Fourier series, which automatically satisfy boundary conditions are used. For non-periodic problem, Chebyshev or Legendere polynomials are more natural choise [1].

1.1.0.2 Various minimization strategies

The choice of the test functions χ_n distinguishes between the three most commonly used spectral schemes, namely

1. *Galerkin method.*
   The test functions $\chi_n(x)$ are the same as the trial functions and each $\phi_n(x)$ satisfy the boundary condition $\mathscr{B}\phi_n = 0$. Since $\phi_n = \chi_n$ for $n = 0, \ldots, N$, Eq. (1.5) is equivalent to

$$\int_\Omega \phi_n R = 0 \Leftrightarrow \int_\Omega \phi_n \left(\mathscr{L}\widetilde{u} - f\right) = 0 \Leftrightarrow \int_\Omega \phi_n \sum_{k=0}^{N} a_k \phi_k = \int_\Omega \phi_n f \Leftrightarrow$$

$$\sum_{k=0}^{N} L_{nk} a_k = \int_\Omega \phi_n f\,,$$

   where $L_{nk} = \int_\Omega \phi_n \mathscr{L} \phi_k$. Solving the obtained linear system one can get all $N+1$ unknown coefficients a_k.
2. *Tau method.* The test functions χ_n are the same as the trial functions, but ϕ_n do not need to satisfy (1.2). The boundary conditions are enforced by an additional set of equations. In order to find this set, let us consider an orthonormal basis $\{\psi_l\}$, $l = 0, \ldots, M$, where $M < N$ on the $\partial\Omega$ and expand $\mathscr{B}\phi_n$ upon it:

$$\mathscr{B}\phi_n = \sum_{l=0}^{M} B_{ln}\psi_l\,.$$

Equation (1.2) then becomes

$$\mathscr{B}u = 0 \Leftrightarrow \sum_{k=0}^{N}\sum_{l=0}^{M} a_k B_{lk}\psi_l = 0 \Leftrightarrow \sum_{k=0}^{N} B_{lk}a_k = 0, \quad l = 0,\ldots,M\,.$$

The resulting linear system of $N+1$ equations consists of $N-M$ first rows of the Galerkin system, presented above, and $M+1$ additional equations for boundary conditions:

$$\sum_{k=0}^{N} L_{nk}a_k = \int_\Omega \phi_n f, \qquad n = 0,\ldots,N-M-1,$$
$$\sum_{k=0}^{N} B_{lk}a_k = 0, \qquad l = 0,\ldots,M\,.$$

3. *Collocation (pseudospectral) method.* The test functions are represented by a delta functions at special points x_n, called *collocation points*, i.e., $\chi_n = \delta(x - x_n)$. In other words this approach requires Eq. (1.1) to be satisfied exactly at the collacation points x_n. Then the condition (1.5) reads:

$$\int_\Omega \chi_n R = 0 \Leftrightarrow \int_\Omega \delta(x-x_n)R = 0 \Leftrightarrow \mathscr{L}\widetilde{u}(x_n) = f(x_n) \Leftrightarrow$$
$$\sum_{k=0}^{N} a_k \mathscr{L}\phi_k(x_n) = f(x_n)\,, \qquad n = 0,\ldots,N\,.$$

The boundary conditions can be imposed as in the tau method.

To sum up: Galerkin and tau methods are implemented in terms of the expansion coefficients, whereas the collocation method is implemented in terms of the physical space values of the unknown function.

1.2 Fourier and Chebyshev series

1.2.1 The Fourier System

A Fourier series is an expansion of a periodic function in terms of an infinite sum of sines and cosines. Consider a periodic integrable function $f(x)$, $x \in [-\pi, \pi]$. The numbers

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(nt)\,dt\,, \qquad n \geq 0\,,$$
$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(nt)\,dt\,, \qquad n \geq 1$$

are called *the Fourier coefficients* of f. The *Fourier series* of the function $f(x)$ is given by

$$\boxed{f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}\Big(a_n\cos(nx) + b_n\sin(nx)\Big)\,.}$$

If the function $f(x)$ is periodic on some interval $[-L, L]$, a simple change of variables

$$x' = \frac{xL}{\pi}$$

can be used to transform the interval of integration. In this case the Fourier series read

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}\Big(a_n\cos\big(\frac{n\pi x'}{L}\big) + b_n\sin\big(\frac{n\pi x'}{L}\big)\Big)$$

with

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x')\cos\big(\frac{n\pi x'}{L}\big)\,dx'\,, \qquad n \geq 0\,,$$
$$b_n = \frac{1}{L}\int_{-L}^{L} f(x')\sin\big(\frac{n\pi x'}{L}\big)\,dx'\,, \qquad n \geq 1\,.$$

One of the main questions is to decide when Fourier series converge, and when the sum is equal to the original function. If a function is *square-integrable* on the interval $[-\pi, \pi]$, then the Fourier series converges to the function *at almost every point*. In particular, the Fourier series converges *absolutely* and *uniformly* to $f(x)$ whenever its derivative is square-integrable. A piecewise regular function that has a finite number of finite discontinuities and a finite number of extrema can be expanded in a Fourier series which converges to the function at continuous points and the mean of the positive and negative limits at points of discontinuity (a Dirichlet condition, see, e.g., [12]). As a result, near points of discontinuity the n'th partial sum of the Fourier series has large oscillations and a so-called *Gibbs phenomenon* or "rinding" occurs [13, 6, 12].

1.2.1.1 Exponential Fourier series

The notion of a Fourier series can also be extended to complex coefficients. Consider a real-valued function $f(x)$. Then using Euler's formula we can write

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}\,,$$

where Fourier coefficients are given by

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}dx.$$

The Fourier coefficients a_n and b_n can be found as

$$\begin{aligned} a_n &= c_n + c_{-n}, \qquad n = 0, 1, 2, \dots, \\ b_n &= i(c_n - c_{-n}), \qquad n = 1, 2, \dots. \end{aligned}$$

For a function periodic in $[-\frac{L}{2}, \frac{L}{2}]$ one obtains

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{in2\pi x}{L}}$$

with

$$c_n = \frac{1}{2L}\int_{-L}^{L} f(x)e^{-\frac{in2\pi x}{L}}dx.$$

These equations are the basis for *the Fourier transform*, which is obtained by transforming c_n from a discrete variable to a continuous one as the length $L \to \infty$.

1.2.1.2 Fourier Transformation

The Fourier transform can be considered as a generalization of the complex Fourier series in the limit $L \to \infty$. Replacing the discrete coefficient c_n with the continuous $F(k)dk$, $n/L \mapsto k$ and changing the sum to an integral one obtains for an integrable function $f(x)$

$$f(x) = \int_{-\infty}^{\infty} F(k)e^{2\pi ikx}dk, \tag{1.6}$$

$$F(k) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ikx}dx. \tag{1.7}$$

Here,

$$\boxed{F(k) = \mathscr{F}[f(x)](k) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ikx}dx} \tag{1.8}$$

is called *the forward Fourier transform*, and

$$\boxed{f(x) = \mathscr{F}^{-1}[F(k)](x) = \int_{-\infty}^{\infty} F(k)e^{2\pi ikx}dk} \tag{1.9}$$

is called *the inverse Fourier transform*. However, other notation can also be found in the literature. Especially physicists prefer to write the Fourier transform, presented above in terms of the angular frequency ω:

$$F(\omega) = \mathscr{F}[f(t)](\omega) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt, \tag{1.10}$$

$$f(t) = \mathscr{F}^{-1}[F(\omega)](t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \tag{1.11}$$

Basic properties of the Fourier transform: Let us consider two integrable functions $f(x)$ and $g(x)$. Then

- *Linearity:* For any complex numbers α and β

$$\mathscr{F}[\alpha f(t) + \beta g(t)] = \alpha \mathscr{F}[f(t)] + \beta \mathscr{F}[g(t)];$$

- *Convolution:*

$$\mathscr{F}[f(t) \cdot g(t)] = \mathscr{F}[f(t)] * \mathscr{F}[g(t)],$$
$$\mathscr{F}[f(t) * g(t)] = \mathscr{F}[f(t)] \cdot \mathscr{F}[g(t)];$$

- *Translation:* For any real t_0

$$\mathscr{F}[f(t - t_0)] = e^{-i\omega t_0} \mathscr{F}[f(t)];$$

- *Scaling:* For all non-zero real numbers a

$$\mathscr{F}[f(at)] = \frac{1}{|a|} F\left(\frac{\omega}{a}\right);$$

- *Derivative:*

$$\mathscr{F}\left[\frac{d^n}{dt^n} f(t)\right] = (i\omega)^n \mathscr{F}[f(t)].$$

Examples

1. *Fourier Transform–Gaussian.*
   Let us consider a function $f(t) = e^{-at^2}$, $\mathrm{Re}(a) > 0$. Then

$$\mathscr{F}[f(t)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-at^2} (\cos(\omega t) - i \sin(\omega t))\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-at^2} \cos(\omega t)\, dt = \frac{1}{2\sqrt{a}} e^{-\omega^2/4a}.$$

   That is, the Gaussian function is its own Fourier transform for some choice of a.
2. *Fourier Transform–Cosine.*
   Consider $f(t) = \cos(at)$. Then one gets:

$$\mathscr{F}[f(t)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega t} \left(\frac{e^{iat} + e^{-iat}}{2}\right) dt = \sqrt{2\pi} \left(\frac{\delta(\omega - a) + \delta(\omega + a)}{2}\right).$$

1.2.1.3 Discrete Fourier Transformation

Now let us consider a generalization to the case of a discrete function. The sequence of N complex numbers $x_0, ..., x_{N-1}$ is transformed into the sequence of N complex numbers $X_0, ..., X_{N-1}$ by *the discrete Fourier transformation (DFT)* according to the formula:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn}, \qquad k = 0, \ldots, N-1. \tag{1.12}$$

The inverse discrete Fourier transform (IDFT) is defined as

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N}kn}, \qquad n = 0, \ldots, N-1. \tag{1.13}$$

The DTF (1.12) is a linear transformation, so one can condiser it as a transformation of the vector $x = (x_0, x_1, \ldots, x_{N-1})^T$ to the vector $X = (X_0, X_1, \ldots, X_{N-1})^T$ of the same lenght via the relation

$$X = \hat{A}x,$$

where

$$\hat{A}_{mn} = \exp\left(-2\pi i \frac{(m-1)(n-1)}{N}\right).$$

That is,

$$\hat{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & \ldots & 1 \\ 1 & e^{-\frac{2\pi i}{N}} & e^{-\frac{4\pi i}{N}} & e^{-\frac{6\pi i}{N}} & \ldots & e^{-\frac{2\pi i}{N}(N-1)} \\ 1 & e^{-\frac{4\pi i}{N}} & e^{-\frac{8\pi i}{N}} & e^{-\frac{12\pi i}{N}} & \ldots & e^{-\frac{2\pi i}{N}2(N-1)} \\ 1 & e^{-\frac{6\pi i}{N}} & e^{-\frac{12\pi i}{N}} & e^{-\frac{18\pi i}{N}} & \ldots & e^{-\frac{2\pi i}{N}3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-\frac{2\pi i}{N}(N-1)} & e^{-\frac{2\pi i}{N}2(N-1)} & e^{-\frac{2\pi i}{N}3(N-1)} & \ldots & e^{-\frac{2\pi i}{N}(N-1)^2} \end{pmatrix} \tag{1.14}$$

Note taht the vectors $e^{\frac{2\pi i}{N}kn}$ form an orthogonal basis over the set of N-dimensional complex vectors:

$$\sum_{n=0}^{N-1} \left(e^{\frac{2\pi i}{N}kn}\right)\left(e^{-\frac{2\pi i}{N}k'n}\right) = N\,\delta_{kk'},$$

where $\delta_{kk'}$ is the Kronecker delta. In addition, if the DFT (1.12) is evaluated for all integers k then the resulting infinite sequence is a periodic extension of the DFT, periodic with period N, i.e.

$$X_{k+N} = \sum_{n=0}^{N-1} x_n\, e^{-\frac{2\pi i}{N}(k+N)n} = \sum_{n=0}^{N-1} x_n\, e^{-\frac{2\pi i}{N}kn} \underbrace{e^{-2\pi i n}}_{1} = \sum_{n=0}^{N-1} x_n\, e^{-\frac{2\pi i}{N}kn} = X_k\,.$$

If $x_0,\ldots,x_{N-1}$ are real numbers then the DFT (1.12) obeys the symmetry:

$$X_{N-k} = \overline{X}_k\,,$$

where the overline denotes complex conjugation. The subscripts are interpreted modulo N. In fact,

$$X_{N-k} = \sum_{n=0}^{N-1} x_n\, e^{-\frac{2\pi i}{N}(N-k)n} = \sum_{n=0}^{N-1} x_n\, e^{\frac{2\pi i}{N}kn} \underbrace{e^{-2\pi i n}}_{1} = \sum_{n=0}^{N-1} \overline{x_n\, e^{-\frac{2\pi i}{N}kn}} = \overline{X_k}\,.$$

As a result of the above relation, a periodic function will contain transformed peaks in not one, but two places. This happens because the periods of the input data become split into "positive" and "negative" frequency complex components. Therefore, the DFT output for real inputs is half redundant, and one obtains the complete information by only looking at roughly half of the outputs $X_0,\ldots,X_{N-1}$. The next point is that the component X_0 is always real for real data. The DFT can be computed efficiently using *a Fast Fourier transform (FFT)* algorithm.With the FFT, the resulting scheme takes $\mathcal{O}(N\,logN)$ arithmetic operations instead of $\mathcal{O}(N^2)$ for the computing a DFT of N points directly.

1.2.1.4 Fast Fourier Transform

The most common FFTs are based on the co-called *Cooley-Tukey algorithm*, named after J. W. Cooley and J. Tukey [3]. However, later it was discovered [7] that the authors had re-invented the algorithm, known to C. F. Gauss around 1805, who used it to interpolate the trajectories of asteroids [5].

The simplest and most common form of the Cooley-Tukey algorithm is based on the idea of Danielson and Lanczos [4] to divide a DFT (1.12) of size N into two interleaved DFTs of size $N/2$, one of those is formed from the even-numbered points of original N, whereas another one from the odd-numbered points, i.e.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} = \sum_{n=0}^{N/2-1} x_{2n}\, e^{-\frac{2\pi i}{N}(2n)k} + \sum_{n=0}^{N/2-1} x_{2n+1}\, e^{-\frac{2\pi i}{N}(2n+1)k} =$$
$$\sum_{n=0}^{N/2-1} x_{2n}\, e^{-\frac{2\pi i}{N/2}nk} + e^{-\frac{2\pi i}{N}k} \sum_{n=0}^{N/2-1} x_{2n+1}\, e^{-\frac{2\pi i}{N/2}nk} = X_k^e + W^k X_k^o\,. \quad (1.15)$$

Here, X_k^e and X_k^o denote the DFT's of the even- and odd-indexed inputs, respectively and the complex constant $W = e^{\frac{2\pi i}{N}}$ stands for *a twiddle factor*. Notice that although k in the last equation varies from 0 to $N-1$, both transforms X_k^e and X_k^o are periodic in k with length $N/2$,

$$X_{k+N/2}^e = X_k^e\,, \qquad X_{k+N/2}^o = X_k^o\,.$$

In addition, for the twiddle factor W

$$W^{k+N/2} = e^{\frac{-2\pi i}{N}(k+N/2)} = e^{-\pi i} W^k = -W^k .$$

That is, the whole DFT can be written as

$$X_k = \begin{cases} X_k^e + W^k X_k^o , & k < N/2 , \\ X_{k-N/2}^e - W^{k-N/2} X_{k-N/2}^o , & k > N/2 . \end{cases}$$

Assuming that N is an integer power of 2, e.g., $N = 2^p$, one can repeat the reduction procedure, described above recursively, i.e,

$$X_k = X_k^e + W^k X_k^o = X_k^{ee} + W^k X_k^{eo} + W^k X_k^{oe} + W^{2k} X_k^{oo} =$$

$$\vdots$$

$$p \,\text{steps}$$

$$\vdots$$

$$= \underbrace{X_k^{ee\cdots e} + \cdots + W^{(\cdots)} X_k^{eooe\cdots o} + \cdots + W^{pk} X_k^{oo\cdots o}}_{N} .$$

Note that on the last step of recursion we have subdivided the data to transforms of length one, i.e, for every even- and odd- pattern there is a one-point transform that is just equals to the one of the input numbers x_n [9],

$$X_k^{eooe\cdots o} = x_n \qquad \text{for some } n .$$

The relation between n and the corresponding even- and odd pattern can be found be use of the so-called *bit reversal algorithm* [9], namely one reverse the pattern of even's and odd's and suppose $e := 0$, $o := 1$ and get the value of n in binary form. That is, the whole scheme can be formulated as follows [9]:

- Consider the input vector x_n and rearrange it into bit-reversed order;
- Combine adjacent elements to get two-point transform;
- Combine adjacent pairs to get four-point transform, etc.;
- Repeat till both halves of the whole data set are combined into the final transform.

Notice that each combination takes $\mathcal{O}(N)$ operations and one has $\mathcal{O}(\log N)$ combinations, so the whole algorithm is of order $\mathcal{O}(N \log N)$.

1.2.2 The Advection Equation

Let us consider a one-dimensional advection equation

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \tag{1.16}$$

Here $u = u(x,t)$, $x \in \mathbb{R}$, c is a nonzero constant velocity. Equation (1.16) describes the motion of a scalar u as it is advected by a known velocity field. The unique solution of (1.16) is determined by a given initial condition $u(x,0) = u_0(x)$ as

$$u(x,t) = u_0(x - ct). \tag{1.17}$$

The solution (1.17) is just an initial function $u_0(x)$ shifted by ct to the right (for $c > 0$) or to the left (for $c < 0$). Our goal is to solve Eq. (1.16) on the domain $x \in [0, 2\pi]$ with periodic boundary conditions, i.e., $u(0,t) = u(2\pi,t)$ by means of the Galerkin method (see Section 1.1.0.2). First of all we rewrite Eq. (1.16) in *the weak form*, i.e., for any test function $\chi(x,t)$

$$\langle \partial_t u | \chi \rangle + c\langle \partial_x u | \chi \rangle = 0\,,$$

where $\langle f,g\rangle := \int_0^{2\pi} f(x)\overline{g(x)}\,dx$ following inner product notation. Choosing the trigonometric polynomial, presented in Section 1.2 as the trial functions, $\phi_k(x) = \exp(ikx)$, the approximated solution $\widetilde{u}$ of (1.16) is represented as

$$\widetilde{u}(x,t) = \sum_{k=-N/2}^{N/2} \widehat{u}_k(t)\, e^{ikx}\,.$$

According to the Galerkn method (see section (1.1.0.2)) the trial functions $\phi_k(x)$ and the test functions $\chi(x)$ are essentialy the same. This reduces the problem in question to

$$\langle \partial_t \widetilde{u} | e^{ikx} \rangle + c\langle \partial_x \widetilde{u} | e^{ikx} \rangle = 0, \qquad \forall t > 0, \forall k = -N/2, \ldots, N/2.$$

Using the orthogonality relation $\langle e^{ilx} | e^{ikx} \rangle = 2\pi\delta_{lk}$, where δ_{lk} is the Kronecker delta, we simplify the relation above for each k to

$$\langle \partial_t \sum_{l=-N/2}^{N/2} \widehat{u}_l(t) e^{ilx} | e^{ikx} \rangle + c\langle \partial_x \sum_{l=-N/2}^{N/2} \widehat{u}_l(t) e^{ilx} | e^{ikx} \rangle = 0 \Leftrightarrow$$

$$\boxed{\frac{d\widehat{u}_k(t)}{dt} + ikc\widehat{u}_k(t) = 0, \quad \forall t > 0, \quad \forall k = -N/2, \ldots, N/2\,.}$$

With Fourier transformed initial conditions $\widehat{u}_k(0) = \frac{1}{2\pi}\langle u_0(x) | e^{ikx} \rangle$ this coupled system of ordinary differential equations involves a standart initial value problem and can be integrated in time (using, e.g., a Runge–Kutta technique (see Appendix A)) to find a solution.