Spectral Methods for Time-Dependent Problems

David Gottlieb and Jan S. Hesthaven

Division of Applied Mathematics
Brown University
Providence
RI 02912

1

Modes, Nodes and Spectral Codes

2

From Local to Global Approximation

When seeking approximate solutions to partial differential equations, a crucial operation is the approximation of spatial derivatives, e.g., classic finite difference methods are often designed to yield exact differentiation if the solution is a low order polynomial. Indeed, let us consider the smooth function, $u(x)$, specified at $N+1$ discrete grid points, $x_0, \ldots, x_N$. Throughout the text we shall assume that N is even.

A $2m$'th local interpolating polynomial to $u(x)$ in the neighborhood of x_j is obtained as

$$u(x) = \sum_{|k| \leq m} u_{j+k} L_{j+k}(x) \quad ,$$

where we have the grid function, $u_{j+k} = u(x_{j+k})$, and the Lagrange interpolation polynomial, $L_{j+k}(x)$, given as

$$L_{j+k}(x) = \prod_{\substack{|l| \leq m \\ l \neq k}} \frac{x - x_{j+l}}{x_{j+k} - x_{j+l}} \quad . \tag{2.1}$$

We shall seek a 2nd order polynomial representation of $u(x)$, i.e., corresponding to $m = 1$ in Eq.(2.1). Assuming, for simplicity, that the grid is equidistant, we obtain

$$\begin{aligned} u(x) = \; & \frac{1}{2\Delta x^2}(x - x_j)(x - x_{j+1})u_{j-1} - \\ & \frac{1}{\Delta x^2}(x - x_{j-1})(x - x_{j+1})u_j + \\ & \frac{1}{2\Delta x^2}(x - x_{j-1})(x - x_j)u_{j+1} \quad , \end{aligned} \tag{2.2}$$

where $\Delta x = x_j - x_{j-1}$ represents the constant distance between the grid points.

To approximate the derivative of $u(x)$ at the grid point, x_j, we utilize the local polynomial approximation, Eq.(2.2), and recover

$$\left.\frac{du}{dx}\right|_{x_j} \simeq \frac{u_{j+1} - u_{j-1}}{2\Delta x} \quad .$$

One can recognize this as the classic centered finite difference formula of 2nd order accuracy.

For problems processing a significant spatial variation it is clear that using a low order local approximation will require a very fine grid to ensure an acceptable accuracy. This translates to severe requirements on the computational resources when addressing problems of interest to science and engineering.

This naturally poses the question as to whether alternative schemes, overcoming this need for a fine grid, can be formulated. Indeed, as an extreme alternative to the local approximation utilized in finite difference schemes, one can think of approximating functions and their derivatives using a global method, i.e., an approach in which all available information is utilized to construct the approximation and its derivatives In between the local and the global approximation schemes reside a large number of methods, generally known as high-order accurate methods, i.e., methods for which the local solution is assumed to have a large degree of smoothness and, thus, is well represented by a high-order local polynomial.

Example 1. Consider the scalar hyperbolic equation

$$\begin{aligned} \frac{\partial u}{\partial t} &= -2\pi \frac{\partial u}{\partial x} \quad , \\ u(0,t) &= u(2\pi,t) \quad , \\ u(x,0) &= \exp\left[\sin(x)\right] \quad , \end{aligned} \tag{2.3}$$

where the smooth function, $u(x,t) \in C^{\infty}[0,2\pi]$, is periodic and the initial condition is periodically extended.

The exact solution to Eq.(2.3) is given as

$$u(x,t) = \exp\left[\sin(x - 2\pi t)\right] \quad ,$$

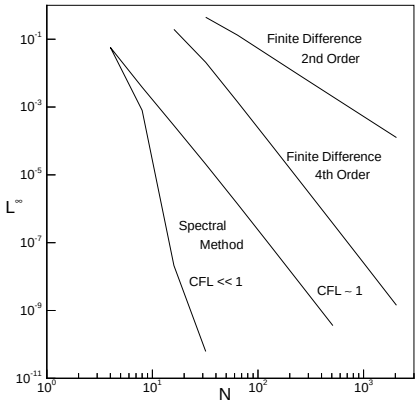

figure 2.1. The maximum pointwise error, L^∞, measured at t=π obtained using a 2nd order, a 4th order and global spectral scheme as a function of the total number of points, N.

i.e., the initial condition is propagating towards increasing x at the speed of the propagation, 2π.

We solve this equation at an equidistant grid

$$x_j = j\Delta x = \frac{2\pi j}{N+1} \quad , \quad j \in [0, \ldots, N] \quad .$$

In a method-of-lines approach we use a 4th order explicit Runge-Kutta scheme to advance the solution in time with the time-step taken to be well below the stability limit.

To approximate the spatial derivatives at the grid-points, x_j, we shall consider 3 different schemes.

Local Finite Difference Scheme: The second order centered finite difference approximation to the spatial derivative is

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \frac{u_{j+1} - u_{j-1}}{2\Delta x} \quad ,$$

as recovered from Eq.(2.1) with $m = 1$.

High-Order Finite Difference Scheme: A fourth order centered finite difference scheme on the form

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \frac{1}{12\Delta x}\left(u_{j-2} - 8u_{j-1} + 8u_{j+1} - u_{j+2}\right) \quad .$$

This scheme appears from Eq.(2.1) with $m = 2$ and evaluating the derivative of the interpolation polynomial at the grid point, x_j.

Global Scheme: The final scheme appears, as we shall see shortly, by taking $m = \infty$ in Eq.(2.1). In this case, the approximation of the derivative at the grid points is evaluated by a matrix-vector product as

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \sum_{i=0}^{N} \tilde{\mathrm{D}}_{ji} u_i \quad ,$$

where the entries of the matrix operator is

$$\tilde{\mathrm{D}}_{ji} = \begin{cases} \frac{(-1)^{j+i}}{2} \left[\sin\left(\frac{(j-i)\pi}{N+1}\right)\right]^{-1} & i \neq j \\ 0 & i = j \end{cases} \quad . \tag{2.4}$$

Let us first consider the dependence of the maximum pointwise error, L^∞, on the number of grid points, N. In Fig. 2.1 we plot this error measured at $t = \pi$ for an increasing number of grid points. It is clear that increasing the order of the method used for approximating the spatial derivative has a significant effect on the error. Indeed, the error obtained with $N = 2048$ using the 2nd order finite difference scheme is the same as that computed using the 4th order method with $N = 128$ and the global infinite order method with only $N = 12$. Moreover, by lowering Δt for the latter method one can obtain even more accurate results, i.e., the error in the global scheme is dominated by time-stepping errors rather than the errors in the spatial derivative. To be fair, one should keep in mind that the 4th order accurate method as well as the global method require more work per grid point to compute the spatial derivative. We shall return to a quantitative discussion of these aspects shortly.

Let us now restrict the attention to a comparison between the popular local 2nd order finite difference scheme and the global method. In Fig. 2.2 we show the result obtained after long time integration. Again, we clearly observe that the global scheme is superior in performance to the local scheme, even though the latter scheme employs 20 times as many grid points and is significantly slower.

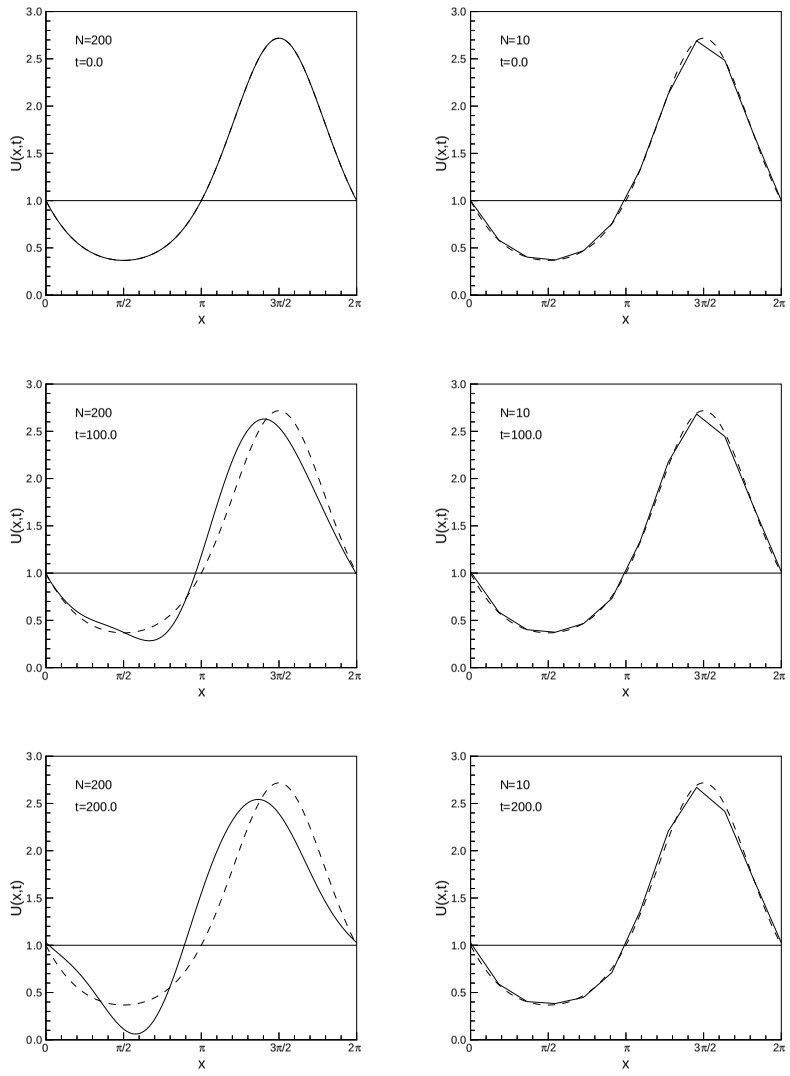

figure 2.2. An illustration of the impact of using a global method for problems requiring long time integration. On the left we show the solution of Eq.(2.3) as computed using a 2nd order centered finite difference scheme. On the right we show the same problem solved using a global method. The full lines represent the computed solution, while the dashed line represents the exact solution.

2.1 Analysis of Finite Difference Schemes

The previous example illustrates that global methods appear to be superior to local methods not only when very high spatial resolution is needed but also when long time integration is required. While the former property can be attributed to the use of additional information in the approximation of the derivative, the latter observation is perhaps more surprising and illustrates why we need to look deeper into the theory of global methods to appreciate their full potential.

In this section we utilize phase error analysis, introduced in [62], to explain the observations made in the previous section. As we shall find, the analysis confirms that high-order/global methods is the appropriate, and quite possibly the only, choice when very accurate solutions and/or long time integration is required.

2.1.1 Phase Error Analysis.

To analyze the phase error associated with a spatial approximation scheme, let us again consider the linear wave problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= -c\frac{\partial u}{\partial x} \ , \\ u(0,t) &= u(2\pi,t) \ , \\ u(x,0) &= \exp(ikx) \ , \end{aligned} \tag{2.5}$$

where $i = \sqrt{-1}$, $k = 2\pi/\lambda$ is the wavenumber with λ representing the wavelength. The solution to Eq.(2.5) is a rightward traveling wave

$$u(x,t) = \exp(ik(x-ct)) \ . \tag{2.6}$$

Note that c has the dimension of velocity and is known as the phase velocity of the wave.

We continue as in Ex. 1 and introduce an equidistant grid

$$x_j = \frac{2\pi j}{N+1} = j\Delta x \ , \ \ j \in [0,\ldots,N] \ ,$$

with the associated grid vector, $\boldsymbol{u} = (u_0,\ldots,u_N)^T$, where $u_j = u(x_j)$.

If we define the central finite difference operator

$$\mathcal{D}_n u(x_j) = \frac{u(x_j+n\Delta x) - u(x_j-n\Delta x)}{2n\Delta x} = \frac{u_{j+n}-u_{j-n}}{2n\Delta x} \ ,$$

then a general $2m$'th order approximation to the spatial derivative of $u(x,t)$ at x_j takes the form

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \sum_{n=1}^{m} \alpha_n^m \mathcal{D}_n u_j \quad , \tag{2.7}$$

where the weights, α_n^m, are given as

$$\alpha_n^m = -2(-1)^n \frac{(m!)^2}{(m-n)!(m+n)!} \quad . \tag{2.8}$$

Suppose now that we consider the semi-discrete version of Eq.(2.5), i.e., keeping time as a continuous variable and using a $2m$'th order approximation to discretize the spatial dimension, we recover a scheme for updating the grid vector, u_j, as

$$\begin{aligned} \frac{du_j}{dt} &= -c \sum_{n=1}^{m} \alpha_n^m \mathcal{D}_n u_j \quad , \\ u_j(0) &= \exp(ikx_j) \quad . \end{aligned} \tag{2.9}$$

We can interpret the grid vector, $\boldsymbol{u}$, as a vector of grid point values of an interpolating trigonometric polynomial, $v(x,t)$, i.e.,

$$v(x,t) = \sum_{|n| \le N/2} \tilde{v}_n(t) \exp(ikx) \quad ,$$

where $\tilde{v}_n(t)$ are constrained such that $v(x_j,t) = u_j(t)$. Thus, solving Eq.(2.9) amounts to

$$\begin{aligned} \frac{\partial v}{\partial t} &= -c \sum_{n=1}^{m} \alpha_n^m \mathcal{D}_n v(x,t) \quad , \\ v(x,0) &= \exp(ikx) \quad . \end{aligned} \tag{2.10}$$

If $v(x,t)$, which is a continuous function, satisfies Eq.(2.10), the solution to Eq.(2.9) is given by $v(x_j,t)$. However, the solution to Eq.(2.10) can be obtained directly on the form

$$v(x,t) = \exp(ik(x - c_m(k)t)) \quad . \tag{2.11}$$

We shall term $c_m(k)$ the numerical phase velocity. Note that contrary to the solution, Eq.(2.6), of original problem, Eq.(2.5), the solution, Eq.(2.11), to a discrete wave problem, Eq.(2.10) is dispersive, i.e., $c_m(k)$ is a function of the wavenumber, k, as a consequence of the introduction of the grid.

As there is no difference in the amplitude of the two solutions, $u(x,t)$ in Eq.(2.6) and $v(x,t)$ in Eq.(2.11), the different behavior illustrated in Fig. 2.2 must be due to differences in the propagation, i.e., the phase velocity.

Following [62] we measure the difference between the actual solution, $u(x,t)$, and the approximate solution, $v(x,t)$, as the leading order term of the relative error

$$\begin{aligned}\left|\frac{u(x,t)-v(x,t)}{u(x,t)}\right| &= |1-\exp(ik(c-c_m(k))t)| \\ &\simeq |k(c-c_m(k))t| = e_m(k) \quad ,\end{aligned}$$

which quite naturally is termed the phase error.

The computation of the phase error for various schemes allows us to pose, and answer, important questions related to accuracy and efficiency of the various schemes. In particular, we can identify the most efficient scheme guaranteeing a certain level of accuracy at a specific time.

For simplicity we shall mainly concern ourselves with the different schemes studied numerically in Ex. 1, although the validity of the techniques extends this simple example.

2.1.2 Finite Order Finite Difference Schemes.

Let us apply these new concepts to the analysis of the two different finite difference scheme discussed in Ex. 1. We begin by considering the 2nd order finite difference schemes for which Eq.(2.10) becomes

$$\begin{aligned}\frac{\partial v(x,t)}{\partial t} &= -c\frac{v(x+\Delta x,t)-v(x-\Delta x,t)}{2\Delta x} \\ v(x,0) &= \exp(ikx) \quad .\end{aligned}$$

Seeking a solution of the form Eq.(2.11) yields the numerical phase velocity

$$c_1(k) = c\frac{\sin(k\Delta x)}{k\Delta x} \quad .$$

If we continue by assuming that

$$k\Delta x = 2\pi\frac{\Delta x}{\lambda} \ll 1 \quad ,$$

,i.e., a highly resolved problem, a Taylor expansion yields

$$c_1(k) = c\left(1 - \frac{(k\Delta x)^2}{6} + \mathcal{O}((k\Delta x)^4)\right) \quad ,$$

confirming the 2nd order spatial accuracy of the scheme.

For the high-order 4th order scheme considered in Ex. 1, the approximation to Eq.(2.10) is

$$\frac{\partial v(x,t)}{\partial t} = -c\frac{v(x-2\Delta x,t) - 8v(x-\Delta x,t) + 8v(x+\Delta x,t) - v(x+2\Delta x,t)}{12\Delta t} \quad .$$

Seeking a solution of the form Eq.(2.11) results in a numerical phase velocity on the form

$$c_2(k) = c\frac{8\sin(k\Delta x) - \sin(2k\Delta x)}{6k\Delta x} \quad .$$

Again considering the limit of $k\Delta x \ll 1$ we recover

$$c_2(k) = c\left(1 - \frac{(k\Delta x)^4}{30} + \mathcal{O}((k\Delta x)^6)\right) \quad ,$$

illustrating the expected 4th order accuracy.

Using the numerical wave velocities, $c_1(k)$ and $c_2(k)$, for the 2nd and 4th order schemes, respectively, we have

$$\begin{aligned} e_1(k,t) &= kct\left|1 - \frac{\sin(k\Delta x)}{k\Delta x}\right| \quad , \\ e_2(k,t) &= kct\left|1 - \frac{8\sin(k\Delta x) - \sin(2k\Delta x)}{6k\Delta x}\right| \quad . \end{aligned} \tag{2.12}$$

To measure the accuracy of a particular scheme the actual number of grid points, N, is less important as the resolution of the scheme clearly depends on the solution. To reflect this, let us introduce

$$p = \frac{\lambda}{\Delta x} = \frac{2\pi}{k\Delta x} \quad .$$

It is worth realizing that it takes a minimum of two points per wavelength to uniquely specify a wave, i.e., p has a theoretical minimum of 2, while $p \gg 1$ reflects a highly resolved wave.

Let us also introduce the number, $\nu = ct/\lambda$, as the number of times the solution returns to itself under the assumption of periodicity. Introducing this notation into Eq.(2.12) yields

$$\begin{aligned} e_1(p,\nu) &= 2\pi\nu \left| 1 - \frac{\sin(2\pi p^{-1})}{2\pi p^{-1}} \right| \quad , \\ e_2(p,\nu) &= 2\pi\nu \left| 1 - \frac{8\sin(2\pi p^{-1}) - \sin(4\pi p^{-1})}{12\pi p^{-1}} \right| \quad . \end{aligned} \tag{2.13}$$

A leading order approximation to Eq.(2.13) is

$$\begin{aligned} e_1(p,\nu) &\simeq \frac{\pi\nu}{3}\left(\frac{2\pi}{p}\right)^2 \quad , \\ e_2(p,\nu) &\simeq \frac{\pi\nu}{15}\left(\frac{2\pi}{p}\right)^4 \quad . \end{aligned} \tag{2.14}$$

Hence, the phase error is directly proportional to the number of periods of time, ν, i.e., the error grows linearly in time.

To arrive at a practical measure, assume that we can accept an error, ε_p, after ν periods of evolution and denote by $p_m(\varepsilon_p, \nu)$ the number of points per wavelength required to ensure that the phase error is bounded by ε_p. From Eq.(2.14) we directly obtain such bounds on $p_m(\varepsilon_p, \nu)$ as

$$\begin{aligned} p_1(\varepsilon,\nu) &\geq 2\pi\sqrt{\frac{\nu\pi}{3\varepsilon_p}} \quad , \\ p_2(\varepsilon,\nu) &\geq 2\pi\sqrt[4]{\frac{\pi\nu}{15\varepsilon_p}} \quad , \\ p_3(\varepsilon,\nu) &\geq 2\pi\sqrt[6]{\frac{\pi\nu}{70\varepsilon_p}} \quad . \end{aligned} \tag{2.15}$$

To illustrate the general trend we have also included the result for a 6th order central finite difference scheme discussed in more detail in the

exercises.

Example 2. Let us consider the implication of these estimates for a few special cases.

$\underline{\varepsilon_p = 0.1}$: For this relatively large error one obtains

$$p_1 \geq 20\sqrt{\nu} \ , \ \ p_2 \geq 7\sqrt[4]{\nu} \ , \ \ p_3 \geq 6\sqrt[6]{\nu} \ .$$

We recall that the 4th order scheme is twice as expensive as the 2nd order scheme, so not much is gained for short time integration. However, as ν increases the 4th order scheme clearly becomes more attractive. For this low accuracy there is little reason to use the 6'th order scheme.

$\underline{\varepsilon_p = 0.01}$: Requiring this error one obtains

$$p_1 \geq 64\sqrt{\nu} \ , \ \ p_2 \geq 13\sqrt[4]{\nu} \ , \ \ p_3 \geq 8\sqrt[6]{\nu} \ ,$$

Here we expect a significant advantage of using the 4th order scheme, even for short time integration, while the advantage of the 6'th order scheme remains marginal unless very long time integration is required.

$\underline{\varepsilon_p = 10^{-5}}$: This approximately corresponds to the minimum error of the 2nd order scheme shown in Fig. 2.1. We obtain

$$p_1 \geq 643\sqrt{\nu} \ , \ \ p_2 \geq 43\sqrt[4]{\nu} \ , \ \ p_3 \geq 26\sqrt[6]{\nu}$$

which corresponds reasonably with what is observed in Fig. 2.1 and confirms that high order methods are superior when high accuracy is required, even for short time integration.

While it generally is accepted that high-order methods yield superior accuracy, one often encounters doubts regarding the efficiency of such methods. To briefly address this, let us define a measure of work, W_m, as

$$W_m = 2mp_m \frac{t}{\Delta t} = 2mp_m \frac{p_m \nu}{CFL_m} \ ,$$

where $CFL_m = c\Delta t/\Delta x$ refers to the maximum CFL number for stabil-

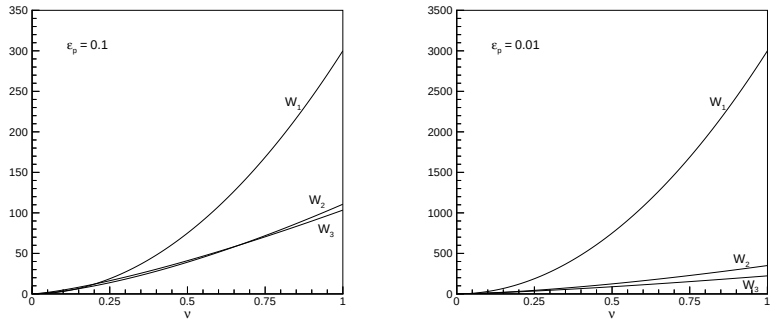

figure 2.3. The growth of the work function, W_m, for various finite difference schemes is given as a function of time, ν, in terms of periods. On the left we show the growth for a required phase error of $\varepsilon_p = 0.1$ while the right shows the result of a similar computation, however with $\varepsilon_p = 0.01$, i.e., a maximum phase error of less than 1 %.

ity, i.e., W_m represent a measure of the amount of work per wavelength using a $2m$'th order scheme during the required number of time-steps.

Assuming that a 4'th order explicit Runge-Kutta method is used for the temporal integration, hence defining the values of CFL_m, the estimated work for a 2nd, a 4th and a 6th order central finite difference scheme is given as

$$W_1 \simeq 30\nu\frac{\nu}{\varepsilon_p} \quad , \quad W_2 \simeq 35\nu\sqrt{\frac{\nu}{\varepsilon_p}} \quad , \quad W_3 \simeq 48\nu\sqrt[3]{\frac{\nu}{\varepsilon_p}} \quad . \tag{2.16}$$

In Fig. 2.3 we illustrate the approximate work associated with the different schemes as a function of accuracy and time. While we expected the high-order methods to be the most appropriate choice when considering only accuracy, this confirms that also for problems exhibiting unsteady behavior should one consider the use of high-order methods to minimize the work required to solve the problem at a prescribed accuracy.

2.1.3 Infinite Order Finite Difference Schemes.

Taking the estimates, Eq.(2.15), to the limit of of $m \to \infty$ suggests that the required number of grid points, p_∞, approaches a constant independent of ν and ε_p. This is also reflected in Eq.(2.16), which yields

$W_\infty \propto \nu$, indicating that the work depends only linearly on time. In other words, the required number of points per wavelength, p_∞, should be independent of ν as well as ε_p.

To make this argument a bit more quantitative, let us first recall that the $2m$'th accurate scheme, Eq.(2.7), can also be expressed as

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \mathcal{D}_1 \sum_{n=0}^{m-1} (-1)^n \tau_{2n} \left(\Delta x^2 \mathcal{D}_+ \mathcal{D}_-\right)^n u_j \quad ,$$

where

$$\mathcal{D}_+ u_j = \frac{u_{j+1} - u_j}{\Delta x} \quad , \quad \mathcal{D}_- u_j = \frac{u_j - u_{j-1}}{\Delta x} \quad ,$$

is the standard upwind and downwind difference operator, respectively. The constants, τ_{2m}, are given explicitly by the series [62]

$$(\arcsin x)^2 = 2x^2 \sum_{n=0}^{\infty} \frac{2^{2n} \tau_{2n}}{2n+2} x^{2n} \quad .$$

For this to hold at $x = 1$, it is clear that

$$2^{2n} \tau_{2n} = (1 + \alpha_n)^{2n} \quad ,$$

with α_n vanishing as n approaches infinity. In other words, $2^{2n}\tau_{2n}$, approaches a constant for large values of n.

Proceeding as previously, we recover the expression for the phase error

$$e_m(p, \nu) = 2\pi\nu \left[1 - \frac{\sin(2\pi p^{-1})}{2\pi p^{-1}} \sum_{n=0}^{m-1} 4^n \tau_{2n} \sin^{2n}\left(\frac{\pi}{p}\right)\right] \quad .$$

To leading order, this yields a phase error as

$$e_m(p, \nu) \simeq 2\pi\nu 4^m \tau_{2m} \sin^{2m}\left(\frac{\pi}{p}\right) \quad .$$

However, as $4^m \tau_{2m}$ is bounded for m approaching infinity, we recover that $\sin(\pi p^{-1}) < 1$ suffices to guarantee that the phase error vanishes for large m. In other words

$$\lim_{m \to \infty} p_m(\varepsilon_p, \nu) = 2 \quad ,$$

i.e., the infinite accuracy finite difference scheme achieves the minimal

number of points regardless of the accuracy and time requirements.

Despite the asymptotic nature of these arguments it is noteworthy that the results discussed in Ex. 1 conform very well with the above observations, i.e., the infinite order finite difference scheme and the global scheme discussed in that example appear to be closely connected. To further elaborate on this connection, let us rewrite the $2m$'th order approximation as

$$\begin{aligned}\left.\frac{\partial u}{\partial x}\right|_{x_j} &= \sum_{n=1}^{m} \alpha_n^m \frac{u_{j+n} - u_{j-n}}{2n\Delta x} \\ &= \frac{N+1}{4\pi}\left(\sum_{n=1}^{m} \frac{\alpha_n^m}{n} u_{j+n} + \sum_{n=-1}^{-m} \frac{\alpha_{-n}^m}{n} u_{j+n}\right) \\ &= \frac{N+1}{4\pi} \sum_{n=-m}^{m} \frac{\beta_n^m}{n} u_{j+n} \quad ,\end{aligned}$$

where we have introduced the new weight

$$\beta_n^m = \begin{cases} \alpha_n^m & n \neq 0 \\ 0 & n = 0 \end{cases} \quad ,$$

and used that $\alpha_n^m = \alpha_{-n}^m$.

Consider now the infinite order finite difference approximation, i.e., the case of $m \to \infty$. Using Eq.(2.8) this implies

$$\beta_n^\infty = \begin{cases} -2(-1)^n & n \neq 0 \\ 0 & n = 0 \end{cases} \quad ,$$

and the approximation becomes

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} = \frac{N+1}{4\pi} \sum_{n=-\infty}^{\infty} \frac{\beta_n^\infty}{n} u_{j+n} \quad .$$

The 2π-periodicity of the solution $u(x,t)$ to Eq.(2.3) is reflected in the grid function as

$$u_{j+n} = u_{j+n+p(N+1)} \quad , \quad p = 0, \pm 1, \pm 2 \ldots \quad .$$

This yields the approximation

$$\begin{aligned}\left.\frac{\partial u}{\partial x}\right|_{x_j} &= \frac{N+1}{4\pi}\sum_{n=-j}^{N-j}\left(\sum_{p=-\infty}^{\infty}\frac{\beta^{\infty}_{n+p(N+1)}}{n+p(N+1)}\right)u_{j+n} \\ &= \frac{1}{2\pi}\sum_{n=-j}^{N-j}-(-1)^n\left(\sum_{p=-\infty}^{\infty}\frac{(-1)^p}{p+n/(N+1)}\right)u_{j+n} \\ &= \frac{1}{2\pi}\sum_{n=-j}^{N-j}-(-1)^n\frac{\pi}{\sin(\pi n/(N+1))}u_{j+n} \quad ,\end{aligned}$$

where the last step assumes that $n \neq 0$. In the special case of $n = 0$ the sum over p vanished identically. Introducing the substitution $i = j + n$, we obtain

$$\begin{aligned}\left.\frac{\partial u}{\partial x}\right|_{x_j} &= \sum_{i=0}^{N}-\frac{1}{2}(-1)^{i-j}\left[\sin\left(\frac{\pi}{N+1}(i-j)\right)\right]^{-1}u_i \\ &= \sum_{i=0}^{N}\frac{1}{2}(-1)^{j+i}\left[\sin\left(\frac{\pi}{N+1}(j-i)\right)\right]^{-1}u_i \quad ,\end{aligned}$$

for $i \neq j$ while the diagonal entry vanishes.

Hence, we obtain the remarkable result [23, 24] that the infinite order finite difference approximation of the spatial derivative of a periodic function can be implemented exactly through the use of the differentiation matrix, $\tilde{\mathrm{D}}$. We recall that this was exactly the approach exploited in Ex. 1. While this certainly is an interesting observation, explaining the good agreement with the phase error analysis, it is even more remarkable that the exact same formulation can be interpreted as Fourier spectral methods as we shall discuss in the following.

2.2 The Fourier Spectral Method

Rather than starting with the finite difference formula, Eq.(2.9), and identifying $v(x,t)$ as a trigonometric polynomial, let us assume that $u(x,t)$ can be represented as

$$u(x,t) = \sum_{|n|\leq N/2}\tilde{u}_n(t)\exp(inx) \quad , \tag{2.17}$$

where the expansion coefficients, $\tilde{u}_n$, must be determined such that

$$u(x_j,t) = \sum_{|n|\leq N/2} \tilde{u}_n(t)\exp(inx_j) \quad , \tag{2.18}$$

is a solution to Eq.(2.5) at the grid points, x_j. As usual, we assume that N is even.

The first issue to address is how to obtain the expansion coefficients, $\tilde{u}_n(t)$, such that Eq.(2.18) holds. For that we shall need the following result

Lemma 1. Consider the equidistant grid given as

$$x_j = \frac{2\pi j}{N+1} \quad , \quad j \in [0,\ldots,N] \quad .$$

The complex exponential function obeys the orthogonality relation

$$\frac{1}{N+1}\sum_{j=0}^{N}\exp(ipx_j) = \begin{cases} 1 & p=(N+1)m,\ m=0,\pm1,\pm2,\ldots \\ 0 & \text{otherwise} \end{cases} \quad .$$

Proof: We rewrite the series as

$$\begin{aligned} \frac{1}{N+1}\sum_{j=0}^{N}\exp(ipx_j) &= \frac{1}{N+1}\sum_{j=0}^{N}\exp(i2\pi mj) \\ &= \frac{1}{N+1}\sum_{j=0}^{N}[\exp(i2\pi m)]^j \quad . \end{aligned} \tag{2.19}$$

If m is an integer, we immediately recover that $\exp(i2\pi m)=1$ and hence the first part of the result.

For m being a non-integer, we have

$$\frac{1}{N+1}\sum_{j=0}^{N}[\exp(i2\pi m)]^j = \frac{1}{N+1}\frac{1-r^{N+1}}{1-r} \quad ,$$

where $r=\exp(i2\pi m)\neq 1$. We then utilize the geometric series as

$$\frac{1-r^{N+1}}{1-r} = \frac{1-(\exp(i2\pi m))^{N+1}}{1-\exp(i2\pi m)} = \frac{1-(\exp(i2\pi(N+1)))^m}{1-\exp(i2\pi m)} = 0 \quad ,$$

since N is an integer. QED

Using Lemma 1, we immediately obtain the expansion coefficients through a discrete inner product as

$$\frac{1}{N+1}\sum_{j=0}^{N} u(x_j,t)\exp(-inx_j) = \frac{1}{N+1}\sum_{j=0}^{N}\sum_{|k|\le N/2}\tilde{u}_k(t)\exp(i(k-n)x_j)$$
$$= \sum_{|k|\le N/2}\tilde{u}_k(t)\left[\frac{1}{N+1}\sum_{j=0}^{N}\exp(i(k-n)x_j)\right]$$
$$= \tilde{u}_n(t) \quad . \tag{2.20}$$

To realize that Eq.(2.20) enforces Eq.(2.18), simply substitute the former into the latter to obtain

$$u(x,t) = \sum_{|n|\le N/2}\tilde{u}_n(t)\exp(inx)$$
$$= \sum_{j=0}^{N} u(x_j,t)\left[\frac{1}{N+1}\sum_{|n|\le N/2}\exp(in(x-x_j))\right] = \sum_{j=0}^{N} u(x_j,t)h_j(x) \quad ,$$

where

$$h_j(x) = \frac{1}{N+1}\frac{\sin\left[\frac{1}{2}(N+1)(x-x_j)\right]}{\sin\frac{1}{2}(x-x_j)} \quad ,$$

is obtained by summing the series directly. It is easily shown that $h_j(x)$ indeed is the interpolation polynomial with $h_j(x_k) = \delta_{jk}$, hence yielding Eq.(2.18).

Let us now return to the problem of computing the solution to Eq.(2.5) using the trigonometric polynomials. The approximation to the derivative is obtained directly from Eq.(2.18) as

$$\frac{\partial u}{\partial x} \simeq \sum_{|n|\le N/2} in\tilde{u}_n\exp(inx) \quad . \tag{2.21}$$

Assuming that the solution takes the form of a trigonometric polynomial, Eq.(2.18), we have

$$\sum_{|n|\le N/2}\left(\frac{d\tilde{u}_n}{dt} + inc\tilde{u}_n\right)\exp(inx) = 0 \quad ,$$

by inserting Eq.(2.18) into Eq.(2.5). This yields $N+1$ equations for the $N+1$ expansion coefficients, $\tilde{u}_n(t)$, as

$$\tilde{u}_n(t) = \exp(-inct)\tilde{u}_n(0) \ ,$$

with the initial conditions being

$$\tilde{u}_n(0) = \frac{1}{N+1}\sum_{j=0}^{N} \exp(ikx_j)\exp(-inx_j) = \begin{cases} 1 & k-n = p(N+1) \\ 0 & \text{Otherwise} \end{cases} \ .$$

Thus, for $|k| > N/2$, the initial condition is completely misrepresented, a phenomenon known as aliasing, and we get an order one error. However, if $|k| \leq N/2$, we have

$$\tilde{u}_n(0) = \begin{cases} 1 & k = n \\ 0 & k \neq n \end{cases} \ ,$$

such that the solution to Eq.(2.5) is

$$\begin{aligned} u(x,t) &= \sum_{|n|\leq N/2} \tilde{u}_n(t)\exp(inx) \\ &= \sum_{|n|\leq N/2} \tilde{u}_n(0)\exp(-inct)\exp(inx) \\ &= \exp(ik(x-ct)) \ . \end{aligned}$$

Consequently, we obtain the rather unusual, and perhaps surprising, result that the error is either of order one or the solution is exact.

If we relate the performance of the global trigonometric method to the number of points per wavelength, p, as done previously, we recover the exact solution provided $|k| \leq N/2$. In other words we have

$$2 \leq \frac{N}{k} = \frac{N}{2\pi/\lambda} < \frac{\lambda}{\Delta x} = p_{N/2} \ .$$

This implies that $p_{N/2} \to 2$ as N approaches infinity, i.e., the Fourier spectral method recovers the optimal minimal value for p_m for large N as was the case for the infinite order finite difference scheme considered in Ex. 1. The natural question that arises is whether the two schemes are indeed related or rather two separate roads to schemes both requiring only two points per wavelength.

To investigate this issue further, let us express the temporal derivative of $u(x,t)$ at x_j as

$$\frac{du_j}{dt} = \sum_{l=0}^{N} u_l(t) \left.\frac{dh_l}{dx}\right|_{x_j} = \sum_{l=0}^{N} \tilde{\mathrm{D}}_{jl} u_l(t) \quad ,$$

where the entries of the matrix $\tilde{\mathrm{D}}$ are

$$\tilde{\mathrm{D}}_{jl} = \begin{cases} \frac{(-1)^{j+l}}{2} \left[\sin\left(\frac{\pi}{N+1}(j-l)\right)\right]^{-1} & i \neq j \\ 0 & i = j \end{cases} \quad .$$

Hence, the Fourier spectral method is mathematically equivalent to the infinite order finite difference scheme [23, 24], although they were derived by very different means.

Furthermore, one obtains the exact same result by summing

$$\tilde{\mathrm{D}}_{jl} = \frac{1}{N+1} \sum_{|n| \leq N/2} in \exp\left(i \frac{2\pi n}{N+1}(j-l)\right) \quad ,$$

which appears directly by inserting the expression for $\tilde{u}_n(t)$, Eq.(2.20), into the Fourier approximation to the spatial derivative, Eq.(2.21).

This duality between point-space formulation, termed the nodal form, and corresponding coefficient-space method, referred to as the modal form, shall prove very fruitful in the subsequent chapters as it allows for different formulations and means of analysis of otherwise equivalent schemes. Furthermore, it has dramatic implications for the implementation of such methods.

2.2.1 Success and Failure.

The remarkable resolution power of the Fourier spectral method discussed above is achievable only for certain special cases as we shall illustrate through a couple of examples prior to engaging in a more thorough discussion.

Example 3. Consider the heat equation with constant coefficients

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} \quad , \\ u(0,t) &= u(\pi,t) = 0 \quad , \\ u(x,0) &= f(x) \quad . \end{aligned} \tag{2.22}$$

Since we have homogeneous boundary conditions, this problem can be solved using standard Fourier techniques to recover

$$u(x,t) = \sum_{n=1}^{\infty} \exp(-n^2 t)\hat{f}_n \sin(nx) \quad , \tag{2.23}$$

where

$$f(x) = \sum_{n=1}^{\infty} \hat{f}_n \sin(nx) \quad .$$

To solve this problem approximately, it seems reasonable to seek a solution of the form

$$u_N(x,t) = \sum_{n=1}^{N} \hat{u}_n(t) \sin(nx) \quad .$$

Indeed, by following the exact same approach as for the infinite series, we obtain the approximate solution

$$u_N(x,t) = \sum_{n=1}^{N} \exp(-n^2 t)\hat{f}_n \sin(nx) \quad .$$

We observe that the numerical approximation reproduces the first N terms of the expansion exactly as was the case when solving the wave equation. We recover the L^2-error as

$$\left(\frac{2}{\pi}\int_0^{\pi} |u(x,t) - u_N(x,t)|^2 \, dx\right)^{1/2} = \left(\sum_{n=N+1}^{\infty} \hat{f}_n^2 \exp(-2n^2 t)\right)^{1/2} \quad .$$

The dominating error term yields

$$\exp(-(N+1)^2 t)\sqrt{\sum_{n=N+1}^{\infty} \hat{f}_n^2} \quad ,$$

which, even for slowly decaying $\hat{f}_n$, decays exponentially in time.

As for the wave equation we found that the solution to the heat equation, subject to homogeneous boundary conditions, can be approximated very well using expansions based on trigonometric polynomials.

However, changing the problem slightly may result in a very different behavior.

Example 4. Consider again the heat equation, although slightly altered compared to the previous example, as

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + 1 \ , \\ u(0,t) &= u(\pi,t) = 0 \ , \\ u(x,0) &= f(x) \ . \end{aligned} \tag{2.24}$$

As before we look for a solution of the form

$$u_N(x,t) = \sum_{n=1}^{N} \hat{u}_n(t) \sin(nx) \ .$$

The equation for the expansion coefficients, $\hat{u}_n(t)$, is now given as

$$\frac{d\hat{u}_n}{dt} = -n^2 \hat{u}_n + \hat{a}_n \ .$$

Here $\hat{a}_n$ represents the expansion coefficients of the constant function, 1, in terms of $\sin(nx)$ and are given as

$$\hat{a}_n = \frac{2(1-(-1)^n)}{\pi n} \ .$$

Now solving for $\hat{u}_n(t)$ yields the result

$$\hat{u}_n(t) = \hat{f}_n \exp(-n^2 t) + \frac{2(1-(-1)^n)}{\pi n^3} \left(1 - \exp(-n^2 t)\right) \ .$$

One easily shows that the L_2 error in this case is

$$\left(\frac{2}{\pi} \int_0^{\pi} \left| u(x,t) - u_N(x,t) \right|^2 dx \right)^{1/2} \leq C \frac{1}{N^{5/2}} \ ,$$

i.e., the scheme is only slightly better than a 2nd order finite difference scheme.

The source of this result is the constant function, 1, which does not have a rapidly converging expansion in the function $\sin(nx)$, thus destroying the rapid global convergence.

As we have seen, global spectral methods, when properly constructed, has remarkable numerical properties, unsurpassed by any other scheme for solving general partial differential equations. However, the proper construction of the schemes is nontrivial and greatly affect the overall performance of the scheme. It is the study and understanding of the criteria underlying these choices that we shall devote ourselves to in the following chapters.

Exercises

1. Consider the central finite difference approximation to the spatial derivative of $u(x)$ at the grid point x_j on the form

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \sum_{n=1}^{m} \alpha_n^m \mathcal{D}_n u_j \quad ,$$

where

$$\mathcal{D}_n u(x_j) = \frac{u(x_j + n\Delta x) - u(x_j - n\Delta x)}{2n\Delta x} = \frac{u_{j+n} - u_{j-n}}{2n\Delta x} \quad ,$$

and Δx is the grid size of the equidistant grid.
Prove that for the approximation to be of order $2m$, the weights, α_n^m, takes the form

$$\alpha_n^m = -2(-1)^n \frac{(m!)^2}{(m-n)!(m+n)!} \quad .$$

2. Show that the 6th order accurate central finite difference approximation is given as

$$\left.\frac{du}{dx}\right|_{x_j} = \frac{-u_{j-3} + 9u_{j-2} - 45u_{j-1} + 45u_{j+1} - 9u_{j+2} + u_{j+3}}{60\Delta x} \quad .$$

3. (Continued). Considering the 6th order approximation, show that the numerical wave speed is given as

$$c_3(k) = c\frac{45\sin(k\Delta x) - 9\sin(2k\Delta x) + \sin(3k\Delta)}{30k\Delta} \quad ,$$

and that the leading order phase error is given as

$$e_3(p,\nu) = \frac{\pi\nu}{70}\left(\frac{2\pi}{p}\right)^6 \quad .$$

Based on this this, show that

$$p_3(\varepsilon_p,\nu) \geq 2\pi\sqrt[6]{\frac{\pi\nu}{70\varepsilon_p}} \quad ,$$

and compute the number of points required to ensure $\varepsilon_p = 0.1$ and $\varepsilon_p = 0.01$.
Compare with Ex. 2. When is it advantageous to use a 6th order scheme.

4. Using von Neumann analysis, show that

$$\Delta t \leq C_{\mathrm{RK}}\Delta x \quad , \quad \Delta t \leq C_{\mathrm{RK}}\Delta x \left[\left(1 + \sqrt{\frac{1}{6}}\right)\sin\left(\arccos\left(1 - \sqrt{\frac{3}{2}}\right)\right)\right]^{-1} \quad ,$$

yields necessary and sufficient conditions for stability of the 2nd and 4th order central finite difference approximation.
Here $C_{\rm RK}$ depends on the choice of Runge-Kutta method only.

5. (Continued). For a 4th order Runge-Kutta scheme, $C_{\rm RK} = \sqrt{8}$. Use this to derive the scalings in Eq.(2.8).

6. Show that the $2m$'th order difference formula, Eq.(2.7), can also be written on the from

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} \simeq \mathcal{D}_1 \sum_{n=0}^{m-1} (-1)^n \tau_{2n} \left(\Delta x^2 \mathcal{D}_+ \mathcal{D}_-\right)^n u_j \quad ,$$

where

$$\mathcal{D}_+ u_j = \frac{u_{j+1} - u_j}{\Delta x} \quad , \quad \mathcal{D}_- u_j = \frac{u_j - u_{j-1}}{\Delta x} \quad .$$

You do not need to relate the coefficients, α_n^m and τ_{2n}, in the two formulas.

7. (Continued) Using the limiting expression

$$\frac{\partial}{\partial x} = \mathcal{D}_1 \sum_{n=0}^{\infty} (-1)^n \tau_{2n} \left(\Delta x^2 \mathcal{D}_+ \mathcal{D}_-\right)^n \quad ,$$

show by considering the testfunction, $\exp(ikx)$, that

$$(\arcsin x)^2 = 2x^2 \sum_{n=0}^{\infty} \frac{2^{2n} \tau_{2n}}{2n+2} x^{2n} \quad .$$

8. Prove that $h_j(x)$ is

$$h_j(x) = \frac{1}{N+1} \sum_{|n| \le N/2} \exp(in(x - x_j)) = \frac{1}{N+1} \frac{\sin\left[\frac{1}{2}(N+1)(x-x_j)\right]}{\sin \frac{1}{2}(x - x_j)} \quad ,$$

and that $h_j(x_l) = \delta_{jl}$.

9. Show that the entries of the Fourier differentiation matrix, $\tilde{D}$ are given as

$$\tilde{D}_{jl} = \left.\frac{dh_l}{dx}\right|_{x_j} = \begin{cases} \frac{(-1)^{j+l}}{2} \left[\sin\left(\frac{\pi}{N+1}(j-l)\right)\right]^{-1} & i \neq j \\ 0 & i = j \end{cases} \quad .$$

10. (Continued) Prove that the entries of the Fourier differentiation matrix, $\tilde{D}$, can also be obtained by directly summing the series

$$\tilde{\mathrm{D}}_{jl} = \frac{1}{N+1} \sum_{|n| \leq N/2} in \exp\left(i \frac{2\pi n}{N+1}(j-l)\right) \quad .$$

11. Using the Fourier differentiation matrix, compute the derivative of the following functions

    (a) $f(x) = \exp(\cos(4x))$.
    (b) $f(x) = \cos(10x)$.
    (c) $f(x) = \cos(x/2)$.
    (d) $f(x) = x$.

    All functions are defined on $[0, 2\pi]$.
    Compute the pointwise error (L_∞) and the global error (L_2), for increasing values of N and discuss the different behaviors and convergence rates. Cam you explain the differences.

12. Test the accuracy of the Fourier differentiation matrix on the function

$$u(x) = \exp(k \sin x) \quad ,$$

    in the interval $x \in [0, 2\pi]$.
    Take

$$k = 2, 4, 6, 8, 10, 12 \quad ,$$

    and measure the relative pointwise error. Determine the minimum N for all values of k that ensures a maximum error less than 10^{-5}.

13. Use the Fourier differentiation matrix to approximate spatial derivatives at an equidistant grid and use this to solve the problem considered in Ex. 1. Use a 4th order Runge-Kutta scheme for the temporal integration.

    Compare the computational results with those in Ex. 1.

3

Elements of Convergence Theory

The single most important property of a numerical scheme, useful for solving partial differential equations, is that the solution approximates that of the continuous partial differential equation and that the level of accuracy improves as we refine the grid in time and space. Such behavior is known as convergence and is nothing short of the holy grail of the analysis of numerical methods for partial differential equations.

Prior to engaging in a detailed discussion of the convergence of spectral approximations to partial differential equations, we shall need to fix the mathematical framework and introduce a number of key concepts, central to the subsequent developments.

We will, with few exceptions, focus our attention on the development and analysis of schemes for solving the general initial boundary value problem (IBVP)

$$\frac{\partial u(\boldsymbol{x},t)}{\partial t} = \mathcal{F}(\boldsymbol{x},t,u(\boldsymbol{x},t)) + f(\boldsymbol{x},t) \ , \quad \boldsymbol{x} \in \mathsf{D} \ , \ t \geq 0 \ , \tag{3.1}$$

where $u(\boldsymbol{x},t) : \mathsf{D} \times \mathsf{R}_+ \rightarrow \mathsf{R}$, D is the domain of interest confined by the boundary $\delta\mathsf{D}$ and $\mathcal{F}$ represents an operator that may depend on space, $\boldsymbol{x}$, and time, t, as well as the solution, $u(\boldsymbol{x},t)$, and derivatives thereof. We have also introduced the forcing function, $f(\boldsymbol{x},t) : \mathsf{D} \times \mathsf{R}_+ \rightarrow \mathsf{R}$. The boundary conditions are given as

$$\mathcal{B}u(\boldsymbol{x},t) = h(\boldsymbol{x},t) \ , \quad \boldsymbol{x} \in \delta\mathsf{D} \ , \ t > 0 \ , \tag{3.2}$$

where $\mathcal{B}$ signifies the boundary operator and the initial conditions are specified as

$$u(\boldsymbol{x},0) = g(\boldsymbol{x}) \ , \quad \boldsymbol{x} \in \mathsf{D} \ , \ t = 0 \ . \tag{3.3}$$

In much of what follows we shall refer to the boundary conditions, the initial conditions and the force function as the data of the problem.

Attempting a direct convergence analysis of a particular scheme for the numerical approximation of the general nonlinear IBVP is at best very complicated and in most cases impossible. Thus, in the following we shall discuss and motivate alternative avenues that, if not completely resolving the question of convergence, at least can illuminate the key problems and properties of the numerical approximation to the IBVP.

3.1 Wellposedness of the Initial Boundary Value Problem

To come to an understanding of the wellposedness of the general IBVP, it is natural to first consider the equivalent Cauchy problem, i.e., we disregard the effect of the boundary conditions and assume the problem to be embedded in an infinite space. We furthermore assume that the solution, $u(\boldsymbol{x},t)$, exists and is unique and require that it depends smoothly on the data of the problem. In other words, small perturbations of the initial data implies only small perturbations on the solution.

To make these statement more rigorous, assume that $u(\boldsymbol{x},t)$ as well as $f(\boldsymbol{x},t)$ and $g(\boldsymbol{x})$ all belong to a Hilbert space, H, endowed with the norm, $\|\cdot\|_{L^2_w[\mathsf{D}]}$, for all $t \in [0,T]$. Consider a perturbed problem for $v(\boldsymbol{x},t) \in \mathsf{H}$ as

$$\frac{\partial v(\boldsymbol{x},t)}{\partial t} = \mathcal{F}(\boldsymbol{x},t,v(\boldsymbol{x},t)) + f(\boldsymbol{x},t) + \delta f(\boldsymbol{x},t) \ , \quad \boldsymbol{x} \in \mathsf{D} \ , \ t \geq 0 \ , \tag{3.4}$$

with the initial conditions

$$v(\boldsymbol{x},0) = g(\boldsymbol{x}) + \delta g(\boldsymbol{x}) \ , \quad \boldsymbol{x} \in \mathsf{D} \ , \ t = 0 \ , \tag{3.5}$$

We shall use the following definition [61]

Definition 1 (Wellposedness I). *Assume that a unique solution exists to the Cauchy problem given by Eq.(3.1) for given initial data and for all $t \in [0,T]$. Then the problem is wellposed if there exists a unique solution to the perturbed problem, Eqs.(3.4)-(3.5), for which*

$$\sup_{t\in[0,T]} \|\delta f\|_{L^2_w[\mathsf{D}]} + \|\delta g\|_{L^2_w[\mathsf{D}]} \leq \varepsilon \ ,$$

for any $\varepsilon > 0$, such that

$$\sup_{t\in[0,T]} \|u(t) - v(t)\|_{L^2_w[\mathsf{D}]} \leq C(T) \left\{ \sup_{t\in[0,T]} \|\delta f\|_{L^2_w[\mathsf{D}]} + \|\delta g\|_{L^2_w[\mathsf{D}]} \right\} \ ,$$

where the constant, $C(T)$, can depend on T but not on the initial data.

Establishing wellposedness for the general nonlinear operator is tremendously complicated and in many cases not currently possible. This is further complicated by realizing that the choice of the norm in Def. 1 plays a crucial role, e.g., a problem being wellposed in one norm may well be illposed in another norm, an example of which we shall see shortly.

To continue our discussion on wellposedness beyond this point, we shall reduce the complexity of the general problem. Clearly, wellposedness of the nonlinear problem is closely related to that of the linearized problem as stated in [61]

Linearization Principle: A nonlinear problem is wellposed at $u(\boldsymbol{x}, t)$ if the linear problems obtained by linearizing for all functions in the neighborhood of $u(\boldsymbol{x}, t)$ are wellposed.

Hence, we can relate the issue of wellposedness of a set of variable coefficient, linear problems to that of the original nonlinear problem with the former providing necessary but not sufficient conditions for wellposedness of the latter. Let us furthermore assume that the solution, $u(\boldsymbol{x}, t)$, has a minimum degree of smoothness, i.e., $u(\boldsymbol{x}, t) \in \mathsf{C}[\mathsf{D}]$. This allows us to state the

Localization Principle: If all constant coefficient problems are wellposed and the solution can be bounded solely by the initial data then the corresponding variable coefficient problem is also wellposed.

While this result, motivating the analysis of frozen coefficient problems, is invalid for the general variable coefficient problem it provides necessary but not sufficient conditions for strictly hyperbolic, parabolic and mixed type operators[61], to which we shall devote most of our attention.

Motivated by the above line of arguments, although not very rigorous in nature, we shall mainly focus on constant or variable coefficient problems as they remain the only type of problems for which a somewhat general theory can be developed. While certainly not rigorous, the two above principles suggests that results do carry over from the linear case to the fully non-linear case or rather that the analysis of the former

may shed some light on the properties of the latter, and much harder, problem.

Let us therefore introduce the linear, constant coefficient form of the original IBVP, Eqs.(3.1)-(3.3), as

$$\begin{aligned} \frac{\partial u(\boldsymbol{x},t)}{\partial t} &= \mathcal{L}u(\boldsymbol{x},t) + f(\boldsymbol{x},t) \ , \quad && \boldsymbol{x} \in \mathsf{D} \ , \ t \geq 0 \ , \\ \mathcal{B}u(\boldsymbol{x},t) &= h(\boldsymbol{x},t) \ , && \boldsymbol{x} \in \delta\mathsf{D} \ , \ t > 0 \ , \\ u(\boldsymbol{x},0) &= g(\boldsymbol{x}) \ , && \boldsymbol{x} \in \mathsf{D} \ , \ t = 0 \ , \end{aligned} \tag{3.6}$$

where $\mathcal{L}$ and $\mathcal{B}$ are independent of time and space and the data are assume to be in $\mathsf{C}[\mathsf{D}]$ at all times.

The complexity of this general problem can be reduced considerably without sacrificing the validity of the subsequent analysis. Let us first consider the effect of the inhomogeneous boundary conditions, $h(\boldsymbol{x},t)$, and introduce the transformation

$$v(\boldsymbol{x},t) = u(\boldsymbol{x},t) - \phi(\boldsymbol{x},t)h(\boldsymbol{x},t) \ ,$$

where $\phi(\boldsymbol{x},t)$ is chosen such that $v(\boldsymbol{x},t)$ vanishes at the boundary at all times. This yields the transformed problem

$$\frac{\partial v(\boldsymbol{x},t)}{\partial t} = \mathcal{L}v(\boldsymbol{x},t) + \left[f(\boldsymbol{x},t) + h(\boldsymbol{x},t)\mathcal{L}\phi(\boldsymbol{x},t) - \frac{\partial \phi(\boldsymbol{x},t)h(\boldsymbol{x},t)}{\partial t} \right] \ .$$

Provided only that $\phi(\boldsymbol{x},t)$ and $h(\boldsymbol{x},t)$ are functions of bounded variation in $t \in [0,T]$, wellposedness of the problem subject to general boundary conditions follows directly from wellposedness of a homogeneous boundary value problem, subject to a different forcing function.

Through a similar line of arguments and the use of the transformation

$$v(\boldsymbol{x},t) = u(\boldsymbol{x},t) - e^{-t}g(\boldsymbol{x}) \ ,$$

we have

$$\frac{\partial v(\boldsymbol{x},t)}{\partial t} = \mathcal{L}v(\boldsymbol{x},t) + \left[f(\boldsymbol{x},t) + e^{-t}(g(\boldsymbol{x}) + \mathcal{L}g(\boldsymbol{x})) \right] \ ,$$

subject to homogeneous initial conditions. As for the boundary conditions, the issue of wellposedness follows directly from the wellposedness of a very similar problem, albeit subject to a different forcing.

However, the impact of the general forcing, $f(\boldsymbol{x},t)$, on the wellposed-

ness of the problem can be understood by recalling that if the solution to Eq.(3.6) with $f(\boldsymbol{x},t)=0$ takes the form

$$u(\boldsymbol{x},t) = \exp\left(\mathcal{L}t\right) g(\boldsymbol{x}) \quad ,$$

then the solution solution to the inhomogeneous problems is given as

$$u(\boldsymbol{x},t) = \exp\left(\mathcal{L}t\right) g(\boldsymbol{x}) + \int_0^t \exp\left(\mathcal{L}\tau\right) f(\boldsymbol{x},\tau)\, d\tau \quad .$$

This results is known as Duhamel's principle and remains valid also for the general variable coefficient and nonlinear problems [**?**].

Hence, we can recast the general linear IBVP into one with homogeneous initial and boundary values, and neglect the forcing terms in the analysis of wellposedness as it follows from that of the homogeneous problem. It should be cautioned, however, that if the homogeneous problem is illposed, this approach does not in general allow us to state anything about the inhomogeneous problem.

Through this rather long line of arguments we have realized that quite a lot can be said about the solution to general nonlinear IBVP by studying the much simpler linear homogeneous IBVP on the form

$$\begin{aligned} \frac{\partial u(\boldsymbol{x},t)}{\partial t} &= \mathcal{L}u(\boldsymbol{x},t) \quad , \quad && \boldsymbol{x} \in \mathsf{D} \quad , \; t \geq 0 \quad , \\ \mathcal{B}u(\boldsymbol{x},t) &= 0 \quad , && \boldsymbol{x} \in \delta\mathsf{D} \quad , \; t > 0 \quad , \\ u(\boldsymbol{x},0) &= g(\boldsymbol{x}) \quad , && \boldsymbol{x} \in \mathsf{D} \quad , \; t = 0 \quad , \end{aligned} \tag{3.7}$$

which shall be the main subject of our study. In much of what follows we shall therefore consider the linear IBVP for which wellposedness is defined as

Definition 2 (Wellposedness II). *Assume that a solution exists to the problem, Eq.(3.7). Then the problem is wellposed for $t \in [0,T]$ in $L^2_w[\mathsf{D}]$ provided only that*

$$\sup_{t\in[0,T]} \|u(t)\|_{L^2_w[\mathsf{D}]} \leq C(T) \|g\|_{L^2_w[\mathsf{D}]} \quad ,$$

where the constant, $C(T)$, can depend on T but not on the initial data.

3.2 Consistency, Stability, and Convergence

We are now ready to return to the discussion of convergence of a numerical scheme, serving as an approximation of a general initial boundary value problem. Based on the discussion in the last section it seems reasonable to focus the treatment to linear, variable coefficient problems. One should be aware that such an analysis provides only necessary, but not sufficient, conditions for convergence of the approximation to the general nonlinear IBVP.

Let us for simplicity begin by restricting the attention to the one-dimensional linear, constant coefficient initial scalar boundary value problem

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= \mathcal{L}u(x,t) \ , \quad && x \in \mathsf{D} \ , \ t \geq 0 \ , \\ \mathcal{B}u(x,t) &= 0 \ , && x \in \delta\mathsf{D} \ , \ t > 0 \ , \\ u(x,0) &= g(x) \ , && x \in \mathsf{D} \ , \ t = 0 \ , \end{aligned} \tag{3.8}$$

where $\mathcal{L}$ is independent of time as well as space. We shall also subsequently assume that the boundary operator, $\mathcal{B}$, is included in the operator, $\mathcal{L}$. The extension to the multi-dimensional scalar case is straightforward provided the domain of interest is simple, e.g., convex with a Lipschitz boundary. The generalization to systems of equations is considerably more complex and we refer to [?] for an detailed discussion of these complications. In Chap. 8 we shall revisit this within the context of spectral methods for conservation laws.

Let us here assume that the solution, $u(x,t)$, belongs to a Hilbert space, H, endowed with a norm, $\|\cdot\|_{L^2_w[\mathsf{D}]}$, in which the problem is wellposed according to Def. 2. The boundary operator, $\mathcal{B}$, restricts the allowable solution space to $\mathsf{B} \subset \mathsf{H}$, where the Hilbert subspace, $\mathsf{B} \subset \mathsf{H}$, is constructed from all $u(x,t) \in \mathsf{H}$ for which $\mathcal{B}u(x,t) = 0$ on $\delta\mathsf{D}$. Wellposedness implies that the operator, $\mathcal{L}$, is a bounded operator from H into B, i.e., $\mathcal{L}[\mathsf{D}] : \mathsf{H} \rightarrow \mathsf{B}$.

The formulation of any numerical schemes for the solution of partial differential equations involves two essential steps

- Choosing a finite dimensional space, B_N, to approximate the continuous space, B. B_N is the space in which to seek approximate solutions and this choice defines the method, e.g., finite difference, finite volume, spectral etc.

- Defining a projection operator, $\mathcal{P}_N[\mathsf{D}] : \mathsf{H} \to \mathsf{B}_N$. This choice specifies the way in which the equation is satisfied, e.g., Galerkin, collocation etc.

Here N is a measure of the dimension of the dense subspace, $\mathsf{B}_N \in \mathsf{B}$, and of the projection operator, $\mathcal{P}_N$.

The projection is often defined through the method of weighted residuals (MWR), by enforcing that the numerical solution, $u_N(x,t) \in \mathsf{B}_N$, satisfies

$$\begin{aligned} &\frac{\partial u_N}{\partial t} - \mathcal{L}_N u_N = 0 \ , \\ &u_N(0) - g_N = 0 \ , \end{aligned} \tag{3.9}$$

where we have introduced the approximated operator, $\mathcal{L}_N[\mathsf{D}] : \mathsf{B} \to \mathsf{B}_N$, defined as $\mathcal{L}_N = \mathcal{P}_N \mathcal{L} \mathcal{P}_N$, and $g_N = \mathcal{P}_N g$.

The aim of the convergence analysis is to derive conditions for the convergence of u_N to u as N tends to infinity for any $t \in [0,T]$. However, a direct comparison between u_N and u is difficult as they occupy different spaces, leaving ambiance as to how to measure the difference. It is more natural to compare u_N and the projection, $\mathcal{P}_N u$, as they belong to the same space, B_N, endowed with the norm $\|\cdot\|_{L^2_w[\mathsf{D}]}$. It is important to realize that any numerical scheme produces a projection of the solution, $\mathcal{P}_N u$, rather than the solution, u, itself which is generally not available.

In what remains, we shall simply assume that

$$\forall t \in [0,T] \ : \ \|u(t) - \mathcal{P}_N u(t)\|_{L^2_w[\mathsf{D}]} \to 0 \quad \text{as} \quad N \to \infty \ . \tag{3.10}$$

Estimating this generally involves knowledge about the regularity of the solutions to the partial differential equation. While this topic is of great importance in the theory of partial differential equations, it is also well beyond the scope of this text. We shall simply assume that the solution has sufficient smoothness and remains bounded to ensure that Eq.(3.10) holds.

The convergence rate, however, of this may well be different from that of

$$\forall t \in [0,T] \ : \ \|u_N(t) - \mathcal{P}_N u(t)\|_{L^2_w[\mathsf{D}]} \to 0 \quad \text{as} \quad N \to \infty \ ,$$

which measures the difference between the projection of the exact solution and the numerical solution.

The projection of Eq.(3.8) yields

$$\frac{\partial \mathcal{P}_N u}{\partial t} = \mathcal{P}_N \mathcal{L} u \quad . \tag{3.11}$$

Let us recall the identity

$$\forall u_N \in \mathsf{B}_N \; : \; \mathcal{P}_N u_N = u_N \quad ,$$

i.e., the projection of a function in B_N is the identity operation. Combining Eq.(3.9) and Eq.(3.11) yields the error equation

$$\frac{\partial}{\partial t} \left(\mathcal{P}_N u - u_N \right) = \mathcal{L}_N \left(\mathcal{P}_N u - u_N \right) + \mathcal{P}_N \mathcal{L} \left(\mathcal{I} - \mathcal{P}_N \right) u \quad . \tag{3.12}$$

Hence, if $\mathcal{P}_N u(0) - u_N(0) = 0$ and the truncation error,

$$\mathcal{P}_N \mathcal{L} \left(\mathcal{I} - \mathcal{P}_N \right) u \quad , \tag{3.13}$$

vanishes, we recover that the error, $\mathcal{P}_N u - u_N$, is zero for all $t \in [0, T]$.

Let us now define the concept of convergence as

Definition 3 (Convergence). *An approximation is convergent if*

$$\forall t \in [0, T] : \| \mathcal{P}_N u(t) - u_N(t) \|_{L^2_w[\mathsf{D}]} \to 0 \quad as \quad N \to \infty \quad ,$$

for all $u(0) \in \mathsf{B}$ *and* $u_N(0) \in \mathsf{B}_N$.

A direct approach to proving convergence of a specific scheme is, in general, hard. However, there fortunately is an alternative avenue along which to proceed.

We recall Eq.(3.12) and define

Definition 4 (Consistency). *An approximation is consistent if*

$$\begin{array}{l} \| \mathcal{P}_N \mathcal{L} \left(\mathcal{I} - \mathcal{P}_N \right) u \|_{L^2_w[\mathsf{D}]} \to 0 \\ \| \mathcal{P}_N u(0) - u_N(0) \|_{L^2_w[\mathsf{D}]} \to 0 \end{array} \quad as \quad N \to \infty \quad ,$$

for all $u \in \mathsf{B}$ *and* $u_N(0) \in \mathsf{B}_N$.

This essentially requires that the truncation error introduced by the approximation vanishes as N approaches infinity.

Let us also define

Definition 5 (Stability). *An approximation is stable if*

$$\forall N : \| \exp(\mathcal{L}_N t)\|_{L^2_w[\mathsf{D}]} \leq C(t) \ ,$$

with the associated operator norm

$$\| \exp(\mathcal{L}_N t)\|_{L^2_w[\mathsf{D}]} = \sup_{u \in \mathsf{B}} \frac{\| \exp(\mathcal{L}_N t)u\|_{L^2_w[\mathsf{D}]}}{\|u\|_{L^2_w[\mathsf{D}]}} \ ,$$

and $C(t)$ is independent of N and bounded for any $t \in [0, T]$.

This guarantees that the solution remains bounded as N approaches infinity. Stability of the approximation is clearly closely related to the question of wellposedness for the partial differential equation.

These concepts are connected through one of the principal results in the convergence theory of the numerical approximation of linear partial differential equations.

Theorem 1 (Lax-Richtmyer Equivalence Theorem). *A consistent approximation to a linear wellposed partial differential equation is convergent if and only if it is stable.*

Proof: We will just outline the proof of this important result. Let us first establish that consistency and stability it suffices to guarantee convergence. Consider the error equation, Eq.(3.12),

$$\frac{\partial}{\partial t} \left(\mathcal{P}_N u - u_N \right) = \mathcal{L}_N \left(\mathcal{P}_N u - u_N \right) + \mathcal{P}_N \mathcal{L} \left(\mathcal{I} - \mathcal{P}_N \right) u \ .$$

Using Duhamel's principle yields

$$\begin{aligned} \mathcal{P}_N u(x,t) - u_N(x,t) = \ & \exp\left[\mathcal{L}_N t\right] \left(\mathcal{P}_N u(x,0) - u_N(x,0) \right) \\ & + \int_0^t \exp\left[\mathcal{L}_N (t-s)\right] \mathcal{P}_N \mathcal{L} \left[\mathcal{I} - \mathcal{P}_N \right] u(s) \, ds \ , \end{aligned}$$

which is valid provided only that the truncation error has sufficient smoothness. Introducing the $L^2_w[\mathsf{D}]$ norm and the triangle inequality we recover

$$\|\mathcal{P}_N u(t) - u_N(t)\|_{L^2_w[\mathsf{D}]} \leq \ C(t) \|\mathcal{P}_N u(0) - u_N(0)\|_{L^2_w[\mathsf{D}]}$$

$$+\int_0^t C(t-s)\|\mathcal{P}_N\mathcal{L}\left[\mathcal{I}-\mathcal{P}_N\right]u(s)\|_{L_w^2[\mathsf{D}]}\,ds \quad .$$

Provided u is dense in B, uniformly bounded in t, and the approximation is stable and consistent according to Def. 4 and Def. 5, the truncation error vanishes as N approaches infinity, thus establishing convergence. Indeed, we observe that the rate of convergence is the same as that of the truncation error.

Conversely, to prove that convergence implies stability, we recall that convergence implies that

$$\begin{aligned}\Big|\,\|\exp(\mathcal{L}t)u\|_{L_w^2[\mathsf{D}]} \;-\; \|\exp(\mathcal{L}_N t)u\|_{L_w^2[\mathsf{D}]}\,\Big| \\ \leq \|\exp(\mathcal{L}t)u-\exp(\mathcal{L}_N t)u\|_{L_w^2[\mathsf{D}]}\to 0 \quad ,\end{aligned}$$

for $N\to\infty$. Since $\|\exp(\mathcal{L}t)u\|_{L_w^2[\mathsf{D}]}$ is bounded due to wellposedness, see Def. 2, this indicates that convergence indeed implies stability. However, there are subtleties as $\|\exp(\mathcal{L}_N t)u\|_{L_w^2[\mathsf{D}]}$ may depend on u as well as t. We shall not address this issue further but refer to [??] for a complete proof of the equivalence theorem. QED

A few remarks regarding the use of the equivalence theorem is in place. We have indicated the proof of the theorem in Hilbert spaces, utilizing the inner product norms. The original result, on the other hand, is valid for solutions in Banach spaces, i.e., the result remains valid in all of the $L_w^p[\mathsf{D}]$ spaces. It is crucial to appreciate, however, that the consistency, stability and wellposedness of the problem has to be established in equivalent spaces, i.e., using equivalent norms, for the theorem to remain valid. The problem is the issue of wellposedness which may well be lost when changing norm. To appreciate this, consider the following example [**?**].

Example 5. Consider the linear wave equation

$$\frac{\partial u}{\partial t}=-\frac{\partial u}{\partial x} \quad , \quad x\in[-1,1] \quad ,$$

with $u(-1,t)=0$ and the initial conditions being

$$u(x,0)=\begin{cases}1-\varepsilon^{-1}|x| & |x|\leq 1-\varepsilon \\ 0 & |x|>\varepsilon\end{cases} \quad ,$$

where $\varepsilon > 0$.

Let us consider the norm

$$\|u(t)\|_\alpha^2 = \int_{-1}^{1} u^2(t)(1-x^2)^\alpha \, dx \quad ,$$

with $\alpha > -1$. Note that $\alpha = 0$ reflects the classical energy (L^2) norm.

One can easily show that

$$\|u(0)\|_\alpha^2 \propto \varepsilon \quad .$$

The solution at $t = 1$ is

$$u(x,1) = \begin{cases} 0 & x \le 1-\varepsilon \\ 1 - \varepsilon^{-1} + \varepsilon^{-1}|x| & 1-\varepsilon < x < 1 \end{cases} \quad .$$

Evaluating the norm shows

$$\|u(1)\|_\alpha^2 \propto \varepsilon^{\alpha+1} \quad .$$

For wellposedness, we must require that

$$\|u(t)\|_\alpha = \|\exp(\mathcal{L}t)u(0)\|_\alpha \le C(t)\|u(0)\|_\alpha \quad ,$$

where $C(t)$ can depend on the time but not on the initial conditions.

However, using the above results, we have

$$\|\exp(\mathcal{L}t)\|_\alpha \ge \frac{\|u(1)\|_\alpha}{\|u(0)\|_\alpha} \propto \varepsilon^{\alpha/2} \quad .$$

Hence, for $-1 \le \alpha \le 0$, one can not bound the operator by any finite constant and the problem is illposed. For $\alpha \ge 0$, the problem is wellposed.

This partly explains why the application of the theorem traditionally has been restricted to problems in Hilbert spaces, as wellposedness, including uniqueness and existence, as well as stability and consistency in most cases is harder to establish in the $L_w^p[\mathsf{D}]$-spaces.

The power of the Lax-Richtmyer equivalence theorem lies in the realization of a natural splitting of the convergence analysis of a numerical approximation scheme into the less difficult issues of consistency and stability. In what follows we shall rely heavily on this result to facilitate the analysis of spectral approximations to partial differential equations.

3.3 The Spectral Approximation

What separate spectral methods from all other methods for solving partial differential equations is the implicit assumption that the solution, $u(x,t) \in \mathsf{B}$, can be expressed as a series expansion of global and smooth polynomial trial functions, $\phi_n(x)$, defined on D as

$$u(x,t) = \sum_{n=0}^{\infty} \hat{u}_n(t)\phi_n(x) \quad , \tag{3.14}$$

with the truncated approximation

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t)\phi_n(x) \quad . \tag{3.15}$$

The conditions on $u(x,t)$ ensuring that such an expansion exists remains unknown for general $\phi_n(x)$. However, for special choices of the trial, or basis, functions one can establish necessary conditions for the existence. For now, we will simply assume that the series exists and later return to the question of existence for specific examples of $\phi_n(x)$.

The choice of the trial functions is of great importance for the development of a good method, e.g., recall Ex. 4 in Chapter 2, since they also define the subspace, B_N, in which we seek the approximate solution, u_N. Clearly, if B_N, is a poor approximation to B we can not expect u_N to be a good approximation to u.

In what remains we assume that $\phi_n(x) \in \mathsf{H}$ belongs to a polynomial family, including the trigonometric polynomials, that is complete in H and orthogonal under the associated inner product. In this setting the finite dimensional subspace, B_N, is of dimension $N+1$ and is spanned by a subset of a polynomial family that is dense in B_N.

In the particular case where

$$\mathsf{B}_N = \text{span}\{\phi_n(x)\}_{n=0}^{N} \quad ,$$

we realize

$$u - u_N \perp u_N \quad , \tag{3.16}$$

i.e., $u - u_N$ forms an orthogonal complement to the subspace, B_N.

So far we have not concerned ourselves with the question of how to recover the expansion coefficients, $\hat{u}_n(t)$, such that Eq.(3.14) remains true. There are two essentially different methods for doing so as we

shall discuss in the following.

3.3.1 The Continuous Approximation

Consider the truncated expansion

$$\mathcal{P}_N u(x,t) = \sum_{n=0}^{N} \hat{u}_n(t)\phi_n(x) \quad .$$

Since we assume that the trial functions, $\phi_n(x)$, form a complete and orthogonal system with respect to the weight, $w(x)$, we have

$$(\phi_n, \phi_m)_w = \gamma_n \delta_{nm} \quad ,$$

where $\gamma_n = (\phi_n, \phi_n)_w$, from which we recover the expansion coefficients

$$\hat{u}_n(t) = \frac{1}{\gamma_n} \int_{\mathsf{D}} u(x,t)\overline{\phi_n(x)} w(x)\, dx \quad . \tag{3.17}$$

Note that this formulation is grid free, i.e., we are working solely in a continuous framework and, motivated by this observation, we shall term $\hat{u}_n(t)$ the continuous expansion coefficients.

Completeness of the system, $\phi_n(x)$, in H is equivalent to the property that for all $u(x,t) \in \mathsf{H}$ we have

$$\|u - \mathcal{P}_N u\|_{L^2_w[\mathsf{D}]} \to 0 \quad \text{as} \quad N \to \infty \quad ,$$

i.e., $\mathcal{P}_N u$ converges to u in the mean. While this ensures consistency it expresses nothing about the quality of the approximation for finite N, i.e., the convergence rate.

To understand this, we recall Bessel's inequality, which in the case of an orthogonal basis becomes an equality, as

$$\|u\|_{L^2_w[\mathsf{D}]} = \left(\sum_{n=0}^{\infty} \gamma_n \hat{u}_n^2 \right)^{1/2} \quad .$$

Recalling that the truncation error is in the complement of the approximation, Eq.(3.16) we recover

$$\|u - \mathcal{P}_N u\|_{L^2_w[\mathsf{D}]} = \left(\sum_{n=N+1}^{\infty} \gamma_n \hat{u}_n^2 \right)^{1/2} \quad .$$

Hence, the approximation error depends solely on the decay of the expansion coefficients, which again depend on the actual orthogonal family being applied, the regularity of u, and the weight function, $w(x)$.

3.3.2 The Discrete Approximation

The computation of the continuous expansion coefficients involves, as expressed in Eq.(3.17), the evaluation of a continuous inner product. For general solutions, $u(x,t)$, this is very hard or even impossible to evaluate and certainly impractical for real problems.

To overcome this problem, let us introduce a set of distinct grid points, x_j, and search for a polynomial, $\mathcal{I}_N u(x,t) \in \mathsf{B}_N$, satisfying

$$\forall x_j : \mathcal{I}_N u(x_j,t) = \sum_{n=0}^{N} \tilde{u}_n \phi_n(x_j) \quad , \tag{3.18}$$

i.e., we require the approximation to $u(x,t)$ be an interpolation. The remaining question is how to obtain the discrete expansion coefficients, $\tilde{u}_n(t)$, such that Eq.(3.18) is satisfied.

Let us define the discrete weighted inner product

$$[u,v]_w = \sum_{j=0}^{N} u(x_j) v(x_j) w_j \quad , \tag{3.19}$$

and assume that $u(x), v(x) \in C[\mathsf{D}]$. We shall furthermore assume that

$$[u,v]_w = (u,v)_w \quad , \quad u,v \in \mathsf{B}_N \quad , \tag{3.20}$$

i.e., the grid points, x_j, and the discrete weights, w_j, are chosen such that the discrete inner product is identical to the usual continuous inner product for all functions in B_N. To thoroughly understand the implications of the assumptions expressed in Eqs.(3.19)-(3.20) we shall need to develop the theory of Gauss integration, the discussion of which we postpone to Chapter [**?**]. At this point the equality is simply a postulate although we saw an example of such a summation rule in Lemma 1.

However, armed with this assumption we recover the discrete expansion coefficients on the form

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n}[u(x), \phi_n(x)]_w = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) \phi_n(x_j) w_j \quad , \tag{3.21}$$

where w_j represents the discrete weights and we have $\tilde{\gamma}_n = [\phi_n, \phi_n]_w$.

3.3.3 A Comparison

Let us briefly compare the two different sets of expansion coefficients, $\hat{u}_n$ and $\tilde{u}_n$. While the computation of the former requires the evaluation of an integral, it introduces no grid points. Contrary to that, the evaluation of $\tilde{u}_n$ involves the definition of a grid and a quadrature to compute the expansion coefficients through a summation. This later approach is clearly better suited for a computer.

The use of a grid, however, introduces an additional source of error. To realize this, let us consider the relation between the two sets of expansion coefficients, assuming that $u(x,t)$ is at least continuous, i.e., $u(x,t) \in C[\mathsf{D}]$. Then the discrete expansion coefficients, $\tilde{u}_n$, can be expressed using the continuous expansion coefficients, $\hat{u}_n$, as

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} \left(\sum_{l=0}^{\infty} \hat{u}_l \phi_l(x_j) \right) \phi_n(x_j) w_j = \hat{u}_n + \frac{1}{\tilde{\gamma}_n} \sum_{l=N+1}^{\infty} \hat{u}_l [\phi_l, \phi_n]_w \quad .$$

The last term does not vanish as $\phi_l(x)$ is in the complement of B_N for $l \geq N$ in which case Eq.(3.20) is no longer valid. Summing over all modes we obtain

$$\begin{aligned} \mathcal{I}_N u(x) &= \sum_{n=0}^{N} \tilde{u}_n \phi_n(x) \\ &= \sum_{n=0}^{N} \hat{u}_n \phi_n(x) + \sum_{l=N+1}^{\infty} \hat{u}_l \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} [\phi_l, \phi_n]_w \phi_n(x) \\ &= \mathcal{P}_N u(x) + \mathcal{R}_N u(x) \quad , \end{aligned}$$

where the last term, $\mathcal{R}_N u(x)$, represents the difference between the two approximations. This difference, known as the static aliasing error, is a direct consequence of the introduction of a grid. Another interpretation is that it is caused by the loss of accuracy in the evaluation of the integral by a summation. This causes high frequency variations in the function to appear as low frequency variations in the approximation due to insufficient resolution. Since the aliasing error is in the complement of the continuous approximation, $\mathcal{P}_N u$, we recover the following error estimate

$$\|u - \mathcal{I}_N u\|_{L^2_w[\mathsf{D}]} = \|u - \mathcal{P}_N u\|_{L^2_w[\mathsf{D}]} + \|\mathcal{R}_N u\|_{L^2_w[\mathsf{D}]} \ .$$

Hence, to establish convergence of the discrete expansion we need to understand the impact of the aliasing error in addition to the behavior of the continuous approximation.

3.3.4 Revisiting the Discrete Approximation

It is worth while returning to the discrete approximation as it allows for an alternative, yet equivalent, formulation. To appreciate this, let us recall that the discrete expansion coefficients, $\tilde{u}_n$, are defined to ensure that the approximation is an interpolation, i.e., $\mathcal{I}_N u(x_j) = u(x_j)$ as discussed in Sec. 3.3.2. In other words, $\mathcal{I}_N u(x)$ represents an Nth order polynomial, specified at $N+1$ grid points, x_j. As this polynomial clearly is unique we may equally well express the interpolation on the form

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) L_j(x) \ ,$$

where the Lagrange interpolation polynomial, $L_j(x)$, based on the grid points, x_j, is given as

$$L_j(x) = \frac{Q(x)}{(x - x_j)Q'(x_j)} \ , \quad Q(x) = \prod_{j=0}^{N} (x - x_j) \ . \tag{3.22}$$

We recall that $L_j(x_k) = \delta_{jk}$, ensuring the interpolation property of $\mathcal{I}_N u$ which is now nothing else than a global polynomial on which we can perform any operation much in the spirit of the finite difference schemes discussed in Chapter **??**. Indeed, we may differentiate the interpolating polynomial to obtain an approximation to the spatial derivative of $u(x)$ at the grid points, x_j, as

$$\left.\frac{du}{dx}\right|_{x_j} \simeq \left.\frac{d(\mathcal{I}_N u)}{dx}\right|_{x_j} = \sum_{k=0}^{N} u(x_k) \left.\frac{dL_k(x)}{dx}\right|_{x_j} \ ,$$

with the derivative of the Lagrange polynomial at the collocation points being given as

$$\left.\frac{dL_k(x)}{dx}\right|_{x_j} = \begin{cases} \frac{Q'(x_j)}{(x_j-x_k)Q'(x_k)} & k \neq j \\ \frac{1}{2}\frac{Q''(x_k)}{Q'(x_k)} & k = j \end{cases} \quad .$$

An alternative way of expressing the interpolation polynomial is by inserting Eq.(3.21) into Eq.(3.18) to obtain

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) L_j(x) = \sum_{j=0}^{N} u(x_j) \left(w_j \sum_{n=0}^{N} \frac{1}{\gamma_n} \phi_n(x_j)\phi_n(x) \right) \quad .$$

Due to the uniqueness of the interpolation polynomial we recover

$$L_j(x) = w_j \sum_{n=0}^{N} \frac{1}{\gamma_n} \phi_n(x_j)\phi_n(x) \quad . \tag{3.23}$$

While the construction of $L_j(x)$ from Eq.(3.22) is possible for any set of $N+1$ distinct grid points, the formulation in Eq.(3.23) is clearly more restrictive in that it involves a sum over a particular orthogonal basis. Hence, one can certainly find examples of $L_j(x)$ that can not be expressed in the form of Eq.(3.23).

However, as we shall concern ourselves with approximations originating from orthogonal expansions with associated quadrature nodes and weights, both formulations are of relevance. Indeed, we shall see that the duality in expressing the discrete approximation in terms of expansion coefficients or in terms of Lagrange interpolation polynomials shall be of very significant use for the analysis of spectral methods as well as the more practical aspects.

3.4 Method of Weighted Residuals

So far we have focused the discussion on the construction of the finite dimensional subspace, B_N, and how to recover the approximations to $u(x,t)$ in the finite dimensional space. However, to complete the specification of the numerical scheme we need to discuss the equally important question of how to satisfy the partial differential equation.

Consider again the linear scalar problem

$$\frac{\partial u}{\partial t} = \mathcal{L} u \quad , \tag{3.24}$$

with appropriate initial conditions and assume that the boundary operator, $\mathcal{B}$, is included in $\mathcal{L}$ and that $u \in \mathsf{B}$.

To recover a semi-discrete approximation we use the method of weighted residuals (MWR) to require that the residual, $R_N(x,t)$, is orthogonal in an inner product to a set of test functions, $\psi_n(x)$, as

$$\forall n \in [0,N] \ : \ \ (R_N, \psi_n)_w = \int_{\mathsf{D}} R_N \, \bar{\psi}_n \, w \, dx = 0 \ \ ,$$

where the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \mathcal{L} u_N \ \ ,$$

reflects the error introduced by approximating $u(x,t)$ by $u_N(x,t)$.

The choice of the test-functions defines the way in which we satisfy the equation, i.e., the projection operator $\mathcal{P}_N$, and names the overall scheme. In the following we briefly discuss the three essentially different choices of test functions that lead to the standard formulations of spectral approximation schemes for solving partial differential equations.

3.4.1 Galerkin Approximation

In this classical approach to the construction of a semi-discrete approximation to a partial differential equation, we assume that the solution, $u(x,t)$, can be approximated by a truncated expansion as

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t) \phi_n(x) \ \ .$$

The test-function, $\psi_k(x)$, is defined by

$$(\phi_n, \psi_l)_w = \delta_{nl} \ \ ,$$

i.e., they are essentially the orthogonal basis functions subject only to a slightly different normalization as

$$\psi_n(x) = \frac{1}{\gamma_n} \phi_n(x) \ \ .$$

This specific approach is known as the Galerkin approximation. The MWR argument leads to the following set of equations

$$\left(\frac{\partial u_N}{\partial t} - \mathcal{L}u_N, \frac{\phi_n}{\gamma_n} \right)_w = 0 \quad , \quad n = 0, \ldots, N \quad ,$$

which, using the orthonormality of the trial and test functions, yields

$$\frac{d\hat{u}_n}{dt} = (\mathcal{L}u_N, \psi_n)_w \quad , \quad n = 0, \ldots, N \quad .$$

Hence, the semi-discrete approximation consists of $N+1$ coupled ordinary differential equations, which determine the temporal evolution of the expansion coefficients, $\hat{u}_n(t)$, subject to the initial conditions on the form

$$\hat{u}_n(0) = (g, \psi_n)_w \quad , \quad n = 0, \ldots, N \quad .$$

The formulation of the Galerkin approximation may be viewed in a different way. Assume that at each given time, t, the expansion coefficients, $\hat{u}_n$, are known. Then seek values of the $N+1$ independent quantities, $(\hat{u}_n)_t$, that minimize

$$\left\| \frac{\partial u_N}{\partial t} - \mathcal{L}u_N \right\|_{L^2_w[\mathsf{D}]} \quad ,$$

i.e., it is the solution that minimizes the residual in a weighted least square sense.

The main difficulty associated with a Galerkin formulation emerges when one considers boundary conditions. Since the projection leaves no degrees of freedom through which to impose the boundary conditions, these conditions have to be a part of the basis itself. Moreover, due to the implicit orthogonality of the basis and the test-functions, we need to require that the basis functions obey the boundary conditions individually, i.e., $\phi_n(x)$ all have to satisfy the boundary conditions. This essentially restricts the practical use of the Galerkin approximation to problems with simple boundary conditions, e.g. periodic or homogeneous boundary conditions.

Example 6. Consider the constant coefficient linear problem

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} + b\frac{\partial^2 u}{dx^2} \quad ,$$

$$u(0,t) = u(2\pi,t) \ ,$$
$$u(x,0) = g(x) \ ,$$

in the domain $\mathsf{D} = [0, 2\pi]$. We will discretize the problem using a Fourier-Galerkin method and assume a polynomial solution of the form

$$u_N(x,t) = \sum_{n=-N/2}^{N/2} \hat{u}_n(t) \exp(inx) \ ,$$

where the continuous expansion coefficients, $\hat{u}_n$, are found as

$$\hat{u}_n(t) = \frac{1}{2\pi} \int_0^{2\pi} u(x,t) \exp(-inx) dx \ .$$

To construct the scheme we require that the residual

$$R_N(x,t) = \sum_{n=-N/2}^{N/2} \left(\frac{d\hat{u}_n}{dt} - ina\hat{u}_n + n^2 b\hat{u}_n \right) \exp(inx) \ ,$$

is orthogonal to $\mathsf{B}_N = \text{span}\{\phi_n\}_{n=0}^N$, i.e.,

$$R_N \perp \mathsf{B}_N \Leftrightarrow \forall n \ : \ (R_N, \phi_n)_w = 0 \ .$$

In this special case, we have that $R_N(x,t) \in \mathsf{B}_N$ and we immediately recover the $N+1$ equations to be solved as

$$\forall n \in [-N/2, \ldots, N/2] : \ \frac{d\hat{u}_n}{dt} - ina\hat{u}_n + n^2 b\hat{u}_n = 0 \ ,$$

and the initial conditions as

$$\hat{u}_n(0) = \frac{1}{2\pi} \int_0^{2\pi} g(x) \exp(-inx) dx \ .$$

The simplicity of this scheme is caused by the fact that

$$\mathcal{L} = a\frac{\partial}{\partial x} + b\frac{\partial^2}{\partial x^2} \ ,$$

commutes with the projection, $\mathcal{L}\mathcal{P}_N = \mathcal{P}_N\mathcal{L}$, i.e., the truncation error, Eq.(3.13), vanishes and we recover the projection of the exact solution, $u_N = \mathcal{P}_N u$.

A main drawback of the Galerkin method is that we have to derive the system of ordinary differential equations separately for each individual problem. While this is possible for constant coefficient problems, it may prove hard or even impossible when considering more general variable coefficient or nonlinear problems. Moreover, the Galerkin approach requires that the trial functions obey the boundary conditions individually which may well complicate matters considerably.

3.4.2 Tau Approximation

For many problems, the boundary conditions are sufficiently complicated to make the Galerkin approximation impractical. Let us in the following assume that the boundary operator, $\mathcal{B}u = 0$, require us to enforce N_b boundary conditions, and assume that the solution to the partial differential equation, $u(x,t)$, is approximated as

$$u_N(x,t) = \sum_{n=0}^{N+N_b} \hat{u}_n \phi_n(x) \quad .$$

In contrast to the Galerkin approximation, we shall choose test functions, $\psi_n(x)$, that do not satisfy the boundary conditions individually. However, in applying the MWR argument we shall leave enough degrees of freedom to enforce the boundary conditions. The consequence of this procedure is that we need to specify two sets of test functions.

For the partial differential equation itself we choose the test functions as for the Galerkin approach

$$\psi_n(x) = \frac{1}{\gamma_n}\phi_n(x) \quad , \quad n = 0, \ldots, N \quad ,$$

and require orthogonality between the residual and the space of test functions as

$$\left(\frac{\partial u_N}{\partial t} - \mathcal{L}u_N, \psi_n \right)_w = 0 \quad , \quad n = 0, \ldots, N \quad .$$

This results in $N+1$ coupled ordinary differential equations as

$$\frac{d\hat{u}_n}{dt} = \left(\mathcal{L}u_N, \psi_n \right)_w \quad ,$$

to describe the evolution of the first $N+1$ expansion coefficients.

The N_b remaining coefficients shall be specified to enforce the bound-

ary conditions. For this we define a set of N_b test-functions on the form

$$\psi_n^{\mathrm{BC}}(x) = \frac{1}{\gamma_n}\phi_n(x)\delta_{\delta\mathsf{D}} \ .$$

Here $delta_{\delta\mathsf{D}}$ reflects a function that vanishes everywhere except at $\delta\mathsf{D}$ where it becomes unity.

Applying the MWR condition on the boundary equation, yields

$$(\mathcal{B}u_N, \psi_n^{\mathrm{BC}})_w = 0 \ , \quad n = 0, \ldots, N + N_b \ ,$$

which results in N_b conditions on the form

$$\sum_{n=0}^{N+N_b} \hat{u}_n(t)\ \mathcal{B}\phi_n|_{\delta\mathsf{D}} = 0 \ ,$$

to close the set of $N + 1 + N_b$ equations for $N + 1 + N_b$ unknowns. This constraint is natural as it simply reflects that one requires $\mathcal{B}u_N = 0$.

The name of the method, which was originally proposed by Lanczos [64, 25] originates in the observation that the approximate solution, u_N, is an exact solution to the modified problem

$$\frac{\partial u_N}{\partial t} = \mathcal{L}u_N + \sum_{p=1}^{\infty} \tau_p \phi_{N+p}(x) \ ,$$

Following the approach outlined above yields the same $N + 1$ equations for the expansion coefficients. However, we also obtain the equations for τ_p as

$$\tau_p = -\frac{(\mathcal{L}u_N, \phi_{N+p})_w}{\gamma_{N+p}} \ , \quad p = 1, 2, \ldots \ .$$

Consequently, calculating τ_p one obtains an error estimate which indicate how accurately the posed IBVP is being solved.

Let us finally note that in the trivial case where the trial functions satisfy the boundary conditions individually, the tau method and the Galerkin method are equivalent.

Example 7. Consider the elliptic problem

$$\frac{d^2u(x)}{dx^2} = f(x) \ ,$$

$$u(0) = u(\pi) = 0 \ ,$$

in the domain $\mathsf{D} = [0, \pi]$. We will also assume that $f(x)$ is even, i.e., $f(x) = f(-x)$, and π-periodic, i.e., $f(x) = f(x + \pi)$.

We chose to discretize the problem using a Cosine-tau method and seek a polynomial solution on the form

$$u_N(x) = \sum_{n=0}^{N+2} \hat{u}_n \cos(nx) \ .$$

The expansion coefficients, $\hat{u}_n$, can be found using the orthogonality of the cosine basis as

$$\int_0^\pi \cos(nx)\cos(lx)dx = \frac{\pi}{2}\delta_{nl} \ ,$$

such that

$$\hat{u}_n = \frac{2}{\pi}\int_0^\pi u(x)\cos(nx)dx \ .$$

The expansion of $f(x)$ is found in a similar way.

We observe that since $\cos(0) = 1$ and $\cos(n\pi) = (-1)^n$, none of the basis functions satisfy the boundary conditions.

To construct the approximation we require that the residual

$$R_N(x) = \sum_{n=0}^{N} \left(-n^2\hat{u}_n - \hat{f}_n\right)\cos(nx) \ ,$$

is orthogonal to B_N as for the Galerkin approximation. This yields the first $N + 1$ equations as

$$\forall n \in [0..N] \ : \ -n^2\hat{u}_n = \hat{f}_n \ .$$

The additional equations required to enforce the boundary conditions yield

$$u_N(0) = \sum_{n=0}^{N+2} \hat{u}_n = 0 \ ,$$
$$u_N(\pi) = \sum_{n=0}^{N+2} \hat{u}_n(-1)^n = 0 \ ,$$

providing the 2 equations required to solve for the $N+3$ unknown.

For this problem we obtain the measures of error as

$$\tau_{N+1} = -(N+1)^2 \hat{u}_{N+1} - \hat{f}_{N+1} \ , \quad \tau_{N+2} = -(N+2)^2 \hat{u}_{N+2} - \hat{f}_{N+2} \ ,$$

while $\tau_p = 0$ otherwise.

As for the Galerkin method, a drawback of the tau-method is the need to derive the equations separately for each separate problem. For linear problems, the tau-method yields a very efficient and accurate method for the solution of ordinary differential equations. However, as for the Galerkin approximation, dealing with variable coefficient or non-linear problems is in most cases very hard and often even impossible.

3.4.3 Collocation Approximation

As we have seen, the Galerkin method and the tau method have the distinct disadvantage that one needs to explicitly derive the governing equations separately for each case.

Let us therefore consider an alternative approach known as the collocation method in which we assume that the solution to the partial differential equation, $u(x,t)$, is well approximated by the interpolation polynomial

$$u_N(x,t) = \sum_{n=0}^{N} \tilde{u}_n(t)\phi_n(x) = \sum_{j=0}^{N} u(x_j,t) L_j(x) \ ,$$

where the interpolation is based on some given set of grid points, x_j, as discussed in Sec. 3.3.4.

What separates the collocation approximation from the two previous techniques to satisfy the equation is the introduction of a grid which we define by $N+1$ distinct grid points, y_j, in D. It is important to appreciate that this set of grid points, y_j, may well be different from the set of grid points, x_j, on which the interpolation polynomial is based. The reality is, however, that they very often are chosen to coincide.

We require that the partial differential equation is satisfied exactly at y_j by choosing the test functions as shifted Dirac delta functions

$$\psi_n(x) = \delta(x - y_n) \ , \quad n = 0, \ldots, N \ .$$

Applying the MWR argument yields the equations

$$R_N(y_n,t) = \left[\frac{\partial u_N}{\partial t} - \mathcal{L}u_N\right]\Big|_{y_n} = 0 \ , \quad n = 0,\ldots,N \ .$$

In case the grid points include the boundary, $\delta\mathsf{D}$, we obtain those equations through the boundary operator which has to be obeyed exactly at the boundary point(s).

One may also understand the collocation method by assuming that at each given time, t, the expansion coefficients, $\tilde{u}_n$, are known. Then we seek values of the $N+1$ independent quantities, $(\tilde{u}_n)_t$, that minimize

$$\left[\frac{\partial u_N}{\partial t} - \mathcal{L}u_N, \frac{\partial u_N}{\partial t} - \mathcal{L}u_N\right]_w \ ,$$

i.e., it is the solution that minimizes the residual in the discrete inner product in a least square sense.

Example 8. Let us consider the following variable coefficient problem

$$\begin{aligned} &\frac{\partial u}{\partial t} = \sin x \frac{\partial u}{\partial x} \ , \\ &u(0,t) = u(2\pi,t) \ , \\ &u(x,0) = g(x) \ , \end{aligned}$$

in the domain $\mathsf{D} \in [0,2\pi]$. To discretize the problem using a Fourier-Collocation method we introduce the grid

$$x_j = \frac{2\pi}{N+1} j \ , \quad j \in [0,\ldots,N] \ ,$$

on which we will base the interpolation and satisfy the equation.

We seek a polynomial solution of the form

$$u_N(x,t) = \sum_{j=0}^{N} u_N(x_j,t) h_j(x) \ ,$$

where $h_j(x_i) = \delta_{ij}$ represents the interpolation Lagrange polynomial given as

$$h_j(x) = \frac{1}{N+1} \frac{\sin\left[\frac{N+1}{2}(x-x_j)\right]}{\sin\left[\frac{x-x_j}{2}\right]} \quad .$$

as discussed in Chap. 2. We shall also approximate the spatial derivative of $u(x,t)$ as

$$\frac{\partial u}{\partial x} \simeq \mathcal{I}_N \frac{\partial u_N}{\partial x} \quad ,$$

which yields the spatial derivative at the grid points as

$$\left.\frac{\partial u_N}{\partial x}\right|_{x_j} = \sum_{l=0}^{N} \tilde{\mathrm{D}}_{jl} u_N(x_l,t) \quad ,$$

where the entries of the differentiation matrix, $\tilde{\mathrm{D}}_{jl}$, are given as

$$\left.\frac{dh_j}{dx}\right|_{x_l} = \tilde{\mathrm{D}}_{jl} = \begin{cases} \frac{(-1)^{j+l}}{2}\left[\sin\left(\frac{\pi}{N+1}(j-l)\right)\right]^{-1} & l \neq j \\ 0 & l = j \end{cases} \quad ,$$

Since

$$\mathcal{I}_N\left(sin(x)\frac{\partial u_N}{\partial x}\right)\bigg|_{x_j} = \sin(x_j)\left.\frac{\partial u_N}{\partial x}\right|_{x_j} \quad ,$$

recover the Fourier-Collocation approximation on the form

$$\left.\frac{du_N}{dt}\right|_{x_j} = \sin(x_j)\sum_{l=0}^{N} \tilde{\mathrm{D}}_{jl} u_N(x_l) \quad ,$$

at all the collocation points, x_j.

Contrary to the Galerkin and tau method, we are not required to obtain the equations governing the expansion coefficients. The collocation scheme is different and it is straightforward to deal with variable coefficient or nonlinear problems. This simply reflects that while it is easy to interpolate such terms it may well be hard to project them onto a particular orthogonal space as required to obtain the equations for the expansion coefficients. The key disadvantage of the collocation method is the need for a grid and the associated introduction of the aliasing error.

Exercises

1. Consider the following functions

   - $u(x) = \frac{3}{5-4\cos(x)}$.
   - $u(x) = \sin(x/2)$.
   - $u(x) = x$.

   with $x \in [0, 2\pi]$.
   Derive the continuous Fourier expansion coefficients, i.e.,

   $$\hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(x) \exp(-inx)\, dx \quad ,$$

   and compare them to the discrete expansion coefficients

   $$\tilde{u}_n = \frac{1}{N+1} \sum_{j=0}^{N} u(x_j) \exp(-inx_j) \quad , \quad x_j = \frac{2\pi}{N+1} \quad ,$$

   for several different values of N. The latter summation should be evaluated computationally.

   Discuss the differences and similarities and how it relates to the different functions.

2. Consider the linear constant coefficient problem

   $$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} + b\frac{\partial^2 u}{\partial x^2} \quad , \quad x \in [0, 2\pi] \quad ,$$

   subject to periodic boundary conditions.

   Show that for a Fourier-Galerkin approximation, the residual is

   $$R_N(x,t) = \sum_{n=-N/2}^{N/2} \left(\frac{d\hat{u}_n}{dt} - ina\hat{u}_n + n^2 b\hat{u}_n \right) \exp(inx) \quad .$$

3. Consider the variable coefficient problem

   $$\frac{\partial u}{\partial t} + \sin(x)\frac{\partial u}{\partial x} = 0 \quad , \quad x \in [0, 2\pi] \quad ,$$

   subject to periodic boundary conditions.
   Derive a Fourier-Galerkin approximation. Is $\mathcal{P}_N u = u_N$?.

4. Consider Burgers equation

   $$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = \varepsilon \frac{\partial^2 u}{\partial x^2} \quad , \quad x \in [0, 2\pi] \quad ,$$

subject to periodic boundary conditions.
Derive a Fourier-Galerkin approximation.

5. Consider the variable coefficient problem

$$\frac{\partial u}{\partial t} + \sin(x)\frac{\partial u}{\partial x} = 0 \quad , \quad x \in [0, 2\pi] \quad ,$$

subject to the boundary conditions

$$u(0,t) = u(\pi,t) = 0 \quad .$$

Derive a Fourier-Galerkin approximation.

6. (Continued) Assume that the solution is expressed as

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t)\cos(nx) \quad .$$

Derive a tau approximation.

7. Consider Burgers equation

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = \varepsilon\frac{\partial^2 u}{\partial x^2} \quad , \quad x \in [0, 2\pi] \quad ,$$

subject to periodic boundary conditions.
Derive a Fourier-Collocation approximation.

8. (Continued) Consider Burgers equation on the equivalent form

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \varepsilon\frac{\partial^2 u}{\partial x^2} \quad , \quad x \in [0, 2\pi] \quad ,$$

subject to periodic boundary conditions.
Derive a Fourier-Collocation approximation.
Will the two approximations yield the same results ? Why/why not ?

4

Trigonometric Polynomials

As discussed in the previous chapter, the choice of the appropriating basis function, i.e., the specification of the finite dimensional subspace, B_N, lies at the heart of the design of the spectral method. Indeed, as we learned through an example in Chapter 2, this choice greatly influences the overall performance of the scheme.

If we restrict the attention to problems possessing some degree of periodicity, it seems natural to consider the use of trigonometric polynomials, also known as Fourier series, for the purpose of representing the unknown solutions. However, as we experienced in Chapter 2, even for problems involving some degree of periodicity may result in a disappointing performance of schemes based on trigonometric polynomials.

In this Chapter we shall come to an understanding of exactly what determines the behavior of the approximating series. We will, for the sake of simplicity, consider functions, $u(x)$, of only one variable and defined on $[0, 2\pi]$. We shall also restrict ourselves to functions having a continuous periodic extension, i.e., $u(x) \in C_p^0[0, 2\pi]$. The behavior of trigonometric series for the approximation of piecewise smooth functions, $u(x) \in L^2[0, 2\pi]$ shall be revisited in Chapter 8.

4.1 Continuous Trigonometric Polynomials

The classic continuous series of trigonometric polynomials, also recognized as the Fourier series $F[u]$, for the approximation of a function, $u(x) \in L^2[0, 2\pi]$, is given as

$$F[u] = \hat{a}_0 + \sum_{n=1}^{\infty} \hat{a}_n \cos(nx) + \sum_{n=1}^{\infty} \hat{b}_n \sin(nx) \quad , \tag{4.1}$$

where the expansion coefficients are

$$\hat{a}_n = \frac{1}{\gamma_n}(u(x), \cos(nx))_{L^2[0,2\pi]} = \frac{1}{c_n \pi} \int_0^{2\pi} u(x) \cos(nx)\, dx \quad ,$$

with

$$c_n = \begin{cases} 2 & n = 0 \\ 1 & n > 0 \end{cases} \quad ,$$

and

$$\hat{b}_n = \frac{1}{\gamma_n}(u(x), \sin(nx))_{L^2[0,2\pi]} = \frac{1}{\pi} \int_0^{2\pi} u(x) \sin(nx)\, dx \quad , \quad n > 0 \quad .$$

This follows immediately from the orthogonality of the trigonometric functions in the unweighted inner product

$$(u, v)_{L^2[0,2\pi]} = \int_0^{2\pi} u(x) \overline{v}(x)\, dx \quad ,$$

with the associated norm

$$\|u\|_{L^2[0,2\pi]} = \left(\int_0^{2\pi} |u(x)|^2\, dx \right)^{1/2} \quad .$$

While orthogonality of the polynomials is advantageous, an essential property is $L^2[0, 2\pi]$-completeness. Establishing this for the trigonometric basis is, however, a classical, albeit somewhat complex, result, the proof of which is beyond the scope of the present text. We shall henceforth simply assume the validity of this result and refer to [??] where a complete proof of $L^2[0, 2\pi]$-completeness of the Fourier basis can be found.

Before we move on to study the properties of the approximating series, let us recall that the Fourier series can be expressed differently by introducing the Fourier basis functions

$$\phi_n(x) = \exp(inx) \quad .$$

Clearly, this set of functions is an orthogonal system over the interval $[0, 2\pi]$ with respect to a unity weight-function. By introducing the complex coefficients,

$$\hat{u}_n = \begin{cases} \hat{a}_0 & n = 0 \\ (\hat{a}_n - i\hat{b}_n)/2 & n > 0 \\ (\hat{a}_{-n} + i\hat{b}_{-n})/2 & n < 0 \end{cases} \tag{4.2}$$

the trigonometric series, Eq.(4.1), is seen to be equivalent to

$$F[u] = \sum_{|n|\leq\infty} \hat{u}_n \phi_n(x) \quad . \tag{4.3}$$

The expansion coefficients, $\hat{u}_n$, are obtained directly as

$$\hat{u}_n = \frac{1}{\gamma_n} \left(u, \exp(inx)\right)_{L^2[0,2\pi]} = \frac{1}{2\pi} \int_0^{2\pi} u(x) \exp(-inx) dx \quad .$$

A few notes concerning the Fourier series are in place. In the important special case where $u(x)$ is a real function, we recover that $\hat{a}_n$ as well as $\hat{b}_n$ are real numbers and, consequently, $\hat{u}_{-n} = \overline{\hat{u}_n}$ (see Eq.(4.2)), i.e., we only need half the coefficients to describe the function. Similar reductions are important when the function being approximated possesses certain symmetries. In case the function is even, i.e., $u(x) = u(-x)$, we have $\hat{b}_n = 0$ for all values of n. Consequently, one needs only consider the cosine series. Similarly, if the function is odd, i.e., $u(x) = -u(-x)$, we obtain $\hat{a}_n = 0$ for all n, recovering the sine series.

Let us now return to the convergence behavior of the truncated Fourier series

$$\mathcal{P}_N u(x) = \sum_{|n|\leq N/2} \hat{u}_n \exp(inx) \quad . \tag{4.4}$$

The central issue is how well does this truncated series approximate the function, $u(x) \in L^2[0, 2\pi]$, and in what sense can we talk about convergence of the series. Moreover, we need to come to an understanding of the convergence rate and how this depends on the properties of the function, $u(x)$, being approximated.

We seek an approximation to $u(x)$ in the finite dimensional subspace, $\hat{\mathsf{B}}_N$, defined as

$$\hat{\mathsf{B}}_N = \text{span}\{\exp(inx) | |n| \leq N/2\} \quad , \quad \dim(\hat{\mathsf{B}}_N) = N + 1 \quad .$$

Recall that $\mathcal{P}_N u$ is the orthogonal projection of $u(x)$ onto $\hat{\mathsf{B}}_N$ or, equivalently, $\mathcal{P}_N u$ is the closest element to $u(x)$ in $\hat{\mathsf{B}}_N$ with respect to $L^2[0, 2\pi]$.

Let us define the notion of periodicity.

Definition 6 (Periodicity). *A function, $u(x)$, $x \in [0, 2\pi]$, is periodic if $u(0)$ and $u(2\pi)$ exist and $u(0) = u(2\pi)$.*

Since $\mathcal{P}_N u$ is periodic, a sufficient condition for uniform convergence is that $u(x) \in L^2[0, 2\pi]$ itself is periodic and possesses a minimum amount of smoothness as stated in

Theorem 2. *Every function, $u(x) \in C_p^1[0, 2\pi]$, has a uniformly convergent Fourier series*

$$\|u - \mathcal{P}_N u\|_{L^\infty[0,2\pi]} \to 0 \quad as \quad N \to \infty \ .$$

The condition on smoothness is needed to ensure that

$$\sum_{|n|\leq\infty} |\hat{u}_n| < \infty \ .$$

as we shall discuss in relation with a direct proof given in Sec. 4.3.1.

A more general, but weaker, result is related to convergence in the mean as

Theorem 3. *Every piecewise continuous function, $u(x) \in L^2[0, 2\pi]$, can be expanded in a Fourier series, which is convergent in the mean*

$$\|u - \mathcal{P}_N u\|_{L^2[0,2\pi]} \to 0 \quad as \quad N \to \infty \ .$$

This is equivalent to the statement of $L^2[0, 2\pi]$-completeness of the Fourier basis and implies the existence of Parseval's identity as

$$\|u\|^2_{L^2[0,2\pi]} = 2\pi \sum_{|n|\leq\infty} |\hat{u}_n|^2 \ . \tag{4.5}$$

We note in particular that for $u(x) \in L^2[0, 2\pi]$, the sum on the right hand side is guaranteed to converge.

Utilizing Eq.(4.5) and orthogonality we find the truncation error introduced by the finite expansion as

$$\|u - \mathcal{P}_N u\|^2_{L^2[0,2\pi]} = 2\pi \sum_{|n|>N/2} |\hat{u}_n|^2 \ .$$

Moreover, provided $u(x) \in C_p^1[0,2\pi]$, Theorem 2 and the triangle inequality implies

$$\|u - \mathcal{P}_N u\|_{L^\infty[0,2\pi]} \leq \sum_{|n|>N/2} |\hat{u}_n| \ ,$$

i.e., the error committed by replacing $u(x)$ with its N'th order Fourier series depends solely on how fast the expansion coefficients of $u(x)$ decay. This, in turn, depends on the regularity of $u(x)$ in $[0,2\pi]$ and the periodicity of the function and its derivatives.

To appreciate this, let us assume that $u(x) \in C^0[0,2\pi]$. Provided $n \neq 0$, integration by parts implies

$$\begin{aligned} 2\pi\hat{u}_n &= \int_0^{2\pi} u(x)\exp(-inx)\,dx \\ &= \frac{-1}{in}(u(2\pi) - u(0)) + \frac{1}{in}\int_0^{2\pi} u'(x)\exp(-inx)\,dx \ . \end{aligned}$$

Clearly, if case $u'(x) \in L^2[0,2\pi]$, the integral exists and we recover

$$\hat{u}_n \propto \frac{1}{n} \ .$$

Moreover, if the function, $u(x) \in C_p^0[0,2\pi]$, we recover

$$\hat{u}_n \propto \frac{1}{n^2} \ ,$$

since $\hat{u}_n'$ must at least decay as n^{-1} if $u'(x) \in L^2[0,2\pi]$. Repeating this line of argument we have

Theorem 4. *If a function, $u(x) \in C_p^{m-2}[0,2\pi]$, then then the continuous Fourier expansion coefficients, $\hat{u}_n$, of $u(x)$ decay as*

$$\forall n \neq 0 : \hat{u}_n \propto \frac{1}{n^m} \ .$$

A Lemma of this yields

Lemma 2. If $u(x) \in C_p^\infty[0,2\pi]$ then the continuous Fourier expansion coefficients, $\hat{u}_n$, of $u(x)$ decay faster than any negative power of N.

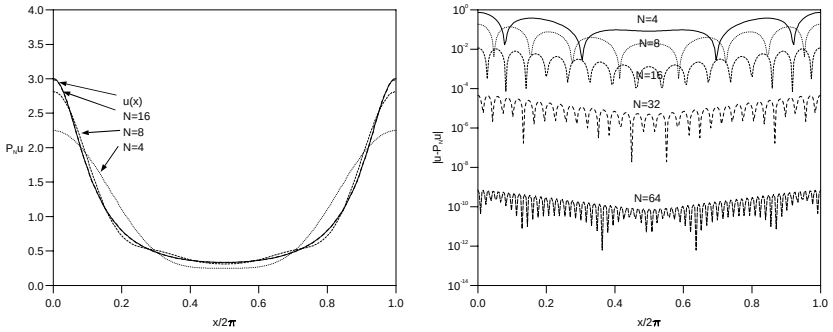

figure 4.1. a) Continuous Fourier series approximation of Example 9 for increasing resolution. b) Pointwise error of the approximation for increasing resolution

A few notes are in place here. First of all, one should realize that the rapid decay of the expansion coefficients, and thus the quickly vanishing truncation error, requires that both smoothness and periodicity of higher derivatives of the function. Furthermore, the asymptotic decay rate of the expansion coefficients is only observed for some $n > n_0$. In case the expansion is truncated below n_0 the approximation may be quite bad. This is true even for a C_p^∞-function. Such behavior is consistent with the results arrived at in Chapter 2 where we realized that the Fourier spectral method is useless if a minimum of two grid points per wavelength is used.

Let us consider a few examples.

Example 9. Consider the function, $u(x) \in C_p^\infty[0, 2\pi]$, defined as

$$u(x) = \frac{3}{5 - 4\cos(x)} \quad .$$

The expansion coefficients can be recovered as

$$\hat{u}_n = 2^{-|n|} \quad .$$

As expected, the expansion coefficients decay faster than any algebraic order of n. In Fig. 4.1 we plot the continuous Fourier series approximation of $u(x)$ and the pointwise error for increasing N.

This example clearly illustrates the fast convergence of the Fourier series

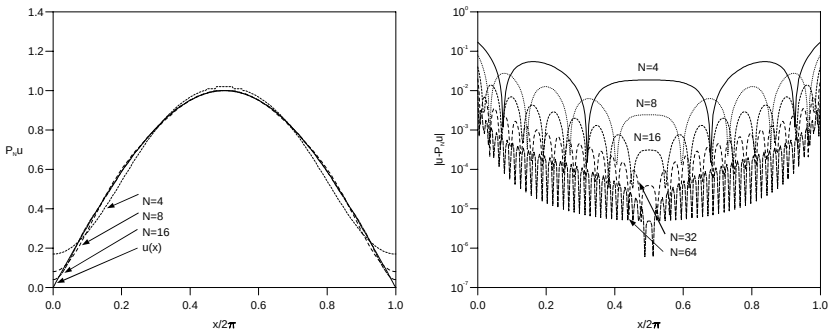

figure 4.2. a) Continuous Fourier series approximation of Example 10 for increasing resolution. b) Pointwise error of approximation for increasing resolution

and also that the convergence of the approximation is very close to being uniform. Note that only for $N > N_0 \sim 16$ do we observe the very fast convergence.

Example 10. Consider now the function

$$u(x) = \sin\left(\frac{x}{2}\right) \quad .$$

Note that $u(x) \in C_p^0[0, 2\pi]$ only. The expansion coefficients are given as

$$\hat{u}_n = \frac{2}{\pi} \frac{1}{1 - 4n^2} \quad ,$$

and we recover quadratic decay in n. In Fig. 4.2 we plot the continuous Fourier series approximation and the pointwise error for increasing N. As expected, we find quadratic convergence except near the endpoints where it is only linear.

Example 10 confirms convergence in the mean. However, we also observe a non-uniform pointwise convergence rate. This is a signature of using Fourier series, indeed of using most global expansions, for the approximation of functions that are not sufficiently smooth. We return to a discussion of this phenomenon, known as the Gibbs phenomenon, in Chapter 8.

4.1.1 Differentiation of the Continuous Expansion.

Representing the unknown function by a series of smooth basis functions imply that the expansion coefficients, $\hat{u}_n$, decay rapidly provided only that the function is sufficiently smooth. However, if the function itself is smooth, so is its derivatives. Thus, one should expect that also the expansion coefficients of the derivatives decay fast. This suggests that we, in contrast to conventional finite difference methods, may evaluate pointwise values of the derivatives with very high accuracy. This aspect is one of the main motivations for the use of spectral methods for solving partial differential equations.

The question is now the following. Given the expansion

$$u(x) = \sum_{|n|\leq\infty} \hat{u}_n \exp(inx) \quad ,$$

is it possible to obtain the expansion coefficients, $\hat{u}_n^{(q)}$, such that

$$\frac{d^q}{dx^q} u(x) = \sum_{|n|\leq\infty} \hat{u}_n^{(q)} \exp(inx) \quad .$$

The answer is, however, recovered directly from Eq.(4.3) as

$$\begin{aligned} u^{(q)}(x) &= \sum_{|n|\leq\infty} \hat{u}_n \frac{d^q}{dx^q} \exp(inx) = \sum_{|n|\leq\infty} (in)^q \hat{u}_n \exp(inx) \\ &= \sum_{|n|\leq\infty} \hat{u}_n^{(q)} \exp(inx) \quad , \end{aligned}$$

provided $u^{(q)}(x) \in C_p^1[0, 2\pi]$ to allow the interchange of the operators. As the basis functions are mutually orthogonal we have

$$\hat{u}_n^{(q)} = (in)^q \hat{u}_n \quad . \tag{4.6}$$

Clearly, if $\hat{u}_n$ decays exponentially, so does $\hat{u}_n^{(q)}$. Further insight into the convergence rate may be gained by realizing that if $u(x) \in C_p^m[0, 2\pi]$ and periodic we have

$$\hat{u}_n \propto \frac{1}{n^{m+2}} \quad \Rightarrow \quad \hat{u}_n^{(q)} \propto \frac{1}{n^{m+2-q}} \quad . \tag{4.7}$$

It is worth while observing that

$$\mathcal{P}_N \frac{d^q}{dx^q} u = \frac{d^q}{dx^q} \mathcal{P}_N u \quad ,$$

i.e., truncation and differentiation commutes for the continuous Fourier series. This implies that the truncation error, Eq.(3.13),

$$\mathcal{P}_N \mathcal{L} \left(\mathcal{I} - \mathcal{P}_N \right) u \quad ,$$

vanishes, explaining that the exact solution of certain types of equations is possible. Note, however, that this is a special result for $\mathcal{L}$ being the constant coefficient differential operator and does not carry over to other problems or methods.

4.2 Discrete Trigonometric Polynomials

The Achilles Heel of the continuous Fourier series method is the need to compute the continuous expansion coefficients through the inner product. In most situations it is indeed neither practical nor possible to evaluate this integral and it is certainly not practical in regards to computational implementations of the Fourier method.

The answer to this problem lies in the approximation of the Fourier integrals by using quadrature formulas, yielding the discrete Fourier coefficients.

Let us recall the definition of the continuous Fourier series

$$\mathcal{P}_N u(x) = \sum_{|n| \leq N/2} \hat{u}_n \exp(inx) \quad , \quad \hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(x) \exp(-inx)\, dx \tag{4.8}$$

In general, the integral can not be computed analytically and we resort to an approximating formula involving a set of grid points. However, the exact position of these grid points plays a crucial role and we shall subsequently split the analysis into a discussion of methods with an even number of grid points and methods with an odd number of grid points. As we shall learn shortly, the two schemes are clearly related but certainly also different in some important ways.

4.2.1 The Even Expansion.

Let us first consider an equidistant grid, consisting of N grid points, $x_j \in [0, 2\pi[$, defined as

$$x_j = \frac{2\pi j}{N} \quad j \in [0, \ldots, N-1] \ .$$

As always we assume that N is even.

One way to approximate the continuous integral is to apply the trapezoidal rule. Thus, we use the values of $u(x)$ at the N grid points to obtain an approximation, $\tilde{u}_n$, to $\hat{u}_n$ as

$$\hat{u}_n \simeq \tilde{u}_n = \frac{1}{N} \sum_{j=0}^{N-1} u(x_j) \exp(-inx_j) \ . \tag{4.9}$$

While the use of this approximation indeed looks innocent, it leads, as we shall realize shortly, to a different numerical scheme when compared to the continuous scheme.

Theorem 5. *The quadrature formula*

$$\frac{1}{2\pi} \int_0^{2\pi} u(x)\, dx = \frac{1}{N} \sum_{j=0}^{N-1} u(x_j) \ ,$$

is exact for any $u(x) \in \hat{\mathsf{B}}_{2N-1}$.

Proof: Assume that $u(x) \in C_p^1[0, 2\pi]$. Then $u(x)$ has a unique representation as

$$u(x) = \sum_{n=-\infty}^{\infty} \hat{u}_n \exp(inx) \ .$$

Let us now first consider the integral in Theorem 5.

$$\frac{1}{2\pi} \int_0^{2\pi} u(x)\, dx = \sum_{|n| \leq \infty} \hat{u}_n \frac{1}{2\pi} \int_0^{2\pi} \exp(inx)\, dx = \hat{u}_0 \ ,$$

due to orthogonality of $\exp(inx)$. Since $u(x) \in C_p^1[0, 2\pi]$ suffices to guarantee that the infinite sum is bounded, this allows for the interchange between integration and the infinite summation.

Considering the other part of the theorem we have

$$\begin{aligned}\frac{1}{N}\sum_{j=0}^{N-1} u(x_j) &= \frac{1}{N}\sum_{j=0}^{N-1}\left(\sum_{|n|\le\infty}\hat{u}_n \exp\left[in\frac{2\pi j}{N}\right]\right)\\ &= \sum_{|n|\le\infty}\hat{u}_n\left(\frac{1}{N}\sum_{j=0}^{N-1}\exp\left[in\frac{2\pi j}{N}\right]\right)\\ &= \hat{u}_0 + \sum_{\substack{|m|\le\infty\\ m\neq 0}}\hat{u}_{Nm} = \hat{u}_0 \quad ,\end{aligned}$$

due to Lemma 1. The last reduction is valid provided only that $\hat{u}_{Nm} \equiv 0$ for $m \neq 0$, i.e., $u(x) \in \hat{\mathsf{B}}_{2N-1}$. QED

Consequently, the trapezoidal rule yields a very good approximation to the inner product. One should note that the quadrature formula remains valid also for

$$u(x) = \sin(Nx) \quad ,$$

but not for $u(x) = \cos(Nx)$.

In what remains we use a slightly different definition of the discrete Fourier transform than appearing directly from the trapezoidal rule for reasons that will become apparent shortly. However, the methods are equivalent in terms of accuracy.

Let us define the complex discrete Fourier transform in $[0, 2\pi]$ as

$$\tilde{u}_n = \frac{1}{N\tilde{c}_n}\sum_{j=0}^{N-1} u(x_j)\exp(-inx_j) \quad , \tag{4.10}$$

with the inversion formula

$$\mathcal{I}_N u(x) = \sum_{|n|\le N/2}\tilde{u}_n \exp(inx) \quad , \tag{4.11}$$

where

$$\tilde{c}_n = \begin{cases} 2 & |n| = N/2 \\ 1 & |n| < N/2 \end{cases} \quad .$$

The need to introduce $\tilde{c}_n$ can be realized by observing that while we have N independent collocation points, we have $N+1$ expansion coefficients.

To resolve this indeterminacy we adopt the convention

$$\tilde{u}_{-N/2} = \tilde{u}_{N/2} \quad .$$

Adopting this notion has consequences for the dimension of the finite dimensional space, $\tilde{\mathsf{B}}_N$. Indeed, we find that

$$\tilde{\mathsf{B}}_N = \text{span}\{(\cos(nx), 0 \le n \le N/2) \cup (\sin(nx), 1 \le n \le N/2 - 1)\} \quad ,$$

with the dimension $\dim(\tilde{\mathsf{B}}_N) = N$. We note the difference from the projection operator, $\mathcal{P}_N$, which projects onto $\hat{\mathsf{B}}_N \neq \tilde{\mathsf{B}}_N$. The implications of this is seen by observing that

$$\mathcal{I}_N \cos\left(\frac{N}{2}x\right) = \cos\left(\frac{N}{2}x\right) \quad , \quad \mathcal{I}_N \sin\left(\frac{N}{2}x\right) = 0 \quad ,$$

since $\sin(Nx/2)$ is not a member of $\tilde{\mathsf{B}}_N$.

The particular definition of the discrete expansion coefficients introduced in Eq.(4.10) has the consequence that the trigonometric polynomial, $\mathcal{I}_N u$, interpolates the function, $u(x)$, at the quadrature nodes of the trapezoidal formula, i.e., $\mathcal{I}_N$ is the interpolation operator.

Theorem 6. *Let the discrete Fourier transform be defined as in Eqs. (4.10)-(4.11). For any periodic function, $u(x) \in C_p^0[0, 2\pi]$, we have*

$$\forall x_j = \frac{2\pi}{N} j \ : \ \mathcal{I}_N u(x_j) = u(x_j) \quad .$$

Proof: Substituting Eq.(4.10) into Eq.(4.11) we obtain

$$\mathcal{I}_N u(x) = \sum_{|n| \le N/2} \left(\frac{1}{N \tilde{c}_n} \sum_{j=0}^{N-1} u(x_j) \exp(-inx_j) \right) \exp(inx) \quad .$$

Exchanging the order of the summations yields

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N-1} u(x_j) g_j(x) \quad ,$$

where

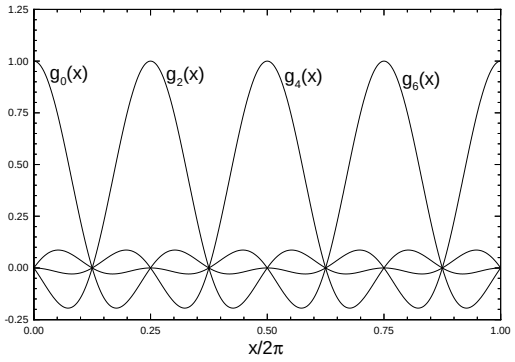

figure 4.3. The interpolation polynomial, $g_j(x)$, for $N = 8$ for various values of j.

$$g_j(x) = \sum_{|n|\le N/2} \frac{1}{N\tilde{c}_n} \exp\left[in(x - x_j)\right]$$
$$= \frac{1}{N} \sin\left[N\frac{x - x_j}{2}\right] \cot\left[\frac{x - x_j}{2}\right] \quad . \tag{4.12}$$

It is easily verified that $g_j(x_i) = \delta_{ij}$ as is also evident from the examples of $g_j(x)$ for $N = 8$ shown in Fig. 4.3.

We still need to show that $g_j(x) \in \tilde{\mathsf{B}}_N$. Clearly, $g_j(x) \in \hat{\mathsf{B}}_N$ as $g_j(x)$ is a polynomial of degree $\le N/2$. However, since

$$\frac{1}{2}\exp\left(-i\frac{N}{2}x_j\right) = \frac{1}{2}\exp\left(i\frac{N}{2}x_j\right) = \frac{(-1)^j}{2} \quad ,$$

and, by convention $\tilde{u}_{-N/2} = \tilde{u}_{N/2}$, we do not get any contribution from the term $\sin(N/2x)$, hence $g_j(x) \in \tilde{\mathsf{B}}_N$. QED

The discrete Fourier series of a function has convergence properties very similar to those discussed for the continuous Fourier series approximation. In particular, the discrete approximation is pointwise convergent for $C_p^1[0, 2\pi]$ functions and convergent in the mean provided only that $u(x) \in L^2[0, 2\pi]$. Moreover, the continuous and discrete approximations share the same asymptotic behavior, in particular having a convergence rate faster than any algebraic order of N^{-1} if $u(x) \in C_p^\infty[0, 2\pi]$.

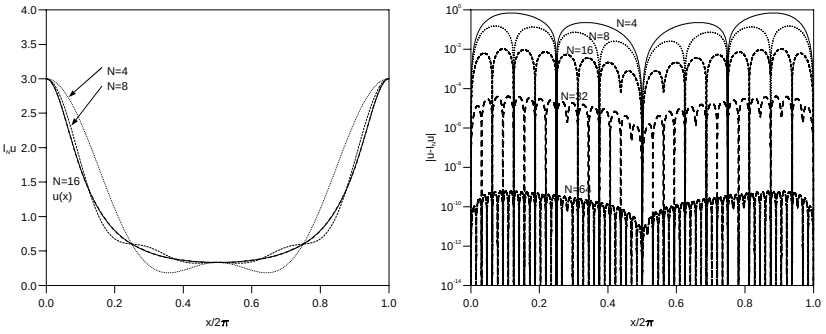

figure 4.4. a) Discrete Fourier series approximation of Ex. 11 for increasing resolution. b) Pointwise error of approximation for increasing resolution

We shall return to the proof of these results in Sec. 4.3.2.

Let us at this point illustrate the behavior of the discrete Fourier series by applying it to the examples considered previously.

Example 11. Consider the function, $u(x) \in C_p^\infty[0, 2\pi]$, defined as

$$u(x) = \frac{3}{5 - 4\cos(x)} \quad .$$

In Fig. 4.3 we plot the discrete Fourier series approximation of u and the pointwise error for increasing N.

This example confirms the spectral convergence of the discrete Fourier series. We note in particular that the approximation error is of the same order as observed for the continuous Fourier series in Ex. 9. The 'spikes' in the pointwise error approaching zero in Fig. 4.4 illustrates the interpolating nature of $\mathcal{I}_N u(x)$, i.e., $\mathcal{I}_N u(x_j) = u(x_j)$ as expected.

Example 12. Consider again the function

$$u(x) = \sin\left(\frac{x}{2}\right) \quad ,$$

and recall that $u(x) \in C_p^0[0, 2\pi]$. In Fig. 4.5 we show the discrete Fourier series approximation and the pointwise error for increasing N. As for the continuous Fourier series approximation we recover a quadratic convergence rate away from the boundary points at which it is only linear.

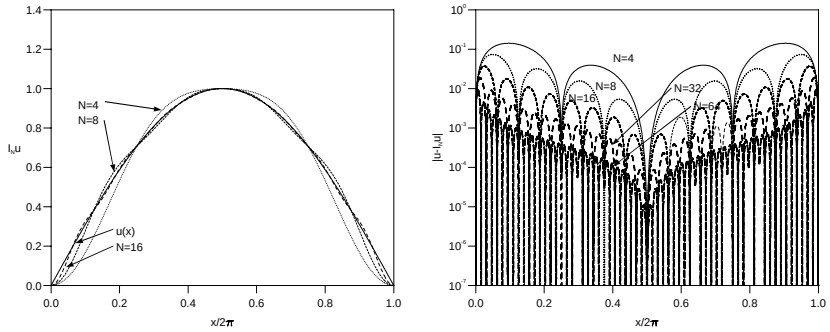

figure 4.5. a) Discrete Fourier series approximation of Ex. 12 for increasing resolution. b) Pointwise error of approximation for increasing resolution

---

4.2.2 The Odd Expansion.

Let us now briefly return to the situation where an odd number of grid points is used. As we saw in the previous section, using an even number of grid points implies that $\tilde{\mathsf{B}}_N \neq \hat{\mathsf{B}}_N$, as we only have N distinct points to determine the $N+1$ expansion coefficients. To construct a collocation method for which $\tilde{\mathsf{B}}_N = \hat{\mathsf{B}}_N$, let us define the grid as

$$x_j = \frac{2\pi}{N+1} j \quad , \quad j \in [0, \ldots, N] \quad , \tag{4.13}$$

in which case the interpolation operator becomes

$$\mathcal{J}_N u(x) = \sum_{|n| \leq N/2} \tilde{u}_n \exp(inx) \quad ,$$

and the expansion coefficients are given as

$$\tilde{u}_n = \frac{1}{N+1} \sum_{j=0}^{N} u(x_j) \exp(-inx_j) \quad . \tag{4.14}$$

Having $N+1$ distinct grid points to determine the $N+1$ expansion coefficients there is no need to impose additional restrictions on $\tilde{u}_n$

We recognize the definition of the expansion coefficients, Eq.(4.14), from the analysis of the infinite accuracy finite difference scheme dis-

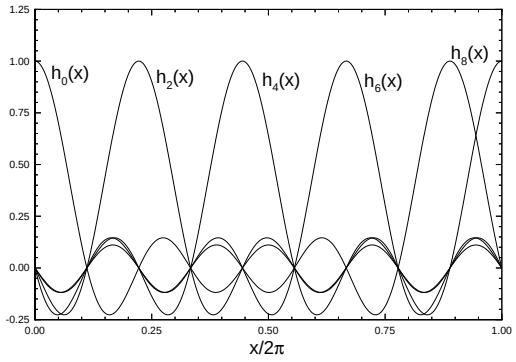

figure 4.6. The interpolation polynomial, $h_j(x)$, for $N = 8$ for various values of j.

cussed in Chapter 2. However, the summation of the series also provides an approximation to the continuous integral with an accuracy as

Theorem 7. *The quadrature formula*

$$\frac{1}{2\pi}\int_0^{2\pi} u(x)\,dx = \frac{1}{N+1}\sum_{j=0}^{N} u(x_j) \quad ,$$

is exact for any $u(x) \in \hat{\mathsf{B}}_{2N+1}$.

The scheme may also, as we have seen previously in Chapter 2, be expressed through the use of an Lagrange interpolation polynomial as

$$\mathcal{J}_N u(x) = \sum_{j=0}^{N} u(x_j) h_j(x) \quad ,$$

where

$$h_j(x) = \frac{1}{N+1}\frac{\sin\left(\frac{N+1}{2}(x-x_j)\right)}{\sin\left(\frac{x-x_j}{2}\right)} \quad . \tag{4.15}$$

One easily shows that $h_j(x_l) = \delta_{jl}$ and that $h_j(x) \in \hat{\mathsf{B}}_N$. Examples of $h_j(x)$ are shown in Fig. 4.6 for $N = 8$.

It may, at first, seem more natural to use this latter method rather than the previous approach utilizing an even number of points, since he former is equivalent to the continuous Fourier method. Historically, however, the even method has received much more interest due to the early availability of fast summation schemes, known as the Fast Fourier Transform, for the number of points being a power of two. However, as we shall discuss in Chapter 9, such fast methods are now available for an even as well as an odd number of grid points provided only that the total number of grid points has a prime-factorization using only small primes.

4.2.3 A First Look at the Aliasing Error.

Let us briefly consider the connection between the continuous Fourier series and the discrete Fourier series based on an even number of grid points. The conclusions of the discussion are, however, equally valid for the case of an odd number of points.

Assuming that the Fourier series converges pointwise, e.g., $u(x) \in C_p^1[0, 2\pi]$, a relation between the two sets of expansion coefficients is given as

$$\tilde{c}_n \tilde{u}_n = \hat{u}_n + \sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+Nm} \quad , \tag{4.16}$$

where the second term is a consequence of the discrete orthogonality of the Fourier basis, Lemma 1.

We observe that the n'th discrete Fourier mode depends not only on the n'th continuous mode of $u(x)$ but also on all higher frequencies. These are indistinguishable at the grid since

$$\exp\left[i(n+Nm)x_j\right] = \exp\left[inx_j\right] \exp\left[i2\pi mj\right] = \exp\left[inx_j\right] \quad .$$

The phenomenon that the $(n + Nm)$'th frequency is misinterpreted as the n'th frequency is termed static aliasing and appears as a result of the introduction of the grid. Another interpretation is that is introduced by the inaccuracy of the integration scheme.

In Fig. 4.7 we illustrate this phenomenon for $N = 8$ and we observe that the $n = -10$ wave as well as $n = 6$ wave can be interpreted as the $n = -2$ wave at the grid.

This aliasing introduces an error since high frequency components of

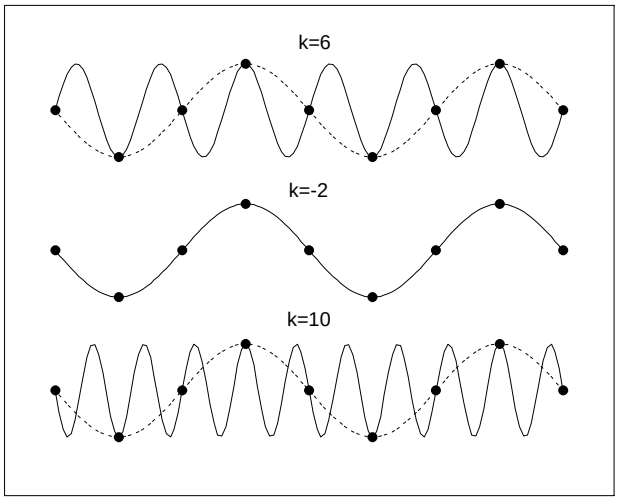

figure 4.7. Illustration of aliasing. The three waves, $n = 6$, $n = -2$ and $n = -10$ are all interpreted as a $n = -2$ wave on an 8-point grid. Consequently, the $n = -2$ appears as more energetic after the discrete Fourier transform than in the original signal.

$u(x)$ are misinterpreted as an additional and artificial contribution to the lower frequency components. The crucial question to ask is how important this effect is, i.e., how does the aliasing error

$$\|\mathcal{R}_N u\|^2_{L^2[0,2\pi]} = \left\| \sum_{n=-N/2}^{N/2} \left(\sum_{\substack{m=-\infty \\ m\neq 0}}^{m=\infty} \hat{u}_{n+Nm} \right) \exp(inx) \right\|^2_{L^2[0,2\pi]},$$

behave as N approaches infinity. As proven in Sec. 4.3.2, the aliasing error is of the same order as the truncation error, $\|u - \mathcal{P}_N u\|_{L^2[0,2\pi]}$ in the limit of large N. Hence, if the function is well approximated the aliasing error is generally negligible and the continuous Fourier series and the discrete Fourier series share similar approximation properties. For poorly resolved or nonsmooth problems, the situation is much more delicate ad we shall return to this concern later.

4.2.4 Differentiation of the Discrete Expansions.

As for the continuous series expansions, we shall need to address the question of how to recover derivatives of the approximated functions

themselves. As we have two computationally different, but mathematically equivalent, methods for expressing the interpolant we also obtain two computationally different ways by which to recover the approximate derivative of the function.

4.2.4.1 Using Expansion Coefficients.

Let us first consider the case where $u^{(q)} \in C_p^1[0, 2\pi]$ and the interpolant is given as

$$\mathcal{I}_N(\frac{d^q}{dx^q}u(x)) = \sum_{|n|\leq N/2} \tilde{u}_n^{(q)} \exp(inx_j) \quad .$$

Following the approach as for the continuous expansion yields

$$\begin{aligned}
\mathcal{I}_N u^{(q)}(x) &= \sum_{|n|\leq N/2} \tilde{u}_n^{(q)} \exp(inx) \\
&\simeq \sum_{|n|\leq N/2} \tilde{u}_n \frac{d^q}{dx^q} \exp(inx) = \sum_{|n|\leq N/2} (in)^q \tilde{u}_n \exp(inx) \\
&= \frac{d^q}{dx^q}\mathcal{I}_N u \quad .
\end{aligned}$$

Since the discrete exponential functions are mutually orthogonal up to the aliasing error, we recover

$$\tilde{u}_n^{(q)} \simeq (in)^q \tilde{u}_n \quad . \tag{4.17}$$

The results related to the continuous Fourier series carry over to the discrete Fourier series with the exception of the commutation of differentiation and interpolation, since in general

$$\mathcal{I}_N \frac{du}{dx} \neq \mathcal{I}_N \frac{d}{dx}\mathcal{I}_N u \quad , \tag{4.18}$$

unless $u(x) \in \tilde{\mathsf{B}}_N$. Consider the case where

$$u(x) = \sin\left(\frac{N}{2}x\right) \quad .$$

Clearly, $\mathcal{I}_N u \equiv 0$ since $u(x)$ is outside $\tilde{\mathsf{B}}_N$, i.e., $d(\mathcal{I}_N u)/dx = 0$. On the other hand, $u'(x) = N/2\cos(Nx/2)$, and $\mathcal{I}_N u'(x) = N/2\cos(Nx/2)$, illustrating Eq. (4.18). Consequently, differentiation can take a function

$u(x)$, originally outside of $\tilde{\mathsf{B}}_N$, into $\tilde{\mathsf{B}}_N$ contrary to the continuous case for which also $u(x) \in \hat{\mathsf{B}}_N$

Likewise, if we consider the scheme based on an odd number of modes, we have

$$\mathcal{J}_N \frac{du}{dx} \neq \mathcal{J}_N \frac{d}{dx} \mathcal{J}_N u \quad ,$$

except if $u \in \hat{\mathsf{B}}_N$. The source of this discrepancy is not the construction of the finite dimensional space but the aliasing error that appears for u not being in the space.

4.2.4.2 The Matrix Method.

Let us consider the approach for computing derivatives utilizing the alternative formulation, i.e., through the use of the Lagrange interpolation polynomials. If we consider the even method we have

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N-1} u(x_j) g_j(x) \quad ,$$

where

$$g_j(x) = \frac{1}{N} \sin\left[N \frac{x - x_j}{2}\right] \cot\left[\frac{x - x_j}{2}\right] \quad ,$$

as shown in Theorem 6. An approximation to the derivative at the collocation points, x_i, is then obtained by differentiating the interpolation directly

$$\left.\frac{d}{dx} \mathcal{I}_N u(x)\right|_{x_l} = \sum_{j=0}^{N-1} u(x_j) \left.\frac{d}{dx} g_j(x)\right|_{x_l} \quad .$$

The entries of the differential operator are given as

$$\mathrm{D}_{ij} = \left.\frac{d}{dx} g_j(x)\right|_{x_i} = \begin{cases} \frac{(-1)^{i+j}}{2} \cot\left[\frac{x_i - x_j}{2}\right] & i \neq j \\ 0 & i = j \end{cases} \quad . \tag{4.19}$$

Important properties of D are

Lemma 3. The Fourier differentiation matrix, D, is skew-symmetric.

Lemma 4. The differentiation matrix, D, is a circulant matrix.

The approximation of higher derivatives follows the exact same route as taken for the first order derivative. The entries of the second order differentiation matrix, $\mathrm{D}^{(2)}$, based on an even number of grid points, are

$$\left.\frac{d^2}{dx^2}g_j(x)\right|_{x_i} = \mathrm{D}^{(2)}_{ij} = \begin{cases} -\frac{(-1)^{i+j}}{2}\left[\sin\left(\frac{x_i-x_j}{2}\right)\right]^{-2} & i \neq j \\ -\frac{N^2+2}{12} & i = j \end{cases} \quad . \tag{4.20}$$

There is, however, a small complication to the computation of higher spatial derivatives in the case of the even approximation. To see this, consider the second order differentiation operator which allows for two different implementations. The first way is straightforward as

$$\mathcal{L}^1_N = \mathcal{I}_N \frac{d^2}{dx^2}\mathcal{I}_N = \mathrm{D}^{(2)} \quad ,$$

corresponding to the differentiation matrix given in Eq. (4.20). Alternatively, we could compute the second order derivative as

$$\mathcal{L}^2_N = \mathcal{I}_N \frac{d}{dx}\mathcal{I}_N \frac{d}{dx}\mathcal{I}_N = \mathrm{DD} \quad ,$$

which corresponds to defining $\mathrm{D}^{(2)} = \mathrm{DD}$, where the entries of D are given in Theorem **??**.

Let us now consider the action of these two operators on the function $u(x) = \cos(N/2x) \in \tilde{\mathsf{B}}_N$. Using $\mathcal{L}^1_N$ we obtain

$$\mathcal{L}^1_N u(x) = \mathcal{I}_N\left[-\left(\frac{N}{2}\right)^2 \cos\left(\frac{N}{2}x\right)\right] = -\left(\frac{N}{2}\right)^2 \cos\left(\frac{N}{2}x\right) \quad ,$$

i.e., the operator preserves the order of the polynomial. The action of $\mathcal{L}^2_N$, however, is

$$\mathcal{L}^2_N u(x) = \mathcal{I}_N \frac{d}{dx}\mathcal{I}_N\left[-\frac{N}{2}\sin\left(\frac{N}{2}x\right)\right] = 0 \quad ,$$

since $\sin(N/2x)$ is outside of $\tilde{\mathsf{B}}_N$ and we find that $\mathcal{L}^2_N$ reduces the order of the polynomial. Thus we have that

$$\mathrm{D}^{(2)} \neq \mathrm{DD} \quad .$$

It is natural to ask which of the two approximations one should use and the general answer is the former, i.e., $\mathcal{L}^1_N$, is the correct choice for reasons of accuracy. However, as we shall find in Chapter 5, there are

cases for which only the use of the $\mathcal{L}_N^2$ allows one to establish stability of the approximation. Note that this discrepancy is a property of the even order differentiations only.

In general, the q'th order differentiation matrix is obtained as

$$\mathrm{D}^{(q)} = \mathcal{I}_N \frac{d^q}{dx^q} \mathcal{I}_N \begin{cases} = (\mathrm{D})^q \text{ for } q \text{ odd} \\ \simeq (\mathrm{D})^q \text{ for } q \text{ even} \end{cases} .$$

We note that for q being odd, the matrices are all skew-symmetric, while for q being even the matrices are symmetric. Independent of q they are circulant.

Let us, for completeness, also give the differentiation matrix for the interpolation based on an odd number of collocation points

$$x_j = \frac{2\pi}{N+1} j \quad , \quad j \in [0, \ldots, N] \quad .$$

We recall that the interpolation operator is expressed in

$$\mathcal{J}_N u(x) = \sum_{j=0}^{N} u(y_j) h_j(x) \quad ,$$

where the interpolation polynomial, $h_j(x)$, is given in Eq. (4.15). From this we obtain the entries of differentiation matrix, $\tilde{\mathrm{D}}$, as

$$\tilde{\mathrm{D}}_{ij} = \begin{cases} \frac{(-1)^{i+j}}{2} \left[\sin\left(\frac{x_i - x_j}{2} \right) \right]^{-1} & i \neq j \\ 0 & i = j \end{cases} ,$$

which we recognize as the differentiation matrix studied in Chapter 2. The properties of $\tilde{\mathrm{D}}$ are similar to those of D, i.e. it is a skew-symmetric and circulant. It should be noted that for the method based on an odd number of points, we have the identity

$$\tilde{\mathrm{D}}^{(q)} = \mathcal{J}_N \frac{d^q}{dx^q} \mathcal{J}_N = (\tilde{\mathrm{D}})^q \quad ,$$

for all values of q.

4.2.4.3 A Comparison.

Let us finally compare the two mathematically equivalent, but computationally very different methods by which to recover approximations to the derivatives. The first method involves the computation of the

expansion coefficients through a summation of the series, obtaining the approximate expansion coefficients for the derivative and then summing once again to obtain the value of the derivative at the collocation points, or any point where the value of the derivative is required. In its most simple implementation this process required $\mathcal{O}(N^2)$ operations. However, the series appearing for some special values of N can be summed faster using the Fast Fourier Transforms which requires only $\mathcal{O}(N \log N)$ operations, making a significant difference for large values of N.

Let us illustrate this first approach by introducing

$$\boldsymbol{u} = [u(x_0), \ldots, u(x_{N-1})]^T \quad , \quad \tilde{\boldsymbol{u}} = [\tilde{u}_{-N/2}, \ldots, \tilde{u}_{N/2}]^T \quad ,$$

being simply the vectors of the grid points values and the discrete expansion coefficients with a connection between the two vectors as

$$\boldsymbol{u} = \mathrm{F}\tilde{\boldsymbol{u}} \quad , \quad \tilde{\boldsymbol{u}} = \mathrm{F}^{-1}\boldsymbol{u} \quad ,$$

The entries of the orthogonal, circulant matrices, F and F^{-1}, are obtained directly from Eqs.(4.10)-(4.11) as

$$\mathrm{F}_{kl} = \exp\left[i\left(l - \frac{N}{2}\right)x_k\right] \quad , \quad \mathrm{F}^{-1}_{kl} = \frac{1}{N\tilde{c}_{k-N/2}} \exp\left[-i\left(k - \frac{N}{2}\right)x_l\right] \quad .$$

If we now introduce the diagonal matrix

$$\mathrm{D}^{c,(q)} = \mathrm{diag}[(-iN/2)^q, \ldots, (-i)^q, 0, i^q, (iN/2)^q] \quad ,$$

corresponding to the continuous differentiation matrix, differentiation at the grid points using the expansion coefficients amounts to

$$\frac{d}{dx}\boldsymbol{u} = \mathrm{F}\mathrm{D}^{c,(q)}\mathrm{F}^{-1}\boldsymbol{u} \quad .$$

What makes this approach attractive is the sparsity of $\mathrm{D}^{c,(q)}$ and the observation that multiplication with F or its inverse can be accomplished in less that $\mathcal{O}(N^2)$ operations.

On the other hand, computing the derivatives at the collocation points using $\mathrm{D}^{(q)}$ involves a matrix-vector product

$$\frac{d}{dx}\boldsymbol{u} = \mathrm{D}^{(q)}\boldsymbol{u} \quad .$$

which is an $\mathcal{O}(N^2)$ operation. One should observe that in this latter

case will we never actually use the expansion coefficients. It seems that the first method is the fastest and should always be used. However, the efficiency of the Fast Fourier Transform is machine dependent and for small values of N is it may be faster to perform the matrix-vector product. Also, since the differentiation matrices are all circulant one need only store one column of the operator, thereby reducing the memory usage to that of the Fast Fourier Transform.

While the computational work associated with the two schemes may be different, the results are equivalent (up to finite precision effects) since

$$\mathrm{D}^{(q)} = \mathrm{F}\mathrm{D}^{c,(q)}\mathrm{F}^{-1} \quad .$$

4.3 Approximation Theory for Smooth Functions

So far we have focused our attention on the general properties of the Fourier expansions, be they based on continuous or discrete expansion coefficients, and paid less attention to a more accurate understanding of the properties of the approximations. It is the purpose of the present section to remedy this negligence.

As we discussed in Chapter 3, the actual rate of convergence of a stable and consistent scheme depends on the truncation error, Eq.(3.13),

$$\mathcal{P}_N \mathcal{L} \left(\mathcal{I} - \mathcal{P}_N \right) u \quad ,$$

which again depends on the particular projection operator, $\mathcal{P}_N$, and the operator, $\mathcal{L}$, being considered. Hence, to establish consistency we need to consider not only the difference between u and $\mathcal{P}_N u$, but also the distance between $\mathcal{L}u$ and $\mathcal{L}u_N$ where the this is measured in some appropriate norm.

Suppose that the operator, $\mathcal{L}$, is linear and of the form

$$\mathcal{L} = a_0(x) + a_1(x)\frac{d}{dx} + a_2(x)\frac{d^2}{dx^2} + \ldots + a_q(x)\frac{d^q}{dx^q} \quad ,$$

where $a_q(x) \in C_p^0[0, 2\pi]$. Provided $u(x)$ is sufficiently smooth, e.g., $u(x) \in C_p^{q-1}[0, 2\pi]$, we have

$$\|\mathcal{L}u\|^2_{L^2[0,2\pi]} = \int_0^{2\pi} \left| \sum_{m=0}^{q} a_m(x) u^{(m)}(x) \right|^2 dx$$

$$\leq \int_0^{2\pi} \sum_{m=0}^{q} |a_m(x)|^2 \left| u^{(m)}(x) \right|^2 \, dx \quad ,$$

using the triangle inequality. Since $a_m(x) \in C_p^0[0, 2\pi]$ it must be uniformly bounded by

$$A = \max_{m \in [0,q]} \|a_m(x)\|_{L^\infty[0,2\pi]} \quad .$$

This implies the bound

$$\|\mathcal{L}u\|^2_{L^2[0,2\pi]} \leq A^2 \int_0^{2\pi} \sum_{m=0}^{q} \left| u^{(m)}(x) \right|^2 \, dx = A^2 \|u\|^2_{H_p^q[0,2\pi]} \quad .$$

Hence, if we identify u with $u - \mathcal{P}_N u$, we can get an estimate of the truncation error by estimating the Sobolev norm on the right hand side of this last expression.

For a periodic function, $u(x) \in L^2[0, 2\pi]$, we know that the continuous Fourier expansion

$$u(x) = \sum_{|n| \leq \infty} \hat{u}_n \exp(inx) \quad ,$$

exists and the expansion coefficients, $\hat{u}_n$, are given as

$$\hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(x) \exp(-inx) \, dx \quad .$$

This implies

$$u^{(m)}(x) = \sum_{|n| \leq \infty} (in)^m \hat{u}_n \exp(inx) \quad ,$$

and enables an alternative expression of the Sobolev q-norm as

$$\begin{aligned} \|u\|^2_{H_p^q[0,2\pi]} &= \sum_{m=0}^{q} \int_0^{2\pi} \left| u^{(m)}(x) \right|^2 \, dx \\ &= 2\pi \sum_{m=0}^{q} \sum_{|n| \leq \infty} |n|^{2m} |\hat{u}_n|^2 = 2\pi \sum_{|n| \leq \infty} \left(\sum_{m=0}^{q} |n|^{2m} \right) |\hat{u}_n|^2 \quad , \end{aligned}$$

where the interchange of the summation is allowed provided $u(x)$ has

sufficient smoothness, e.g., $u(x) \in C_p^q[0, 2\pi]$.

It will prove useful to introduce a new norm, $\|\cdot\|_{W_p^q[0,2\pi]}$, equivalent to $\|\cdot\|_{H_p^q[0,2\pi]}$, as

$$\|u\|_{W_p^q[0,2\pi]} = \left(\sum_{|n|\leq\infty} (1+|n|)^{2q} |\hat{u}_n|^2 \right)^{1/2} ,$$

with the associated Sobolev space, $W_p^q[0, 2\pi]$ of functions for which

$$W_p^q[0, 2\pi] = \{u(x) \in L^2[0, 2\pi] | \|u\|_{W_p^q[0,2\pi]} < \infty\} .$$

The equivalence is realized since

$$\frac{1}{2^{2q}}(1+|n|)^{2q} \leq \sum_{m=0}^{q} n^{2m} \leq q(1+|n|)^{2q} .$$

It is possible to extend the definition of $W_p^q[0, 2\pi]$ to include real values of q as it appears as a power in the norm only.

4.3.1 Results for the Continuous Expansion.

Let us return to the estimation of the approximation error associated with the continuous Fourier series and seek an understanding of the accuracy of the truncated expansion.

We wish to estimate the difference between $\mathcal{L}u$ and $\mathcal{L}\mathcal{P}_N u$ in some appropriate norm, with the projection operator is given as

$$\mathcal{P}_N u(x) = \sum_{|n|\leq N} \hat{u}_n \exp(inx) .$$

Note that we, to simplify the notation, have changed the summation slightly compared to the previously used notation, i.e., we have $|n| \leq N$ instead of $|n| \leq N/2$.

Let us begin by discussing the approximation in the familiar L^2-norm for which we have the following result

Theorem 8. *For any for $u(x) \in H_p^r[0, 2\pi]$, there exists a positive constant C, independent of N, such that*

$$\|u - \mathcal{P}_N u\|_{L^2[0,2\pi]} \leq CN^{-q} \|u^{(q)}\|_{L^2[0,2\pi]} ,$$

provided $0 \leq q \leq r$.

Proof: The proof is easily established since

$$\|u - \mathcal{P}_N u\|^2_{L^2[0,2\pi]} = 2\pi \sum_{|n|>N} |\hat{u}_n|^2 \ ,$$

by Parseval's identity. Furthermore we have

$$\sum_{|n|>N} |\hat{u}_n|^2 = \sum_{|n|>N} \frac{1}{n^{2q}} n^{2q} |\hat{u}_n|^2 \leq N^{-2q} \sum_{|n|>N} n^{2q} |\hat{u}_n|^2 \ ,$$

and the last bracket can be bounded by $\|u^{(q)}\|_{L^2[0,2\pi]}$. QED

This result substantiates the claim put forward in Theorem 4. Moreover, assuming that $u(x) \in C^\infty_p[0,2\pi]$ is analytic we have

$$\|u^{(q)}\|_{L^2[0,2\pi]} \leq Cq! \, \|u\|_{L^2[0,2\pi]} \ ,$$

such that

$$\begin{aligned} \|u - \mathcal{P}_N u\|_{L^2[0,2\pi]} &\leq CN^{-q} \|u^{(q)}\|_{L^2[0,2\pi]} \sim C \frac{q!}{N^q} \|u\|_{L^2[0,2\pi]} \\ &\sim C \left(\frac{q}{N}\right)^q e^{-q} \|u\|_{L^2[0,2\pi]} \sim Ce^{-cN} \|u\|_{L^2[0,2\pi]} \ , \end{aligned}$$

assuming that $q \propto N$. This confirms the potential for exponentially fast convergence and provides a motivation for the title of exponentially accurate schemes often put on spectral methods.

A more general result is

Theorem 9. *For any real r and any real q where $0 \leq q \leq r$, with $u(x) \in W^r_p[0,2\pi]$, there exists a positive constant C, independent of N, such that*

$$\|u - \mathcal{P}_N u\|_{W^q_p[0,2\pi]} \leq C \frac{\|u\|_{W^r_p[0,2\pi]}}{N^{r-q}} \ .$$

Proof: Using Parseval's identity we have

$$\|u - \mathcal{P}_N u\|^2_{W^q_p[0,2\pi]} = \sum_{|n|>N} (1 + |n|)^{2q} |\hat{u}_n|^2 \ .$$

Since $|n|+1 \geq N$, we obtain

$$(1+|n|)^{2q} = \frac{(1+|n|)^{2r}}{(1+|n|)^{2(r-q)}} \leq \frac{(1+|n|)^{2r}}{N^{2(r-q)}} \quad ,$$

for any $q \leq r$. This immediately yields

$$\|u - \mathcal{P}_N u\|^2_{W_p^q[0,2\pi]} \leq C \sum_{|n|>N} \frac{(1+|n|)^{2r}}{N^{2(r-q)}} |\hat{u}_n|^2 \leq C \frac{\|u\|^2_{W_q^r[0,2\pi]}}{N^{2(r-q)}} \quad ,$$

and, thus, the result. QED

A bound on the pointwise error difference between $u(x) \in C_p^q[0,2\pi]$ and its projection, $\mathcal{P}_N u$, is given as

Theorem 10. *For any $q > 0$ and $u(x) \in C_p^q[0,2\pi]$, there exists a positive constant C, independent of N, such that*

$$|u - \mathcal{P}_N u| \leq C \frac{1}{N^{q-1/2}} \left\| u^{(q)} \right\|_{L^2[0,2\pi]} \quad .$$

Proof: Provided $u(x) \in C_p^q[0,2\pi]$, $q > 0$, we have for any $x \in [0,2\pi]$ that

$$|u - \mathcal{P}_N u| = |\sum_{|n|>N} \hat{u}_n \exp(inx)| \leq \sum_{|n|>N} |\hat{u}_n| \quad ,$$

by the triangle inequality.

Using the Cauchy-Schwarz inequality, we recover

$$\begin{aligned} \sum_{|n|>N} |\hat{u}_n| &= \sum_{|n|>N} \frac{1}{n^q} n^q |\hat{u}_n| \\ &\leq \left(\sum_{|n|>N} \frac{1}{n^{2q}} \right)^{1/2} \left(\sum_{|n|>N} n^{2q} |\hat{u}_n|^2 \right)^{1/2} \\ &\leq \frac{1}{N^{q-1/2}} \left\| u^{(q)} \right\|_{L^2[0,2\pi]} \quad , \end{aligned}$$

which completes the proof. QED

We observe that the leading error source in Theorem 10 is determined

solely by the regularity of the function being approximated. As the upper bound is independent of x we recover that for $u(x)$ being analytic, i.e., $u(x) \in C_p^\infty[0, 2\pi]$, the rate of pointwise convergence is faster than any algebraic power of $1/N$. This result is equivalent to that stated in Theorem 8, although here obtained in a stronger norm than $L^2[0, 2\pi]$, ensuring pointwise convergence.

4.3.2 Results for the Discrete Expansion.

For the discrete Fourier method we seek to estimate the difference between $\mathcal{L}u$ and $\mathcal{L}\mathcal{I}_N u$ in some norm. Let us begin by considering the interpolation operator

$$\mathcal{I}_N u = \sum_{n=-N}^{N} \tilde{u}_n \exp(inx) \quad ,$$

associated with an even number of grid points for which the the expansion coefficients given as

$$\tilde{u}_n = \frac{1}{2N\tilde{c}_n} \sum_{j=0}^{2N-1} u(x_j) \exp(-inx_j) \quad , \quad x_j = \frac{2\pi}{2N} j \quad .$$

Rather than deriving the estimates of the approximation error directly, we shall use the results obtained in the previous section and then estimate the difference between the two different expansions, recognized as the aliasing error.

The two sets of expansion coefficients are connected as

Lemma 5. Consider $u(x) \in W_p^r[0, 2\pi]$, where $r > 1/2$. For $|n| \leq N$ we have

$$\tilde{c}_n \tilde{u}_n = \hat{u}_n + \sum_{\substack{|m| \leq \infty \\ m \neq 0}} \hat{u}_{n+2Nm} \quad .$$

Proof: Substituting the continuous Fourier expansion into the discrete expansion yields

$$\tilde{c}_n \tilde{u}_n = \frac{1}{2N} \sum_{j=0}^{2N-1} \sum_{|l| \leq \infty} \hat{u}_l \exp(i(l-n)x_j) \quad .$$

To interchange the two summations we must ensure that

$$\sum_{|l|\leq\infty} |\hat{u}_l| < \infty \ .$$

Convergence of this series is established using

$$\begin{aligned}\sum_{|l|\leq\infty} |\hat{u}_l| &= \sum_{|l|\leq\infty} (1+|l|)^r \frac{|\hat{u}_l|}{(1+|l|)^r} \\ &\leq \left(\sum_{|l|\leq\infty} (1+|l|)^{2r} |\hat{u}_l|^2 \right)^{1/2} \left(\sum_{|l|\leq\infty} (1+|l|)^{-2r} \right)^{1/2} \ ,\end{aligned}$$

where the last expression follows from the Cauchy-Schwarz inequality. As $u(x) \in W_p^r[0,2\pi]$ the first part is clearly bounded. Furthermore, provided $r > 1/2$ the second term converges, hence ensuring boundedness.

Interchanging the order of summation and using orthogonality of the exponential function at the grid yields the result. QED

Let us first consider the behavior of the approximation in the familiar $L^2[0,2\pi]$-space. We have

Theorem 11. *For any $u(x) \in W_p^q[0,2\pi]$ with $q > 1/2$, there exists a positive constant C, independent of N, such that*

$$\|u - \mathcal{I}_N u\|_{L^2[0,2\pi]} \leq C N^{-q} \|u^{(q)}\|_{L^2[0,2\pi]} \ .$$

Proof: We begin by expanding the function, $u(x)$, in the continuous Fourier series and use Parseval's identity to obtain

$$\|u - \mathcal{I}_N u\|^2_{L^2[0,2\pi]} = \sum_{|n|\leq N} |\hat{u}_n - \tilde{u}_n|^2 + \sum_{|n|>N} |\hat{u}_n|^2 \ .$$

Consider first the case where $|n| < N$ such that $\tilde{c}_n = 1$. Theorem 5 implies

$$\sum_{|n|<N} |\hat{u}_n - \tilde{u}_n|^2 = \sum_{|n|<N} \left| \sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm} \right|^2 \ .$$

For the case of $|n| = N$, where $\tilde{c}_N = 2$, we have

$$\sum_{|n|=N} |\hat{u}_n - \tilde{u}_n|^2 \leq$$

$$\sum_{|n|=N} \left|\frac{1}{2}\hat{u}_n\right|^2 + \sum_{|n|=N} \left|\frac{1}{2} \sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm}\right|^2 \leq$$

$$\sum_{|n|=N} |\hat{u}_n|^2 + \sum_{|n|=N} \left|\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm}\right|^2 .$$

This yields

$$\|u - \mathcal{I}_N u\|^2_{L^2[0,2\pi]} \leq \sum_{|n|\geq N} |\hat{u}_n|^2 + \sum_{|n|\leq N} \left|\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm}\right|^2 .$$

The first term is bounded by the result of Theorem 8, representing the truncation error, while the second term measures the aliasing error.

To estimate this, we first note that

$$\left|\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm}\right|^2 = \left|\sum_{\substack{|m|\leq\infty \\ m\neq 0}} |n+2Nm|^q \hat{u}_{n+2Nm} \frac{1}{|n+2Nm|^q}\right|^2 .$$

Using the Cauchy-Schwarz inequality yields

$$\left|\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm}\right|^2 \leq \left(\sum_{\substack{|m|\leq\infty \\ m\neq 0}} |n+2Nm|^{2q} |\hat{u}_{n+2Nm}|^2\right) \left(\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \frac{1}{|n+2Nm|^{2q}}\right) .$$

Since $|n| \leq N$, bounding the second term is ensured by

$$\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \frac{1}{|n+2Nm|^{2q}} \leq \frac{2}{N^{2q}} \sum_{m=1}^{\infty} \frac{1}{(2m-1)^{2q}} = C_1 N^{-2q} \quad ,$$

provided $q > 1/2$. Here, the constant, C_1, is independent of N.

Utilizing that, we have

$$\sum_{|n|\leq N} \left| \sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{n+2Nm} \right|^2 \leq$$
$$\sum_{|n|\leq N} C_1 N^{-2q} \sum_{\substack{|m|\leq\infty \\ m\neq 0}} |n+2mN|^{2q} \, |\hat{u}_{n+2Nm}|^2 \leq$$
$$C_2 N^{-2q} \|u^{(q)}\|^2_{L^2[0,2\pi]} \quad .$$

The total error is thus bounded as

$$\|u - \mathcal{I}_N u\|^2_{L^2[0,2\pi]} \leq C N^{-2q} \|u^{(q)}\|^2_{L^2[0,2\pi]} + C_2 N^{-2q} \|u^{(q)}\|^2_{L^2[0,2\pi]} \quad ,$$

establishing the theorem. QED

Theorem 11 confirms that for $u(x)$ having only half a derivative, e.g., $u(x) \in C^0_p[0,2\pi]$, the approximation error of the continuous expansion and the discrete expansion are of the same order. Furthermore, the rate of convergence depends, in both cases, only on the smoothness of the function being approximated.

A similar result can be obtained in the Sobolev spaces as

Theorem 12. *Let $u(x) \in W^r_p[0,2\pi]$ where $r > 1/2$. Then for any real q for which $0 \leq q \leq r$, there exists a positive constant, C, independent of N such that*

$$\|u - \mathcal{I}_N u\|_{W^q_p[0,2\pi]} \leq C N^{-(r-q)} \|u\|_{W^r_p[0,2\pi]} \quad .$$

Proof: The proof follows that of Theorem 11. Using Parseval's theorem and considering $|n| = N$ and $|n| \neq N$ separately we recover

$$\|u - \mathcal{I}_N u\|^2_{W_p^q[0,2\pi]} = \sum_{|n|\le N} (1+|n|)^{2q} |\hat{u}_n - \tilde{u}_n|^2 + \sum_{|n|>N} (1+|n|)^{2q} |\hat{u}_n|^2$$

$$\le \sum_{|n|\ge N} (1+|n|)^{2q} \left|\hat{u}_n\right|^2 + \sum_{|n|\le N} (1+|n|)^{2q} \left| \sum_{\substack{|m|\le\infty \\ m\neq 0}} \hat{u}_{n+2Nm} \right|^2 .$$

The first term is bounded by Theorem 9.

The effect of the aliasing error can be estimated using

$$\left| \sum_{\substack{|m|\le\infty \\ m\neq 0}} \hat{u}_{n+2Nm} \right|^2 = \left| \sum_{\substack{|m|\le\infty \\ m\neq 0}} (1+|n+2Nm|)^r \hat{u}_{n+2Nm} \frac{1}{(1+|n+2Nm|)^r} \right|^2 ,$$

such that

$$\left| \sum_{\substack{|m|\le\infty \\ m\neq 0}} \hat{u}_{n+2Nm} \right|^2 \le \left(\sum_{\substack{|m|\le\infty \\ m\neq 0}} (1+|n+2Nm|)^{2r} \left|\hat{u}_{n+2Nm}\right|^2 \right) \left(\sum_{\substack{|m|\le\infty \\ m\neq 0}} \frac{1}{(1+|n+2Nm|)^{2r}} \right) .$$

The second factor is again bounded as

$$\sum_{\substack{|m|\le\infty \\ m\neq 0}} \frac{1}{(1+|n+2Nm|)^{2r}} \le \frac{2}{N^{2r}} \sum_{m=1}^{\infty} \frac{1}{(2m-1)^{2r}} = C_1 N^{-2r} ,$$

provided $r > 1/2$ and $|n| \le N$.

Also, since $(1+|n|)^{2q} \le C_2 N^{2q}$ for $|n| \le N$ we recover

$$\sum_{|n|\le N} (1+|n|)^{2q} \left| \sum_{\substack{|m|\le\infty \\ m\neq 0}} \hat{u}_{n+2Nm} \right|^2 \le$$

$$\sum_{|n|\leq N} C_1 C_2 N^{-2(r-q)} \sum_{\substack{|m|\leq\infty \\ m\neq 0}} (1+|n+2mN|)^{2r} \left|\hat{u}_{n+2Nm}\right|^2 \leq$$

$$C_3 N^{-2(r-q)} \|u\|^2_{W_p^r[0,2\pi]} \ .$$

This yields the bound

$$\|u - \mathcal{I}_N u\|^2_{W_p^q[0,2\pi]} \leq C N^{-2(r-q)} \|u\|^2_{W_p^r[0,2\pi]} + C_3 N^{-2(r-q)} \|u\|^2_{W_p^r[0,2\pi]} \ ,$$

and, thus, the result. QED

The results on the behavior of the aliasing error carries directly over to the estimates of errors lost due to lack commutation of interpolation and differentiation.

Lemma 6. Let $u(x) \in W_p^r[0,2\pi]$ where $r > 1$. Then there exists a positive constant, C, independent of N such that

$$\|u' - (\mathcal{I}_N u)'\|_{L^2[0,2\pi]} \leq C N^{-(r-1)} \|u\|_{W_p^r[0,2\pi]} \ .$$

This result confirms that for smooth problems, this error vanishes at approximately the same rate as the truncation error itself.

The results on the errors associated with the interpolation operator, $\mathcal{J}_N$, based on the odd number of grid points, are identical to those given above for $\mathcal{I}_N$ and can be obtained in a similar, albeit simpler, way.

The estimate put forward in Theorem 12 measures the truncation error in terms of the L^2-error of the function and its derivatives. However, the discrete Fourier expansions are constructed by means of interpolation of the function given at the grid points, x_j. It is may therefore seem more natural to measure the truncation error of the function and its derivatives at the grid points. To this end we introduce the grid based version of the Sobolev norms as

$$|||u - \mathcal{I}_N u|||_q = \left[\sum_{m=0}^{q} \frac{1}{2N} \sum_{j=0}^{2N-1} \left| u^{(m)}(x_j) - \mathcal{I}_N^{(m)} u(x_j) \right|^2 \right]^{1/2} , \quad (4.21)$$

for any integer q.

The connection between this norm and the Sobolev q-norm, $\|\cdot\|_{H_p^q[0,2\pi]}$, is given through the trapezoidal rule, Theorem 5, since

$$\frac{1}{2\pi}\int_0^{2\pi} u(x)\,dx - \frac{1}{2N}\sum_{j=0}^{2N-1} u(x_j) = -\sum_{\substack{|m|\leq\infty \\ m\neq 0}} \hat{u}_{m2N} \quad ,$$

i.e., the difference is due solely to the aliasing error.

Consequently, we have the relation

$$|||u - \mathcal{I}_N u|||_q \simeq \|u - \mathcal{I}_N u\|_{H_p^q[0,2\pi]} \simeq \|u - \mathcal{I}_N u\|_{W_p^q[0,2\pi]} \quad ,$$

leading to

Theorem 13. *Let $u(x) \in W_p^r[0,2\pi]$ where $r > 1/2$. Then for any real q where $0 \leq q \leq r$, there exists a positive constant, C, independent of N such that*

$$|||u - \mathcal{I}_N u|||_q \leq CN^{-(r-q)}\|u\|_{W_p^r[0,2\pi]} \quad .$$

Proof: The proof follows that of Theorem 12 for estimating the aliasing error and using Theorem 5 to establish the connection between the norm introduced in Eq.(4.21) and the $W_p^q[0,2\pi]$-norm. QED

Hence, the properties of the approximation carries over to the norms based on the discrete measures. A similar result can be derived for the interpolation operator, $\mathcal{J}_N$.

Exercises

1. Assume that $u(x) \in L^2[0, 2\pi]$, and that

$$\mathcal{P}_N u(x) = \sum_{|n| \le N/2} \hat{u}_n \exp(inx) \quad , \quad \hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(x) \exp(-inx)\, dx \quad .$$

Prove that convergence in the mean, Theorem 3, implies that Parseval's identity

$$\|u\|^2_{L^2[0,2\pi]} = 2\pi \sum_{|n| \le \infty} |\hat{u}_n|^2 \quad ,$$

is true.

2. (Continued) Prove the reverse, i.e., that Parseval's identity implies convergence in the mean.

3. Consider the sequence of functions

$$u^{q+1}(x) = \int_0^{2\pi} u^q(x)\, dx \quad ,$$

for $q = 0, 1, 2...$ and $u^0(x) = x$. Note that while $u^0(x) \in L^2[0, 2\pi]$, one has $u^q \in C_p^{q-1}[0, 2\pi]$ for $q > 0$.
According to Theorem 4, this means that the continuous expansion coefficients

$$\hat{u}_n^q \simeq \frac{1}{n^{q+1}} \quad ,$$

for large values of n.
Confirm that result by computing the expressions for $\hat{u}_n^q$ for a few values of q.

4. (Continued) Evaluate (using a computer) the L^2 and L^∞ error of the expansions, $\mathcal{P}_N$, and use that to confirm Theorems 2 and 3.

5. Assume that $u(x) \in L^2[0, 2\pi]$, and that

$$\mathcal{I}_N u(x) = \sum_{|n| \le N/2} \tilde{u}_n \exp(inx) \quad , \quad \tilde{u}_n = \frac{1}{N\tilde{c}_n} \sum_{j=0}^{N-1} u(x_j) \exp(-inx_j)\, dx \quad .$$

Here

$$x_j = \frac{2\pi}{N} j \quad , \quad c_n = 1 + \delta_{N/2,|n|} \quad .$$

Compute the L^2 and L^∞ errors of the expansion, $\mathcal{I}_N$, of u^q (see Problem 3) and compared with the behavior of the continuous expansion (by solving

Problems 3/4 or referring to Theorems 2 and 3).

6. (Continued) Compute the continuous expansion coefficients also and use those to evaluate the aliasing error directly. Can you confirm that

$$\|\mathcal{I}_N u\| = \|\mathcal{P}_N u\| + \|\mathcal{R}_N u\| \quad .$$

7. (Continued) Repeat the comparison, using the expansion based on an odd number of points. Does using $\mathcal{J}_N u$ make any significant difference ?

8. Prove that

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N-1} u(x_j) g_j(x) \quad ,$$

where

$$g_j(x) = \frac{1}{N} \sin\left[N\frac{x-x_j}{2}\right] \cot\left[\frac{x-x_j}{2}\right] \quad .$$

Prove also that $g_j(x_i) = \delta_{ij}$.

9. (Continued) Plot $g_j(x)$ for $N = 6$ to confirm the Lagrange property.

10. Prove Theorem 7.

11. Show that the entries of differentiation matrix, D, are given as in Eq. 4.19.

12. (Continued) Show that the entries of differentiation matrix, D, can likewise be derived by directly summing the series

$$\mathrm{D_{jl}} = \frac{1}{N} \sum_{n=-N/2}^{N/2} \frac{in}{\tilde{c}_n} \exp\left[in\frac{2\pi}{N}(j-l)\right] \quad .$$

13. Show that D is skew-symmetric, i.e., $\mathrm{D} = -\mathrm{D}^T$.

14. Show that D is circulant, i.e., that it is a Toeplitz matrix ($\mathrm{D}_{i,j} = \mathrm{D}_{i+1,j+1}$) and that it rows/columns wraps around ($\mathrm{D}_{i,N-1} = \mathrm{D}_{i+1,0}$).

15. Show that the entries of $\mathrm{D}^{(2)}$ are as given in Eq.(4.20).

16. Show directly that $\mathrm{DD} \neq \mathrm{D}^{(2)}$ as discussed in the text.

17. Prove that the differentiation matrix, $\tilde{\mathrm{D}}$, associated with the odd expansion has the entries

$$\tilde{\mathrm{D}}_{ij} = \begin{cases} \frac{(-1)^{i+j}}{2}\left[\sin\left(\frac{x_i - x_j}{2}\right)\right]^{-1} & i \neq j \\ 0 & i = j \end{cases} \quad .$$

18. (Continued) Derive the entries for $\tilde{\mathrm{D}}^{(2)}$ and show that $\tilde{\mathrm{D}}\tilde{\mathrm{D}} = \tilde{\mathrm{D}}^{(2)}$.

19. Consider the 4 functions defined on $x \in [0, 2\pi]$.

    (a) $u(x) = |sin(x)|$.
    (b) $u(x) = \exp\left(-\frac{1}{\sin(x)}\right)$.
    (c) $u(x) = (1 + \sin^2(x))^{-1}$.
    (d) $u(x) = \sin(e^{\pi} x)$.

    Compute the derivative of the $u(x)$ using the even method and evaluate the L^2 and the L^∞ error for increasing values of N. Explain the differences in the convergence behavior.

20. (Continued). Plot the distribution of the pointwise error and relate that to the features of the functions. Do you see uniform convergence ? – if not, why not ?.

21. Prove that

$$\frac{1}{2^{2q}}(1 + |n|)^{2q} \leq \sum_{m=0}^{q} n^{2m} \leq q(1 + |n|)^{2q} \quad ,$$

    to establish that $W_p^q[0, 2\pi]$ and $H_p^q[0, 2\pi]$ are equivalent spaces.

22. Prove the equivalent of Theorem 5 for the odd expansion, $\mathcal{J}_N u$.

23. Prove the equivalent of Theorem 11 for the odd expansion, $\mathcal{J}_N u$.

24. Prove the equivalent of Theorem 12 for the odd expansion, $\mathcal{J}_N u$.

5

Fourier Spectral Methods

Understanding the properties of the Fourier series, we are now equipped to consider the formulation of Fourier spectral methods for the solution of partial differential equations. As in the previous chapter we restrict ourselves to problems stated on $[0, 2\pi]$ and assume that the solutions, $u(x)$, can be periodically extended. The assumption of periodicity suggests that we may disregard the τ-method introduced in Chap. 3.4 and focus the attention on Galerkin and Collocation methods. While these methods are equivalent for problems involving only linear, constant coefficient operators and bandlimited initial conditions, the discrepancy is significant in more general cases of variable coefficients or nonlinear problems. As we shall see, these differences are not restricted to issues of implementation only but appear already at the level of the formulation of the schemes.

The second part of this chapter is devoted to an analysis of the stability of some of the semi-discrete schemes discussed in the first part. While the stability of the Galerkin methods is closely related to properties of the partial differential equation itself, the analysis of stability for the collocation method turns out to be considerably more involved.

5.1 The Construction of Fourier Spectral Methods

As the construction of the Galerkin and Collocation schemes is based on fundamentally different principles of satisfying the partial differential equation we discuss the two approaches separately. However, much of the following is centered around examples and we shall strive to directly compare the two approaches.

5.1.1 Fourier-Galerkin Methods

Let us assume that the solution, $u(x,t) \in L^2[0,2\pi]$, is periodic and that we have

$$u(x,t) = \sum_{|n|\leq\infty} \hat{u}_n(t) \exp(inx) \quad .$$

In the Fourier-Galerkin method, we seek solutions, $u_N(x,t) \in \hat{\mathsf{B}}_N$ with $\hat{\mathsf{B}}_N \in \text{span}\{\exp(inx)\}_{|n|\leq N/2}$ to the partial differential equation of the form

$$u_N(x,t) = \sum_{|n|\leq N/2} \hat{u}_n(t) \exp(inx) \quad .$$

We recall that the continuous expansion coefficients, $\hat{u}_n(t)$, are given as

$$\hat{u}_n(t) = \frac{1}{2\pi} \int_0^{2\pi} u(x,t) \exp(-inx)\, dx \quad .$$

Consider the problem

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= \mathcal{L}u(x,t) \quad , \quad x \in [0,2\pi] \ , \ t \geq 0 \ , \\ u(x,0) &= g(x) \quad , \qquad x \in [0,2\pi] \ , \ t = 0 \ , \end{aligned}$$

and let us seek solutions, $u_N(x,t)$, such that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - \mathcal{L}u_N(x,t) \quad ,$$

is orthogonal to $\hat{\mathsf{B}}_N$, i.e.,

$$\forall\, |n| \leq \frac{N}{2} \ : \ \frac{1}{2\pi} \left(R_N, \exp(inx)\right)_{L^2[0,2\pi]} = 0 \quad .$$

The initial conditions are

$$u_N(x,0) = \sum_{|n|\leq N/2} \hat{g}_n \quad , \quad \hat{g}_n = \frac{1}{2\pi} \left(g, \exp(inx)\right)_{L^2[0,2\pi]} \quad .$$

In other words, if we express the residual as

$$R_N(x,t) = \sum_{|n|\leq\infty} \hat{R}_n(t) \exp(inx) \quad ,$$

we recover $N+1$ equations to determine the $N+1$ unknowns, $\hat{u}_n$, representing the solution, $u_n(x,t)$, by requiring that

$$\forall\, |n| \leq \frac{N}{2} \;:\; \hat{R}_n(t) = \frac{1}{2\pi}\int_0^{2\pi} R_N(x,t)\exp(-inx)\,dx = 0 \quad .$$

The crucial point here is the projection of the residual onto $\hat{\mathsf{B}}_N$, a process that can be very difficult and even impossible depending on the equation.

Let us discuss a number of examples of increasing complexity to come to an understanding of the strength and the limitations of the Fourier-Galerkin method. As we shall see, there are indeed classes of problems where the Fourier-Galerkin approach is superior and solves the problem exactly as there are cases where it escapes formulation entirely.

Example 13. Consider the linear constant coefficient problem

$$\frac{\partial u(x,t)}{\partial t} = a\frac{\partial^q}{\partial x^q}u(x,t) \quad ,$$

with the assumption that $u(x,t) \in C_p^\infty[0,2\pi]$, a is a constant, and $q \geq 0$ signifies the order of differentiation.

To recover the approximate solution we seek a trigonometric polynomial,

$$u_N(x,t) = \sum_{|n|\leq N/2} \hat{u}_n(t)\exp(inx) \quad ,$$

such that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - a\frac{\partial^q}{\partial x^q}u_N(x,t) \quad ,$$

is orthogonal to $\hat{\mathsf{B}}_N$.

From Chap. 4.1.1 we recall

$$\frac{\partial^q}{\partial x^q}u_N(x,t) = \sum_{|n|\leq N/2} (in)^q \hat{u}_n(t)\exp(inx) \quad ,$$

and recover the residual directly

$$R_N(x,t) = \sum_{|n|\leq N/2}\left(\frac{d\hat{u}_n(t)}{dt} - a(in)^q\hat{u}_n(t)\right)\exp(inx) \quad .$$

We observe that $R_N(x,t) \in \hat{\mathsf{B}}_N$, i.e., the first $N/2$ terms of the equations can be solved exactly as discussed in Chapter 2.

The $N+1$ ordinary differential equations (ODE) describing the evolution of the continuous expansion coefficients, $\hat{u}_n(t)$, are obtained directly by requiring that

$$\forall\, |n| \leq \frac{N}{2} \;:\; \hat{R}_n(t) = 0 \;\;\Rightarrow\;\; \frac{d\hat{u}_n(t)}{dt} = a(in)^q \hat{u}_n(t) \;\; .$$

In this particular case we recover that $\mathcal{P}_N u(x,t) = u_N(x,t)$ as the truncation error, Eq.(3.13), vanishes identically.

The vanishing truncation error is, as we have discussed earlier, a particular result that does not extend beyond the case of linear, constant coefficient periodic problems.

Example 14. Consider the linear, variable coefficient problem

$$\frac{\partial u(x,t)}{\partial t} = \sin(x) \frac{\partial u(x,t)}{\partial x} \;\; ,$$

where the initial conditions are given through $g(x)$ and the solution, $u(x,t) \in C_p^\infty[0, 2\pi]$.

We seek solutions in the form of a trigonometric polynomial

$$u_N(x,t) = \sum_{|n| \leq N/2} \hat{u}_n(t) \exp(inx) \;\; , \tag{5.1}$$

and require that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - \sin(x) \frac{\partial}{\partial x} u_N(x,t) \;\; ,$$

is orthogonal to $\hat{\mathsf{B}}_N$.

The residual is given as

$$R_N(x,t) = \sum_{|n| \leq N/2} \left(\frac{d\hat{u}_n(t)}{dt} - \frac{1}{2i} \left(\exp(ix) - \exp(-ix) \right) (in) \hat{u}_n(t) \right) \exp(inx) \;\; .$$

If we assume that $\hat{u}_{-(N/2+1)}(t) = \hat{u}_{N/2+1}(t) = 0$ in accordance with the basis assumption on $u_N(x,t)$, Eq.(5.1), the residual becomes

$$
\begin{aligned}
R_N(x,t) = & \sum_{|n|\le N/2} \frac{d\hat{u}_n(t)}{dt}\exp(inx) - \frac{1}{2}\sum_{|n|\le N/2} n\exp[i(n+1)x]\hat{u}_n(t) \\
& + \frac{1}{2}\sum_{|n|\le N/2} n\exp[i(n-1)x]\hat{u}_n(t) \\
= & \sum_{|n|\le N/2} \frac{d\hat{u}_n(t)}{dt}\exp(inx) - \frac{1}{2}\sum_{|n|\le N/2} (n-1)\exp(inx)\hat{u}_{n-1}(t) \\
& + \frac{1}{2}\sum_{|n|\le N/2} (n+1)\exp(inx)\hat{u}_{n+1}(t) \\
& - \frac{N}{4}\left(\exp\left[i\frac{N+2}{2}x\right]\hat{u}_{N/2}(t) + \exp\left[-i\frac{N+2}{2}x\right]\hat{u}_{-N/2}(t)\right) \\
= & \sum_{|n|\le N/2} \left(\frac{d\hat{u}_n(t)}{dt} - \frac{n-1}{2}\hat{u}_{n-1}(t) + \frac{n+1}{2}\hat{u}_{n+1}(t)\right)\exp(inx) \\
& - \frac{N}{4}\left(\exp\left[i\frac{N+2}{2}x\right]\hat{u}_{N/2}(t) + \exp\left[-i\frac{N+2}{2}x\right]\hat{u}_{-N/2}(t)\right) \quad .
\end{aligned}
$$

We note that, contrary to the situation in the previous example, $R_N(x,t)$ is not solely in the space of $\hat{\mathsf{B}}_N$ due to the two extra terms, i.e., $R_N(x,t) \in \hat{\mathsf{B}}_{N+1}$. Hence, requiring that the residual is orthogonal to $\hat{\mathsf{B}}_N$ results in $N+1$ coupled ODE's

$$
\frac{d\hat{u}_n(t)}{dt} - \frac{n-1}{2}\hat{u}_{n-1}(t) + \frac{n+1}{2}\hat{u}_{n+1}(t) = 0 \quad ,
$$

with $\hat{u}_{-(N/2+1)}(t) = \hat{u}_{N/2+1}(t) = 0$ and introduces a truncation error.

In the above variable coefficient problem we find that the projection of the residual vanishes rather than the residual itself as it not contained entirely within $\hat{\mathsf{B}}_N$. However, one should note that since $u(x,t)$ is assumed smooth we know that for large values of N, the expansion coefficients, $\hat{u}_n(t)$, decay exponentially fast in N, and the part of the residual being outside $\hat{\mathsf{B}}_N$ can thus be assumed to be very small provided N is sufficiently large.

As always, the formulation of the Fourier-Galerkin method involves the derivation of the equations for the expansion coefficients of the unknown solution. While this was relatively easy for the particular variable coefficient case considered in the previous example this is not always the

case. Moreover, the resulting equations may be somewhat complicated as we shall find in the next example.

Example 15. Consider the nonlinear problem

$$\frac{\partial u(x,t)}{\partial t} = u(x,t)\frac{\partial u(x,t)}{\partial x} \quad ,$$

where the initial conditions are given through $g(x)$ and the solution, $u(x,t) \in C_p^\infty[0,2\pi]$, local in time.

We seek a solution on the form of a trigonometric polynomial

$$u_N(x,t) = \sum_{|n|\leq N/2} \hat{u}_n(t)\exp(inx) \quad ,$$

and require that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - u_N(x,t)\frac{\partial}{\partial x}u_N(x,t) \quad ,$$

be orthogonal to $\hat{\mathsf{B}}_N$.

Consider first the quadratic nonlinearity

$$\begin{aligned} u_N(x,t)\frac{\partial}{\partial x}u_N(x,t) &= \sum_{|l|\leq N/2}\sum_{|k|\leq N/2} \hat{u}_l(t)(ik)\hat{u}_k(t)\exp\left[i(l+k)x\right] \\ &= \sum_{|k|\leq N/2}\sum_{n=-N/2+k}^{N/2+k} (ik)\hat{u}_{n-k}(t)\hat{u}_k(t)\exp(inx) \quad . \end{aligned}$$

This shows that the residual, $R_N(x,t) \in \hat{\mathsf{B}}_{2N}$, as a consequence of the quadratic nonlinearity and the associated three-wave mixing. Hence, we have $2N+1$ equations but can recover $N+1$ conditions by requiring that $\mathcal{P}_N R_N = 0$. Satisfying the first $N+1$ conditions, which results in a set of ODE's of the form

$$\frac{d\hat{u}_n(t)}{dt} = \sum_{|k|\leq N/2} (ik)\hat{u}_{n-k}(t)\hat{u}_k(t) \quad ,$$

representing the first $N+1$ Fourier coefficients of the solution.

Although we could only derive an approximate scheme for the quadratic

non-linearity, it was nevertheless possible to arrive at the Fourier-Galerkin scheme suitable for solving the problem. This is, in fact, a strike of good fortune caused by the special nonlinearity we considered. If the nonlinearity is stronger, we may well be unable to derive the Fourier-Galerkin equations as illustrated in the last example.

Example 16. Consider the strongly nonlinear problem

$$\frac{\partial u(x,t)}{\partial t} = \exp\left[u(x,t)\right] \frac{\partial u(x,t)}{\partial x} \quad ,$$

where the initial conditions are given through $g(x)$ and the solution, $u(x,t)$ is assumed to be smooth and periodic, at least local in time.

As usual, we seek a trigonometric polynomial

$$u_N(x,t) = \sum_{|n| \le N/2} \hat{u}_n(t) \exp(inx) \quad ,$$

and require that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - \exp\left[u_N(x,t)\right] \frac{\partial}{\partial x} u_N(x,t) \quad ,$$

is orthogonal to $\hat{\mathsf{B}}_N$.

This results in the constraints

$$\frac{d\hat{u}_n(t)}{dt} - \frac{1}{2\pi} \sum_{|k| \le N/2} ik\hat{u}_k(t) * \left(\exp\left[\sum_{|l| \le N/2} \hat{u}_l(t) \exp(ilx) \right], \exp[i(n-k)x] \right)_{L^2[0,2\pi]} = 0 \quad .$$

However, we are unable to evaluate the inner product and, hence, unable to even formulate the Fourier-Galerkin scheme.

To summarize, the grid free Fourier-Galerkin method is very efficient for linear, constant coefficient problems, but tends to become complex even for simple variable coefficient problems and nonlinear problem. The main drawback of the method is the need to derive the system of governing ordinary differential equations, which results from requesting that the residual be orthogonal to $\hat{\mathsf{B}}_N$. Every partial differential equation

will result in a different set of ordinary differential equations and, as experienced through the examples, the evaluation of the inner products is by no means a straightforward task.

5.1.2 Fourier-Collocation Methods

Even if one can derive the Fourier-Galerkin scheme for a particular problem, one often needs to approximate the inner products by sums to evaluate the initial conditions. This will generally introduce an aliasing error which would otherwise be absent from the Galerkin formulation. Given that some aliasing error will be present even in a Fourier-Galerkin scheme, one may as well utilize the introduction of grids to ones advantage. This is exactly what is happening in the collocation methods.

To define a Fourier-Collocation method we must introduce a grid, y_j, at which to require that the residual vanishes identically. It is important to appreciate that this grid need not be the same as the grid, termed x_j, on which the interpolation itself is based.

For the latter we restrict the attention to approximations based on the grid

$$x_j = \frac{2\pi}{N} j \quad , \quad j \in [0, \ldots, N-1] \quad ,$$

where N is assumed even. Keep in mind, however, that everything said about this specific choice holds for schemes based on the odd method also.

We assume that the solution, $u(x,t) \in L^2[0,2\pi]$, is periodic and consider again the general problem

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= \mathcal{L}u(x,t) \quad , \quad && x \in [0,2\pi] \ , \ t \geq 0 \ , \\ u(x,0) &= g(x) \quad , && x \in [0,2\pi] \ , \ t = 0 \ . \end{aligned}$$

In the Fourier-Collocation method we seek solutions, $u_N \in \tilde{\mathsf{B}}_N$, of the form

$$u_N(x,t) = \sum_{|n| \leq N/2} \tilde{u}_n(t) \exp(inx) \quad ,$$

with the discrete expansion coefficients, $\tilde{u}_n(t)$, being

$$\tilde{u}_n(t) = \frac{1}{\tilde{c}_n N} \sum_{j=0}^{N-1} u(x_j, t) \exp(-inx) \ ,$$

and we recall that $\tilde{c}_{-N/2} = \tilde{c}_{N/2} = 2$ and $\tilde{c}_n = 1$ otherwise.

As discussed in Sec. 4.2 the discrete polynomial has a dual expression on the form

$$u_N(x,t) = \sum_{j=0}^{N-1} u(x_j, t) g_j(x) \ ,$$

where $g_j(x)$ represents the Lagrange interpolation polynomial, Eq.(4.12).

We require the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - \mathcal{L} u_N(x,t) \ ,$$

to vanish at the grid points, y_j, i.e.,

$$\forall y_j \ : \ R_N(y_j, t) = 0 \ , \quad j \in [0, \ldots, N-1] \ . \tag{5.2}$$

Note in particular that we do not require the residual to be orthogonal to the subspace, $\tilde{\mathsf{B}}_N$, as was the case in the Fourier-Galerkin method. The requirement, Eq.(5.2), yields N equations to determine the N point values, $u_N(x_j, t)$, of the solution.

Let us now, as we did for the Fourier-Galerkin method, consider a number of examples of increasing complexity. For the purpose of comparison, we will discuss the same problems as for the Fourier-Galerkin methods with the exception of the Ex. 14 which we considered already in Chapter 3. Moreover, except stated explicitly, we shall restrict the discussion to the situation where $y_j = x_j$, i.e., the equations are required to be satisfied at the same set of nodes as those on which the approximation is based.

Example 17. Consider first the linear constant coefficient problem

$$\frac{\partial u(x,t)}{\partial t} = a \frac{\partial^q}{\partial x^q} u(x,t) \ ,$$

assuming that $u(x,t) \in C_p^\infty[0, 2\pi]$, a is a constant and $q \geq 0$ signifies the order of differentiation.

We seek solutions on the form

$$u_N(x,t) = \sum_{|n|\leq N/2} \tilde{u}_n(t)\exp(inx) = \sum_{j=0}^{N-1} u_N(x_j,t)g_j(x) \quad , \tag{5.3}$$

such that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - a\frac{\partial^q}{\partial x^q}u_N(x,t) \quad ,$$

vanishes at a specified set of grid points, y_j.

Let us first assume that $y_j = x_j$, i.e., the residual is required to vanish at the same grid points as the ones on which the approximation is based. Hence, we seek an N'th order polynomial, u_N, such that

$$\begin{aligned}\mathcal{I}_N R_N\big|_{x_j} &= \mathcal{I}_N\left[\frac{\partial u_N(x,t)}{\partial t} - a\frac{\partial^q}{\partial x^q}u_N(x,t)\right]\Bigg|_{x_j} \\ &= \left[\frac{\partial u_N(x,t)}{\partial t} - a\mathcal{I}_N\frac{\partial^q}{\partial x^q}\mathcal{I}_N u_N(x,t)\right]\Bigg|_{x_j} = 0 \quad .\end{aligned}$$

This results in N ordinary differential equations, describing the evolution of $u_N(x_j,t)$, to be solved at the grid points, x_j, on the form

$$\begin{aligned}\frac{du_N(x_j,t)}{dt} &= a\mathcal{I}_N\frac{\partial^q}{\partial x^q}\mathcal{I}_N u_N(x_j,t) \\ &= a\sum_{|n|\leq N/2}(in)^q\tilde{u}_n(t)\exp(inx_j) \\ &= a\sum_{i=0}^{N-1}\mathrm{D}_{ij}^{(q)}u_N(x_i,t) \quad ,\end{aligned}$$

where $\mathrm{D}^{(q)}$ represents the differentiation matrix discussed in Sec. 4.2. Consequently, the scheme consists of solving the ODE's at the grid points only. Note that two different formulations of the ODE's have been given, emphasizing the two computational different, but mathematically equivalent, methods of approximating the derivatives at the grid points.

Let us briefly also consider the case where we require that the residual vanishes at a set of grid points, y_j, which is different from x_j. In this case we recover N equations on the form

$$\frac{du_N(y_j,t)}{dt} = a\sum_{i=0}^{N-1} u_N(x_i,t)\left.\frac{d^q g_i}{dx^q}\right|_{y_j} \quad ,$$

describing the evolution of the unknowns $u_N(y_j,t)$ from which the unknowns, $u_N(x_j,t)$, in the original assumption ,Eq.(5.3), can be obtained by interpolation as

$$u_N(x_j,t) = \sum_{i=0}^{N-1} u_N(y_i,t)\tilde{g}_i(x_j) \quad .$$

Here $\tilde{g}_i(x)$ represents the Lagrange interpolation polynomial based on the grid points y_j.

While the formulation of the collocation method for linear problems is straightforward, it is by turning our attention to nonlinear problems that their advantages shine more brightly.

Example 18. Consider the nonlinear problem

$$\frac{\partial u(x,t)}{\partial t} = u(x,t)\frac{\partial u(x,t)}{\partial x} \quad ,$$

where the initial conditions are given through $g(x)$ and the solution is smooth and periodic in all derivatives, local in time.

Again, we seek a solution on the form

$$u_N(x,t) = \sum_{|n|\le N/2} \tilde{u}_n(t)\exp(inx) = \sum_{j=0}^{N-1} u_N(x_j,t)g_j(x) \quad ,$$

with the condition that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - u_N(x,t)\frac{\partial u_N(x,t)}{\partial x} \quad ,$$

vanishes at the grid points, x_j, as

$$\mathcal{I}_N R_N|_{x_j} = \mathcal{I}_N\left[\frac{\partial u_N(x,t)}{\partial t} - u_N(x,t)\frac{\partial u_N(x,t)}{\partial x}\right]\Bigg|_{x_j}$$

$$= \left[\frac{\partial u_N(x,t)}{\partial t} - \mathcal{I}_N \left(u_N(x,t) \frac{\partial u_N(x,t)}{\partial x} \right) \right] \Bigg|_{x_j} = 0 \ .$$

From this we recover N coupled ODE's

$$\begin{aligned} \frac{du_N(x_j,t)}{dt} &= u_N(x_j,t) \sum_{|n| \le N/2} in\tilde{u}_n(t) \exp(inx_j) \\ &= u_N(x_j,t) \sum_{i=0}^{N-1} \mathrm{D}_{ji} u_N(x_i,t) \ , \end{aligned}$$

which express the global solution to the Burgers equation through the grid point values, $u_N(x_j,t)$.

If we return to Ex. 15 for comparison, it is clear that whereas the Fourier-Galerkin method required the derivation of the complicated equations, there is little difference between formulating a Fourier-Collocation method for the solution of a linear problem or a scheme for a non-linear problem.

Let us finally consider a problem with a nonlinearity so strong that we found ourselves unable to formulate a Fourier-Galerkin method.

Example 19. Consider the strongly nonlinear problem

$$\frac{\partial u(x,t)}{\partial t} = \exp[u(x,t)] \frac{\partial u(x,t)}{\partial x} \ ,$$

where the initial conditions are given through $g(x)$ and the solution, $u(x,t)$, is assumed periodic and smooth, local in time.

We seek solutions on the form

$$u_N(x,t) = \sum_{|n| \le N/2} \tilde{u}_n(t) \exp(inx) = \sum_{j=0}^{N-1} u_N(x_j,t) g_j(x) \ ,$$

and require that the residual

$$R_N(x,t) = \frac{\partial u_N(x,t)}{\partial t} - \exp\left[u_N(x,t)\right] \frac{\partial u_N(x,t)}{\partial x} \ ,$$

vanishes at the grid points, x_j, on which the approximation is based. Thus, we constrain the solution such that

$$\mathcal{I}_N R_N|_{x_j} = \mathcal{I}_N \left[\frac{\partial u_N(x,t)}{\partial t} - \exp\left[u_N(x,t)\right] \frac{\partial u_N(x,t)}{\partial x} \right] \Bigg|_{x_j}$$
$$= \left[\frac{\partial u_N(x,t)}{\partial t} - \mathcal{I}_N \left(\exp\left[u_N(x,t)\right] \frac{\partial u_N(x,t)}{\partial x} \right) \right] \Bigg|_{x_j} = 0 \ ,$$

yielding N coupled ODE's to describe the evolution of the approximate solution as

$$\frac{du_N(x_j,t)}{dt} = \exp\left[u_N(x_j,t)\right] \sum_{|n| \le N/2} in\tilde{u}_n(t) \exp(inx_j)$$
$$= \exp\left[u_N(x_j,t)\right] \sum_{i=0}^{N-1} \mathrm{D}_{ji} u_N(x_i,t) \ ,$$

to be solved at the grid points, x_j.

Again, we find that the application of the Fourier-Collocation method is easy even for problems where the Fourier-Galerkin method fails. This is due to the fact that we can easily interpolate the nonlinear function, $F(u)$, in terms of the point values of $u(x)$, while it may be very hard, and in some cases impossible, to express the Fourier coefficients of $F(u)$ in terms of the expansion coefficients of $u(x)$.

One should keep in mind, however, that the ease of the formulation of the collocation method is obtained at the expense of the introduction of additional sources of error through aliasing and the approximation of the spatial derivatives. While the results of Sec. 4.3.2 suggests that this may be less of a concern in terms of the accuracy of the computed solution, these effects turn out to have a dramatic impact on the stability of the semi-discrete approximation and ultimately the fully discrete scheme when solving the partial differential equation.

5.2 Stability of Fourier Spectral Methods

Understanding the construction of Fourier spectral methods for various partial differential equations we are now ready to undertake the final part of the analysis of the schemes. We have previously convinced ourselves of the superior properties of the Fourier approximations, leaving us confident about the consistency of the schemes. However, the ques-

tion of stability of the schemes formulated in the last sections remains open.

To address this we shall split the analysis into two stages. On one hand, we shall discuss the stability of the semi-discrete approximation in which time, t, is kept as a continuous variable. This analysis relates directly to the stability discussed in the Equivalence Theorem in Sec. 3.2 and is based solely on an understanding of the properties of the spatial approximation of the operators. It is this analysis we shall undertake in this chapter.

The analysis of the fully discrete approximation, including also particular choices for the approximation of the temporal integration, is more involved and we postpone this discussion to Chapter 10.

5.2.1 Stability of the Fourier-Galerkin Method

The stability of the Fourier-Galerkin method is closely related to the wellposedness of the partial differential equation. Let us therefore first discuss some conditions ensuring wellposedness in the spirit of Sec. 3.1, i.e., in an energy sense.

We consider the one-dimensional initial boundary value problem

$$\frac{\partial u}{\partial t} = \mathcal{L}u \quad , \tag{5.4}$$

where $u(x,t) \in \mathsf{H}$ is assumed periodic at all times and proper initial data is supplied. We recall that H is a Hilbert space endowed with the inner product $(\cdot,\cdot)_{L^2[0,2\pi]}$ and the associated norm, $\|\cdot\|_{L^2[0,2\pi]}$.

A condition on wellposedness in the spirit of Def. 2 in Chap. ?? is

Theorem 14. *If the operator $\mathcal{L}$ is semi-bounded then the initial boundary value problem, Eq.(5.4), is wellposed in an energy sense as*

$$\frac{d}{dt}\|u\|^2_{L^2[0,2\pi]} \leq \alpha \|u\|^2_{L^2[0,2\pi]} \quad .$$

Proof: Multiply Eq.(5.4) with $\overline{u}$ to obtain

$$\overline{u}\frac{\partial u}{\partial t} = \overline{u}\mathcal{L}u \quad .$$

Likewise, consider the complex conjugate of Eq.(5.4) and multiply with u to obtain

$$u\overline{\frac{\partial u}{\partial t}} = u\overline{\mathcal{L}u} \ .$$

Adding the two expressions and integrating over $[0, 2\pi]$ yields

$$\begin{aligned}\frac{d}{dt}\|u\|^2_{L^2[0,2\pi]} &= (\mathcal{L}u, u)_{L^2[0,2\pi]} + (u, \mathcal{L}u)_{L^2[0,2\pi]} \\ &= (u, \mathcal{L}^* u)_{L^2[0,2\pi]} + (u, \mathcal{L}u)_{L^2[0,2\pi]} \\ &\leq \alpha \|u\|^2_{L^2[0,2\pi]} \ ,\end{aligned}$$

where the last results follows from semi-boundedness. QED

Example 20. Consider the operator

$$\mathcal{L} = a(x)\frac{\partial}{\partial x} \ ,$$

where $a(x) \in C^1_p[0, 2\pi]$ is real. In this case we derive the condition for semi-boundedness directly by computing the adjoint operator

$$\begin{aligned}(\mathcal{L}u, v)_{L^2[0,2\pi]} &= \int_0^{2\pi} a(x)\frac{\partial u}{\partial x}\overline{v}\, dx \\ &= -\int_0^{2\pi} u\frac{\partial}{\partial x}\left(\overline{a(x)v}\right)\, dx = \left(u, \left[-a(x)\frac{\partial}{\partial x} - \frac{da(x)}{dx}\right] v\right)_{L^2[0,2\pi]} \ ,\end{aligned}$$

by periodicity and integration by parts. Thus,

$$\mathcal{L}^* = -a(x)\frac{\partial}{\partial x} - \frac{da(x)}{dx} \ .$$

The condition on semi-boundedness is

$$\mathcal{L} + \mathcal{L}^* = -\frac{da(x)}{dx} \leq \alpha \quad \Rightarrow \quad \left|\frac{da(x)}{dx}\right| \leq A \ .$$

This result establishes wellposedness of the problems considered in Ex. 13 for $q = 1$ and Ex. 14.

The direct computation of the adjoint operator is in general quite complicated. However, it is often possible to establish conditions for en-

ergy boundedness directly and hence wellposedness in the sense of Def. 2 as illustrated in the following.

Example 21. Consider the operator

$$\mathcal{L} = \frac{\partial}{\partial x} b(x) \frac{\partial}{\partial x} \quad ,$$

where $b(x) \in C_p^1[0, 2\pi]$.

Proceeding as in the proof of Theorem 14, we recover

$$\frac{d}{dt} \|u\|^2_{L^2[0,2\pi]} = - \left(\frac{\partial u}{\partial x}, (b + \overline{b}) \frac{\partial u}{\partial x} \right)_{L^2[0,2\pi]} \quad .$$

If we require that

$$\left(\frac{\partial u}{\partial x}, (b + \overline{b}) \frac{\partial u}{\partial x} \right)_{L^2[0,2\pi]} \geq \sigma \left\| \frac{\partial u}{\partial x} \right\|^2_{L^2[0,2\pi]} \quad ,$$

with $\sigma > 0$, then the problem is clearly wellposed in an energy sense.

A problem obeying such a condition is termed strongly parabolic and, by inspection, we see that a necessary condition for strong parabolicity and, hence, wellposedness is that the real part of $b(x)$ is strictly positive.

Returning to the relation between wellposedness of the partial differential equation and the stability of the Fourier-Galerkin method, we can now state the result.

Theorem 15. *If the operator is semi-bounded, the Fourier-Galerkin scheme is stable.*

Proof: In the Fourier-Galerkin method we employ the expansion

$$u_N(x,t) = \sum_{|n| \leq N/2} \hat{u}_n(t) \exp(inx) \quad ,$$

such that $u_N(x,t) \in \hat{\mathsf{B}}_N$ and we define the residual, $R_N(x,t)$, as

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \mathcal{L} u_N \quad .$$

In general, we recall that

$$R_N(x,t) = \sum_{|n|\leq\infty} \hat{R}_n \exp(inx) \ ,$$

and $R_N(x,t)$ is generally not contained completely in $\hat{B}_N$, i.e., $\hat{R}_n \neq 0$ for $|n| > N/2$.

The Fourier-Galerkin scheme is obtained by requiring

$$\mathcal{P}_N R_N(x,t) = 0 \ \ \Rightarrow \ \ \forall\, |n| \leq \frac{N}{2} \ : \ \hat{R}_n(t) = 0 \ ,$$

which yields

$$\begin{aligned}
&\int_0^{2\pi} \mathcal{P}_N R_N(x,t)\overline{u_N(x,t)}\, dx \\
&= \int_0^{2\pi} \sum_{|n|>N/2} \hat{R}_n(t) \exp(inx) \sum_{|l|\leq N/2} \overline{\hat{u}_l(t)} \exp(-ilx)\, dx = 0 \ .
\end{aligned}$$

Thus, we have the identity

$$\int_0^{2\pi} \mathcal{P}_N \left(\frac{\partial u_N}{\partial t} - \mathcal{L} u_N \right) \overline{u_N}\, dx = 0 \ ,$$

since $\mathcal{P}_N R_N(x,t)$ is orthogonal to $u_N(x,t)$. Following the proof of Theorem 14 we recover

$$\begin{aligned}
\frac{d}{dt}\|u_N\|^2_{L^2[0,2\pi]} &= (u_N, \mathcal{P}_N \mathcal{L} u_N)_{L^2[0,2\pi]} + (\mathcal{P}_N \mathcal{L} u_N, u_N)_{L^2[0,2\pi]} \\
&= (u_N, \mathcal{P}_N \left[\mathcal{L} + \mathcal{L}^*\right] u_N)_{L^2[0,2\pi]} \\
&\leq \alpha \|u_N\|^2_{L^2[0,2\pi]} \ ,
\end{aligned}$$

provided the operator is semi-bounded.

Stability follows immediately as

$$\| \exp\left(\mathcal{L}_N t\right) \|_{L^2[0,2\pi]} \leq \exp\left(\tfrac{1}{2}\alpha t\right) \ .$$

where $\mathcal{L}_N = \mathcal{P}_N \mathcal{L} \mathcal{P}_N$. QED

A necessary and sufficient condition, providing a generalization of Theorem 14, for wellposedness can be stated on the following form [?]

Theorem 16. *Assume that there exists a self-adjoint operator, $\mathcal{H}$, and a constant, $K > 0$, such that*

$$K^{-1}\|u\|^2_{L^2[0,2\pi]} \leq (u, \mathcal{H}u)_{L^2[0,2\pi]} \leq K\|u\|^2_{L^2[0,2\pi]} \quad .$$

Then the initial value problem, Eq.(5.4), is wellposed if and only if there exists a constant, α, such that

$$(u, [\mathcal{H}\mathcal{L} + \mathcal{L}^*\mathcal{H}]\, u)_{L^2[0,2\pi]} \leq \alpha\, (u, \mathcal{H}u)_{L^2[0,2\pi]} \quad .$$

This allows for a generalization of the Theorem 15 as

Theorem 17. *If the operator is wellposed in the generalized sense of Theorem 16, then the Fourier-Galerkin scheme is stable.*

In the case of the Fourier-Galerkin methods it thus suffices to consider the issue of wellposedness as stability follows directly.

5.2.2 Stability of the Fourier-Collocation Method

While the issue of stability for the Fourier-Galerkin is determined entirely by the wellposedness of the partial differential equation this does not carry over to the Fourier-Collocation method.

For the stability theory, a key difference between the two methods is the requirement in the Galerkin method that the residual be orthogonal to the basis in which the approximate solution, $u_N(x,t)$ itself is expressed, while it is only required to vanish pointwise in the collocation scheme. This difference implies that wellposedness is reduced from being a sufficient to being a necessary condition for stability as there is no direct connection between the residual and space in which the solution lives.

Establishing stability of the pseudospectral basically proceeds along two different avenues, both centered around the use of energy methods. Thus, one strives to recover results on energy boundedness of the semi-discrete approximation, much as when wellposedness is considered. However, as we can not rely on the properties of the projection, we shall need to consider different techniques.

Let us begin by briefly discussing the discrete inner products and norms needed in the following as well as their relationship to the continuous inner products and norms. Consider first the discrete inner product

and the associated energy norm

$$[f_N, g_N]_N = \frac{2\pi}{N+1} \sum_{j=0}^{N} f_N(x_j) g_N(x_j) \quad , \|f_N\|_N^2 = [f_N, f_N]_N$$

where $f_N, g_N \in \hat{\mathsf{B}}_N$ and x_j represents the odd grid points. As a consequence of the accuracy of the quadrature, Theorem 7, we have

$$(f_N, g_N)_{L^2[0,2\pi]} = [f_N, g_N]_N \quad , \quad \|f_N\|_{L^2[0,2\pi]} = \|f_N\|_N \quad .$$

Hence, when basing the approximation on the odd number of grid points, the continuous and the discrete inner products and norms can be interchanged.

Returning to the case where the discrete inner product and the associated energy norm is

$$[f_N, g_N]_N = \frac{2\pi}{N} \sum_{j=0}^{N-1} f_N(x_j) g_N(x_j) \quad , \|f_N\|_N = [f_N, f_N]_N$$

where $f_N, g_N \in \tilde{\mathsf{B}}_N$ and x_j signifies the even grid points the situation is a bit more complex. This is a consequence of the quadrature rule, Theorem 5, being unable to exactly integrate polynomials of degree $2N$. Nevertheless, using the fact that $f_N \in L^2[0, 2\pi]$ one easily proves that there exists a $K > 0$ such that

$$K^{-1} \|f_N\|_{L^2[0,2\pi]}^2 \leq \|f_N\|_N^2 \leq K \|f_N\|_{L^2[0,2\pi]}^2 \quad . \tag{5.5}$$

Hence, the continuous and discrete norms are uniformly equivalent. To prove L^2-stability is it therefore suffices to prove stability in the discrete norms.

With this in mind, let us now attempt to derive bounds on the energy. As a first approach, we shall use the properties of the differentiation operators, i.e., they are all symmetric or skew-symmetric as discussed in Sec. 4.2.4. Alternatively, the quadrature rules introduced in Sec. 4.2 may, under certain circumstances, allow us to pass from the semi-discrete case with summations to the continuous case with integrals and, thus, simplify the subsequent analysis. Which one of these two techniques is most appropriate is problem dependent as we shall see in the following discussion. Unless otherwise stated we focus the attention on the Fourier-Collocation methods based on an even number of grid points, discussed

in detail in Sec. 4.2. For the collocation formulation we also generally take the grid points on which collocation scheme is based to similar to those on which the interpolating approximate solution is sought, i.e., $y_j = x_j$ in the context of Sec. 5.1.2.

5.2.2.1 Stability for Hyperbolic Problems

Let of consider the stability of the pseudospectral Fourier approximation to the hyperbolic problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + a(x)\frac{\partial u}{\partial x} = 0 \ , \\ &u(x,0) = g(x) \ , \end{aligned} \tag{5.6}$$

where $u(x) \in L^2[0,2\pi]$ and $g(x) \in L^2[0,2\pi]$ are assumed periodic. The coefficient, a, is assumed real for hyperbolicity while $|a_x|$ must be bounded to ensure wellposedness. For the purpose of illustration let us discuss the question of stability using two different versions of the energy method.

Theorem 18 (Method 1). *The pseudospectral Fourier approximation to the wellposed variable coefficient hyperbolic problem, Eq.(5.6), is stable as*

$$\|\exp\left[\mathrm{AD}t\right]\|_{L^2[0,2\pi]} \leq \frac{\max_x \sqrt{|a(x)|}}{\min_x \sqrt{|a(x)|}} \ ,$$

provided $a(x)$ is strictly bounded away from zero, i.e., $0 < |a(x)| < \infty$.

Proof: The Fourier-Collocation approximation to the variable coefficient hyperbolic problem, Eq.(5.6), is given as

$$\frac{d\boldsymbol{u}(t)}{dt} + \mathrm{AD}\boldsymbol{u}(t) = 0 \ , \tag{5.7}$$

provided we require that the equation is satisfied exactly at the grid points, x_j. Here x_j signifies the even grid points, $\boldsymbol{u}$ the associated grid vector, and A a diagonal matrix with entries, $\mathrm{A}_{jj} = a(x_j)$. Also, D represents the differentiation matrix, Eq.(4.19), and we recall that $\mathrm{D}^T = -\mathrm{D}$.

The direct solution to Eq. (5.7) is given as

$$\boldsymbol{u}(t) = \exp\left[-\mathrm{AD}t\right]\boldsymbol{u}(0) \ ,$$

and stability is guaranteed provided

$$\left\|\exp\left[-\mathrm{AD}t\right]\right\|_{L^2[0,2\pi]} \leq K(t) \ ,$$

which equals

$$\exp\left[-\mathrm{AD}t\right]\exp\left[\left(-\mathrm{AD}\right)^T t\right] \leq K^2(t) \ .$$

Consider first the case where $a(x_j) = a$ for which we obtain

$$\exp\left[-a\mathrm{D}t\right]\exp\left[\left(-a\mathrm{D}\right)^T t\right] = \exp\left[-a\left(\mathrm{D}+\mathrm{D}^T\right)t\right] = 1 \ ,$$

since D is skew-symmetric and commutes with itself.

Consider now the case of $a(x) > 0$ in which case AD no longer commutes with $-\mathrm{DA}$, causing the above procedure to fail. However, as $a(x) > 0$ we have

$$\mathrm{A} = \mathrm{A}^{1/2}\mathrm{A}^{1/2} \ .$$

Using the Taylor expansion of the matrix exponential one realizes that

$$\mathrm{A}^{-1/2}\exp\left[-\mathrm{AD}t\right]\mathrm{A}^{1/2} = \exp\left[-\mathrm{A}^{1/2}\mathrm{D}\mathrm{A}^{1/2}t\right] \ .$$

The key observation to make now is that

$$\left(\mathrm{A}^{1/2}\mathrm{D}\mathrm{A}^{1/2}\right)^T = \mathrm{A}^{1/2}\mathrm{D}^T\mathrm{A}^{1/2} = -\mathrm{A}^{1/2}\mathrm{D}\mathrm{A}^{1/2} \ ,$$

i.e., the new operator $\mathrm{A}^{1/2}\mathrm{D}\mathrm{A}^{1/2}$ is skew-symmetric. This implies

$$\begin{aligned}
\left\|\exp\left[-\mathrm{AD}t\right]\right\|_{L^2[0,2\pi]} &= \left\|\mathrm{A}^{1/2}\exp\left[-\mathrm{A}^{1/2}\mathrm{D}\mathrm{A}^{1/2}t\right]\mathrm{A}^{-1/2}\right\|_{L^2[0,2\pi]} \\
&\leq \left\|\mathrm{A}^{1/2}\right\|_{L^2[0,2\pi]}\left\|\exp\left[-\mathrm{A}^{1/2}\mathrm{D}\mathrm{A}^{1/2}t\right]\right\|_{L^2[0,2\pi]}\left\|\mathrm{A}^{-1/2}\right\|_{L^2[0,2\pi]} \\
&\leq \left\|\mathrm{A}^{1/2}\right\|_{L^2[0,2\pi]}\left\|\mathrm{A}^{-1/2}\right\|_{L^2[0,2\pi]} \\
&\leq \frac{\max_x\sqrt{a(x)}}{\min_x\sqrt{a(x)}} \ ,
\end{aligned}$$

hence establishing stability provided $0 < a(x) < \infty$.

The proof for $a(x) < 0$ is equivalent with the exception that we split A as

$$\mathrm{A} = -|\mathrm{A}|^{1/2}|\mathrm{A}|^{1/2} \ .$$

Since the differentiation matrix, $\tilde{\mathrm{D}}$, discussed in Sec. 4.2.4, for the odd method maintains the skew-symmetry, stability of this method follows from the above. QED

An alternative technique to establish stability may be understood by realizing that if $a(x)$ is uniformly bounded away from zero, A is non-singular and A^{-1} exists. Therefore, by multiplying from the left with $\boldsymbol{u}^T\mathrm{A}^{-1}$ we recover

$$\boldsymbol{u}^T\mathrm{A}^{-1}\frac{d\boldsymbol{u}}{dt} = \frac{1}{2}\frac{d}{dt}\boldsymbol{u}^T\mathrm{A}^{-1}\boldsymbol{u} = \boldsymbol{u}^T\mathrm{D}\boldsymbol{u} = 0 \ ,$$

as D is skew-symmetric and A is diagonal. This establishes stability in the weighted norm, $\boldsymbol{u}^T\mathrm{A}^{-1}\boldsymbol{u}$, known as the elliptic norm. However, this norm is clearly uniformly equivalent to the discrete energy norm since

$$\frac{1}{\max_x \sqrt{|a(x)|}}\boldsymbol{u}^T\boldsymbol{u} \le \boldsymbol{u}^T\mathrm{A}^{-1}\boldsymbol{u} \le \frac{1}{\min_x \sqrt{|a(x)|}}\boldsymbol{u}^T\boldsymbol{u} \ ,$$

hence completing the proof of stability.

Using the quadrature rules discussed in Sec. 4.2 to relate summations with integrations we can establish results equivalent to the above through a different route.

Theorem 19 (Method 2). *The pseudospectral Fourier approximation to the wellposed variable coefficient hyperbolic problem, Eq.(5.6), is stable as*

$$\frac{1}{2}\frac{d}{dt}\boldsymbol{u}^T\mathrm{A}^{-1}\boldsymbol{u} = 0 \ .$$

provided $a(x)$ is strictly bounded away from zero, i.e., $0 < |a(x)| < \infty$.

Proof: Seek a polynomial, $u_N \in \tilde{\mathsf{B}}_N$, that satisfy the equation

$$\left.\frac{\partial u_N}{\partial t}\right|_{x_j} + a(x_j)\left.\frac{\partial u_N}{\partial x}\right|_{x_j} = 0 \ , \tag{5.8}$$

where x_j represents the even grid points and we require that the equation be satisfied on these points also.

Assuming that $a(x)$ is uniformly bounded away from from zero ensures that $a(x)^{-1}$ exists. If we multiply Eq.(5.8) with $a(x_j)^{-1}u_N(x_j)$ and sum over all collocation points we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\sum_{j=0}^{N-1}\frac{1}{a(x_j)}u_N^2(x_j) &= -\sum_{j=0}^{N-1}u_N(x_j)\left.\frac{\partial u_N}{\partial x}\right|_{x_j} \\ &= -\frac{N}{2\pi}\int_0^{2\pi}u_N(x)\frac{\partial u_N}{\partial x}\,dx = 0 \ ,\end{aligned}$$

where we have used the quadrature rule given in Theorem 5 to establish stability along with the fact that u_N is periodic. The proof for an odd number of points can be completed in a similar fashion. QED

For the general case where $a(x)$ changes sign somewhere inside the computational domain, the situation is more complex than reflected in the above. The straightforward way to derive a stable pseudospectral Fourier approximation to Eq.(5.6) is to consider the equation on the skew-symmetric form

$$\frac{\partial u}{\partial t}+\frac{1}{2}a(x)\frac{\partial u}{\partial x}+\frac{1}{2}\frac{\partial a(x)u}{\partial x}-\frac{1}{2}a_x(x)u(x,t)=0 \ ,$$

and seek an approximation to this equation, the stability of which is stated in the following

Theorem 20. *The pseudospectral Fourier approximation to the well-posed variable coefficient hyperbolic equation, Eq.(5.6), expressed on skew-symmetric form is stable as*

$$\frac{1}{2}\frac{d}{dt}\|u_N\|_N^2 \le \tfrac{1}{2}\max_x|a_x(x)|\|u_N\|_N^2 \ ,$$

with $|a_x(x)|$ being bounded due to wellposedness.

Proof: To solve Eq.(5.6) on skew-symmetric form we seek a polynomial, $u_N(x,t)\in\tilde{\mathsf{B}}_N$, satisfying the equation

$$\left.\frac{\partial u_N}{\partial t}\right|_{x_j}+\frac{1}{2}a(x_j)\left.\frac{\partial u_N}{\partial x}\right|_{x_j}+\frac{1}{2}\left.\frac{\partial \mathcal{I}_N[a(x)u_N]}{\partial x}\right|_{x_j}-\frac{1}{2}a_x(x_j)u_N(x_j)=0 \ ,$$

at all grid points, x_j. Multiply with $u_N(x_j)$ and sum over all the collocation points to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\sum_{j=0}^{N-1} u_N^2(x_j) = &-\frac{1}{2}\sum_{j=0}^{N-1} a(x_j)u_N(x_j)\frac{\partial u_N}{\partial x}\Big|_{x_j} \\ &-\frac{1}{2}\sum_{j=0}^{N-1} u_N(x_j)\,\frac{\partial \mathcal{I}_N[a(x)u_N]}{\partial x}\Big|_{x_j} \\ &+\frac{1}{2}\sum_{j=0}^{N-1} a_x(x_j)u_N^2(x_j) \quad .\end{aligned}$$

Considering the second term we observe that $u_N \in \tilde{\mathsf{B}}_N$ and $\mathcal{I}_N\partial\mathcal{I}_N[a(x)u_N(x)]/\partial x \in \hat{\mathsf{B}}_{N-1}$, i.e., the quadrature rule in Theorem 5 is exact and we have

$$\begin{aligned}\frac{1}{2}\sum_{j=0}^{N-1} u_N(x_j)\,\frac{\partial \mathcal{I}_N[a(x)u_N(x)]}{\partial x}\Big|_{x_j} &= \frac{N}{4\pi}\int_0^{2\pi} u_N(x,t)\mathcal{I}_N\frac{\partial \mathcal{I}_N[a(x)u_N(x)]}{\partial x}\,dx \\ &= -\frac{N}{4\pi}\int_0^{2\pi} \mathcal{I}_N[a(x)u_N(x)]\mathcal{I}_N\frac{\partial u_N(x)}{\partial x}\,dx \\ &= -\frac{1}{2}\sum_{j=0}^{N-1} a(x_j)u_N(x_j)\frac{\partial u_N(x)}{\partial x}\Big|_{x_j} \quad ,\end{aligned}$$

assuming only that $a(x)$ and $u_N(x,t)$ are periodic.

Hence, we recover

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u_N\|_N^2 &= \frac{1}{2}\sum_{j=0}^{N-1} a_x(x_j)u_N^2(x_j,t) \\ &\le \tfrac{1}{2}\max_x|a_x(x)|\sum_{j=0}^{N-1} u_N^2(x_j,t) \quad ,\end{aligned}$$

which guarantees stability provided Eq.(5.6) is wellposed. QED

While we may establish stability for the skew-symmetric formulation it is less interesting from a practical point of view as it is twice as expensive as solving the original problem

$$\left.\frac{\partial u_N}{\partial t}\right|_{x_j} + a(x_j)\left.\frac{\partial u_N}{\partial x}\right|_{x_j} = 0 \ .$$

The question of stability of the pseudospectral Fourier approximation to this problem, however, is shrouded in a number of subtleties and a complete answer has only recently been given [?], although partial results for special $a(x)$ has been known for some time [?, ?].

To come to a better understanding of what causes these difficulties, let us consider Eq.(5.6) on the following form

$$\frac{\partial u_N}{\partial t} + \mathcal{N}_1 u_N + \mathcal{N}_2 u_N + \mathcal{N}_3 u_N = 0 \ ,$$

where

$$\mathcal{N}_1 u_N = \frac{1}{2}\mathcal{J}_N\left(a(x)\frac{\partial u_N}{\partial x}\right) + \frac{1}{2}\frac{\partial \mathcal{J}_N a(x) u_N}{\partial x} \ ,$$

is the skew-symmetric form of the operator introduced above,

$$\mathcal{N}_2 u_N = \frac{1}{2}\mathcal{J}_N\left(a(x)\frac{\partial u_N}{\partial x}\right) - \frac{1}{2}\mathcal{J}_N\frac{\partial (a(x)u_N)}{\partial x} \ ,$$

and

$$\mathcal{N}_3 u_N = \frac{1}{2}\mathcal{J}_N\frac{\partial (a(x)u_N)}{\partial x} - \frac{1}{2}\frac{\partial \mathcal{J}_N a(x) u_N}{\partial x} \ .$$

We note that we are considering the scheme based on the odd number of points. This is done for simplicity only as it allows us to pass to the integrals without complications. However, the conclusions we reach remains valid also for the even case.

Let us now consider

$$\frac{1}{2}\frac{d}{dt}\|u_N\|^2_{L^2[0,2\pi]} = -[u_N, \mathcal{N}_1 u_N]_N - [u_N, \mathcal{N}_2 u_N]_N - [u_N, \mathcal{N}_3 u_N]_N \ .$$

As we have shown in Theorem 20, the contribution from the skew-symmetric form, $\mathcal{N}_1 u_N$, vanishes identically. Furthermore, we easily establish that

$$[u_N, \mathcal{N}_2 u_N]_N \leq \tfrac{1}{2}\max_x |a_x(x)| \|u_N\|^2_{L^2[0,2\pi]} \ ,$$

by differentiation once. Inspecting the last term

$$\mathcal{N}_3 u_N = \frac{1}{2}\mathcal{J}_N \frac{\partial(a(x)u_N)}{\partial x} - \frac{1}{2}\frac{\partial \mathcal{J}_N a(x) u_N}{\partial x} \quad ,$$

we see that this measures the error associated with the loss of commutation between differentiation and interpolation. Indeed, if the interpolation, $\mathcal{J}_N$, and differentiation would commute, as for the continuous expansion in the Fourier-Galerkin scheme, this last term would vanish identically.

In Sec. 4.2 we discussed how the loss of commutation is a consequence of the aliasing error, i.e., $\mathcal{N}_3 u_N$ is a direct measure of the effect of aliasing. To understand the impact of this term on the stability of the approximation, we use the bound

$$[u_N, \mathcal{N}_3 u_N]_N \leq C \left(\|u_N\|^2_{L^2[0,2\pi]} + \|\mathcal{N}_3 u_N\|^2_{L^2[0,2\pi]} \right) \quad .$$

As $\|\mathcal{N}_3 u_N\|_{L^2[0,2\pi]}$ can be bounded as

$$\|\mathcal{N}_3 u_N\|_{L^2[0,2\pi]} \leq C N^{1-p} \|u_N^{(p)}\|_{L^2[0,2\pi]} \quad ,$$

by Theorem **??**, with C depending on $a(x)$ and its derivatives, we recover

$$\frac{1}{2}\frac{d}{dt}\|u_N\|^2_{L^2[0,2\pi]} \leq C \left(\|u_N\|^2_{L^2[0,2\pi]} + N^{2-2p} \|u_N^{(p)}\|^2_{L^2[0,2\pi]} \right) \quad ,$$

indicating that as long as the solution, u_N, is sufficiently smooth one can expect the approximation to be stable. However, for poorly resolved problems one may experience a weakly unstable solution with the instability caused solely by aliasing.

These arguments are qualitative in nature and a detailed understanding of the stability has only been obtained recently [**?**]. The main result is stated as

Theorem 21. *The pseudospectral Fourier approximation to the well-posed variable coefficient hyperbolic problem, Eq.(5.6), is weakly unstable as*

$$\|u_N(t)\|_N \leq C(t) N \|u_N(0)\|_N \quad ,$$

where $C(t)$ depends on t but not on N.

This more rigorous analysis confirms that one can only hope for algebraic stability and that the source of this growth is aliasing. Hence, if the solution is well resolved at all relevant times, the errors incurred by aliasing are very small and linear amplification is insufficient to excite the weak instability and, thus, impact the solution.

For poorly resolved phenomena, however, the situation is different as the small scale information, which no longer is insignificant, is destroyed by aliasing and experiences an $\mathcal{O}(N)$ amplification. This spreads to the full spectrum and eventually destroys the accuracy of the solution. The growth is, however, only algebraic.

The difficulty with solving problems in which $a(x)$ changes sign is that such problems often develop very steep gradients in finite time. Take as an example Eq.(5.6) with

$$a(x) = -\sin(x) \quad ,$$

the solution of which is

$$u(x,t) = g\left[2\tan^{-1}\left(e^t \tan\tfrac{x}{2}\right)\right] \quad .$$

This develops a very steep gradient around $x = 0$ as t grows. Hence, even if the initial conditions, $g(x)$, is well resolved the solution will, if a fixed grid is used, appear as poorly resolved at a later time, aliasing will become significant and the weak instability be excited.

5.2.2.2 Entr'acte on a Nonlinear Problem.

The equivalence theorem discussed in Sec. 3.2 supplies the motivate for splitting the discussion of convergence of the semi-discrete approximations to linear problems into that of consistency and stability. However, for nonlinear problems this ceases to be meaningful and one must generally attempt to prove convergence directly.

Nevertheless, a few results in the spirit of the previous discussion can be established for certain nonlinear problems. As an example of this consider the problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \quad , \\ &u(x,0) = g(x) \quad , \end{aligned} \tag{5.9}$$

where we assume that $u(x,t)$ and $g(x)$ are periodic and smooth, local in

time. By expressing it on skew-symmetric form as

$$\frac{\partial u}{\partial t} + \frac{1}{3}u\frac{\partial u^2}{\partial x} + \frac{1}{3}u^2\frac{\partial u}{\partial x} = 0 \ , \tag{5.10}$$

one easily proves

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2[0,2\pi]} &= \int_0^{2\pi} u^2\frac{\partial u}{\partial x}\,dx = \frac{1}{3}\int_0^{2\pi}\left(u^2\frac{\partial u}{\partial x} + u\frac{\partial u^2}{\partial x}\right)dx \\ &= \frac{1}{3}\int_0^{2\pi}\frac{\partial u^3}{\partial x}\,dx = 0 \ ,\end{aligned}$$

where the last equality follows directly from periodicity. Assuming that that a unique solution exists, this establishes wellposedness.

For the Fourier-Collocation approximation of Eq.(5.10) we have the following result

Theorem 22. *The pseudospectral Fourier approximation of the well-posed nonlinear problem, Eq.(5.9), is stable as*

$$\frac{1}{2}\frac{d}{dt}\|u_N\|^2_N = 0 \ ,$$

if expressed on skew-symmetric form, Eq.(5.10).

Proof: Look for a polynomial, $\mathcal{I}_N u(x,t) = u_N(x,t) \in \tilde{\mathsf{B}}_N$, that satisfy Eq.(5.10), in the following way

$$\left.\frac{\partial u_N}{\partial t}\right|_{x_j} + \frac{1}{3}\sum_{i=0}^{N-1}\mathrm{D}_{ji}u_N^2(x_i) + \frac{1}{3}u_N(x_j)\sum_{i=0}^{N-1}\mathrm{D}_{ji}u_N(x_i) = 0 \ ,$$

i.e., we require that the equation be satisfied exactly at the even number of grid points, x_j, on which also the approximation is based. Multiply with $u_N(x_j,t)$ and sum over the grid points to obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\sum_{j=0}^{N-1}\|u_N\|^2_N \\ &= -\frac{1}{3}\sum_{j=0}^{N-1}u_N(x_j)\sum_{i=0}^{N-1}\mathrm{D}_{ji}u_N^2(x_i) - \frac{1}{3}\sum_{j=0}^{N-1}u_N^2(x_j)\sum_{i=0}^{N-1}\mathrm{D}_{ji}u_N(x_i)\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{3}\sum_{j=0}^{N-1}\sum_{i=0}^{N-1}\left[u_N(x_j)\mathrm{D}_{ji}u_N^2(x_i) + u_N^2(x_j)\mathrm{D}_{ji}u_N(x_i)\right] \\
&= -\frac{1}{3}\sum_{j=0}^{N-1}\sum_{i=0}^{N-1}\left[u_N(x_j)\mathrm{D}_{ji}u_N^2(x_i) - u_N^2(x_i)\mathrm{D}_{ij}u_N(x_j)\right] \\
&= 0 \ ,
\end{aligned}
$$

where the last reduction appears by using the skew-symmetry of the differentiation matrix. QED

This establishes stability of the Fourier-Collocation method and also shows that the semi-discrete approximation of the skew-symmetric form maintains the energy conserving property.

However, the resulting scheme is twice as computationally expensive as the simple approximation. While this may be unacceptable for the linear problem, the fact that the semi-discrete approximation, as the original partial differential equations, conserves energy may be crucial.

The nonlinear energy conserving operator considered in the above represents a simple example of a much larger class of problems known as conservation laws. We shall return to such problems and their approximation in more detail in Chapter 8.

5.2.2.3 Stability of Parabolic Equations.

Let us now consider the question of stability for strongly parabolic problems as

$$
\begin{aligned}
&\frac{\partial u}{\partial t} = b(x)\frac{\partial^2 u}{\partial x^2} \ , \\
&u(x,0) = g(x) \ ,
\end{aligned}
\tag{5.11}
$$

where $b(x) > 0$ for wellposedness and $u(x,t)$ as well as $g(x)$ are assumed to be periodic and smooth.

As for the discussion of stability for hyperbolic problems we shall approach the question of stability for Eq.(5.11) in two different ways, yielding the same result.

Theorem 23 (Method 1). *The Fourier-Collocation method approximation to the strongly parabolic problem, Eq.(5.11), is stable as*

$$\frac{1}{2}\frac{d}{dt}\|u_N\|_N^2 \leq 0 \ ,$$

provided $b(x)$ is strictly positive.

Proof: We require the equation to be satisfied on x_j, yielding the the Fourier-Collocation approximation to the parabolic problem as

$$\frac{d\boldsymbol{u}(t)}{dt} = \mathrm{B}\mathrm{D}^{(2)}\boldsymbol{u}(t) \ .$$

Here x_j represents the grid points, $\boldsymbol{u}$ the associated grid vector and B the diagonal positive matrix with entries $\boldsymbol{B}_{jj} = b(x_j)$. Furthermore, $\mathrm{D}^{(2)}$ represents the differentiation matrix of 2nd order as discussed in Sec. 4.2.4. We need to be careful, however, when defining this operator.

As discussed in Sec. 4.2.4, a consequence of using the even method is that $\mathrm{D}^{(2)} \neq \mathrm{D}\cdot\mathrm{D}$ where both D and $\mathrm{D}^{(2)}$ are obtained directly by differentiation of the Lagrange interpolation polynomial, Eq.(4.12). The key difference between the two formulations is that $\mathrm{D}^{(2)}\boldsymbol{u} \in \tilde{\mathsf{B}}_N$ while $(\mathrm{D}\cdot\mathrm{D})\boldsymbol{u} \in \hat{\mathsf{B}}_{N-1}$, i.e., the latter reduces the order of the polynomial. As we shall see shortly, we must choose the latter definition, i.e.,

$$\mathrm{D}^{(2)} \equiv \mathrm{D}\cdot\mathrm{D} \ ,$$

to ensure stability of the Fourier-Collocation scheme. We note that the discrepancy between the two formulations is a consequence of the restricted space, $\tilde{\mathsf{B}}_N$, associated with the even number of collocation points. The problem does not arise when using a method based on an odd number of grid points.

Following the resolution of this, we continue by multiplying with $\boldsymbol{u}^T\mathrm{B}^{-1}$ from the left to recover

$$\begin{aligned}\boldsymbol{u}^T\mathrm{B}^{-1}\frac{d}{dt}\boldsymbol{u} &= \boldsymbol{u}^T\mathrm{D}^{(2)}\boldsymbol{u} = \boldsymbol{u}^T\mathrm{D}\mathrm{D}\boldsymbol{u} \\ &= \left(\mathrm{D}^T\boldsymbol{u}\right)^T(\mathrm{D}\boldsymbol{u}) = -\left(\mathrm{D}\boldsymbol{u}\right)^T(\mathrm{D}\boldsymbol{u}) \leq 0 \ ,\end{aligned}$$

where we use that D is skew-symmetric. Since

$$\frac{1}{\max_x b(x)}\|u_N\|_N^2 \leq \boldsymbol{u}^T\mathrm{B}^{-1}\boldsymbol{u} \leq \frac{1}{\min_x b(x)}\|u_N\|_N^2 \ ,$$

the result follows. QED

Let us also recover the same result using the quadrature rules.

Theorem 24 (Method 2). *The Fourier-Collocation method approximation to the strongly parabolic problem, Eq.(5.11), is stable as*

$$\frac{1}{2}\frac{d}{dt}\|u_N\|_N^2 \leq 0 \quad ,$$

provided $b(x)$ is strictly positive.

Proof: We seek solutions on the form

$$u_N(x,t) = \sum_{j=0}^{N-1} u_N(x_j,t) g_j(x) \quad ,$$

and require the equation to be satisfied at x_j as

$$\left.\frac{\partial u_N(x,t)}{\partial t}\right|_{x_j} = b(x_j) \left.\frac{\partial^2 u_N(x,t)}{\partial x^2}\right|_{x_j} \quad .$$

Multiply with $b(x_j)^{-1} u_N(x_j,t)$ and sum over the collocation points to obtain

$$\frac{1}{2}\frac{d}{dt}\sum_{j=0}^{N-1}\frac{1}{b(x_j)}u_N^2(x_j,t) = \sum_{j=0}^{N-1} u_N(x_j,t) \left.\frac{\partial^2 u_N(x,t)}{\partial x^2}\right|_{x_j} \quad .$$

We realize that the summation on the right hand side is a polynomial of order $2N$ which is beyond the accuracy of the quadrature rule for the even number of points, Theorem 5, and we cannot immediately pass to the integral.

To circumvent this problem we define the second order derivative as

$$\mathcal{I}_N \frac{d}{dx}\mathcal{I}_N \frac{d}{dx}\mathcal{I}_N \quad ,$$

similar to the approach taken in Theorem 23. This ensures that

$$\frac{\partial^2 u_N(x,t)}{\partial x^2} = \mathcal{I}_N \frac{d}{dx}\mathcal{I}_N \frac{d}{dx} u_N(x,t) \in \hat{\mathsf{B}}_{N-1} \quad ,$$

since the quadrature rule is exact and we have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\sum_{j=0}^{N-1}\frac{1}{b(x_j)}u_N^2(x_j,t) &= \sum_{j=0}^{N-1} u_N(x_j,t)\left(\frac{\partial}{\partial x}\mathcal{I}_N\frac{\partial u_N}{\partial x}\bigg|_{x_j}\right) \\
&= \frac{N}{2\pi}\int_0^{2\pi} u_N(x,t)\mathcal{I}_N\left(\frac{\partial}{\partial x}\mathcal{I}_N\frac{\partial u_N}{\partial x}\right)dx \\
&= -\frac{N}{2\pi}\int_0^{2\pi}\mathcal{I}_N\frac{\partial u_N}{\partial x}\mathcal{I}_N\frac{\partial u_N}{\partial x}\,dx \\
&= -\sum_{j=0}^{N-1}\left(\frac{\partial u_N}{\partial x}\bigg|_{x_j}\right)^2 \leq 0 \quad ,
\end{aligned}$$

where we use partial integration, periodicity of $u_N(x,t)$, and

$$\mathcal{I}_N\frac{\partial u_N(x,t)}{\partial x}\bigg|_{x_j} = \frac{\partial u_N(x,t)}{\partial x}\bigg|_{x_j} \quad ,$$

as a property of the interpolation operator. Utilizing the uniform equivalence between the elliptic norm and the usual energy norm as discussed in the proof of Theorem 23 yields the result.

The situation for the odd number of points is simpler as the associated quadrature, Theorem 7, integrates polynomials of order $2N$ exactly and no special definition of $\mathrm{D}^{(2)}$ needs to be considered. QED

Exercises

1. Consider the problem

$$\frac{\partial u}{\partial t} = a(x)\frac{\partial^q u}{\partial x^q} \quad , \quad x \in [0, 2\pi] \quad ,$$

where $u(x,t) \in C_p^\infty[0,2\pi]$. Both $u(x,t)$ and $a(x)$ are general complex numbers.
Assuming that a unique $u(x,t)$ exixts, show what conditions one must place on $a(x)$ to ensure wellposedness. Note that the answer depends on q.

2. Consider the biharmonic problem

$$\frac{\partial u}{\partial t} = -\frac{\partial^4 u}{\partial x^4} \quad , \quad x \in [0, 2\pi] \quad ,$$

where $u(x,t) \in C_p^\infty[0,2\pi]$.
Assuming that $u(x,t)$ exists, show that the problem is wellposed.

3. (Continued) Derive a Fourier-Galerkin scheme and discuss its stability.

4. (Continued) Derive a Fourier-Collocation scheme and discuss its stability.

5. Consider the system of equations

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial v}{\partial x} \quad , \quad x \in [0, 2\pi] \quad , \\ \frac{\partial v}{\partial t} &= \frac{\partial u}{\partial x} \quad , \end{aligned}$$

where $u(x,t) \in C_p^\infty[0,2\pi]$ and $v(x,t) \in C_p^\infty[0,2\pi]$.
Derive a Fourier-Galerkin scheme and prove its stability.

6. (Continued) Derive a Fourier-Collocation scheme and show its stability.

7. Consider the variable coefficient problem

$$\frac{\partial u}{\partial t} = -x\frac{\partial u}{\partial x} \quad , \quad x \in [0, 2\pi] \quad ,$$

where we assume that $u(x,t) \in C_p^\infty$ and smooth initial conditions are given.
Assuming that a unique solution exists, prove that the problem is well-posed.

8. (Continued) Derive a Fourier-Galerkin scheme and discuss its accuracy and stability.

9. (Continued) Consider the problem on skew-symmetric form

$$\frac{\partial u}{\partial t} = \frac{1}{2}\left[-x\frac{\partial u}{\partial x} - \frac{\partial (x\,u)}{\partial x} + u\right] \quad , \quad x \in [0, 2\pi] \quad ,$$

and derive a stable Fourier-Collocation scheme.

10. Consider the nonlinear Schrödinger equation

$$i\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + u|u|^2 = 0 \quad ,$$

where $u(x,t) : [0, 2\pi] \times \mathsf{R}^+ \to \mathsf{C}$. Furthermore, assume that $u(x,t) \in C_p^\infty[0, 2\pi]$.
This equation arises in the modeling of pulse propagation in optical fibers and studies of deep water waves among other things.
Assuming that a unique solution exists, prove that the problem is well-posed.

11. (Continued) Propose a Fourier-Collocation scheme and prove that it is stable in an energy sense.

12. Consider the Korteweg-de Vries equation

$$\frac{\partial u}{\partial t} + (c_0 + c_1 u)\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 \quad ,$$

where $u(x,t) : [0, 2\pi] \times \mathsf{R}^+ \to \mathsf{R}$. Furthermore, assume that $u(x,t) \in C_p^\infty[0, 2\pi]$. Also, c_0 and c_1 are real constants.
This equation arises in the modeling shallow water waves.
Assuming that a unique solution exists, prove that the problem is well-posed.

13. (Continued) Derive a Fourier-Galekin method for the Korteweg-de Vries equation.

14. (Continued) Derive a Fourier-Collocation scheme for the Korteweg-de Vries equation and prove its stability.

6

Orthogonal Polynomials

The last few chapters have focused on a detailed development and analysis of spectral and pseudospectral methods using trigonometric polynomials. While we found such schemes to perform well and deliver highly accurate results for certain special classes of problems, the analysis also revealed that the exponential accuracy of the scheme is achieved only when the solution is periodic. Moreover, the periodicity is required of the solution as well as of all its derivatives. As illustrated in Chapter 4, lacking such higher order periodicity globally impacts the convergence rate of the Fourier series, often reducing the spatial accuracy of the spectral scheme to that of a finite difference scheme.

This need for periodicity naturally limits the application of Fourier methods and inhibits the accurate study of more general non-periodic problems, e.g., boundary layer phenomena in fluid dynamics, scattering and penetration problems in acoustics and electromagnetics, and elastic waves in materials.

We recall from the previous chapters that the requirement of periodicity is a consequence of the choice of basis functions, $\phi_n(x)$. This suggests that to overcome this restriction we should attempt to identify a different basis better suited for the approximation of solutions on finite domains. The central question, the answer of which we shall devote a considerable part of this Chapter, is whether it is indeed possible to identify such a basis, resulting in rapidly converging spectral expansions, independent of the boundary conditions.

Guided by past experiences, we focus the attention on polynomial expansions of the form

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n x^n \quad ,$$

which, as we shall see, turns out to be exactly what we seek, albeit often expressed in a different form. The underlying assumption is that $u(x)$ can be well approximated in the finite dimensional subspace of

$$\mathsf{B}_N = \text{span}\{x^n\}_{n=0}^{N} \quad ,$$

that satisfy the boundary conditions.

Prior to discussing the properties of the expansion itself, however, we shall need to look a bit deeper into the properties of the basis, including attention to completeness and orthogonality. Following this, we subsequently limit the attention to a very special class of polynomial, appearing as eigensolutions to Sturm-Liouville problems. The emphasis shall be on polynomial approximations of continuous functions, $u(x) \in C^0[a,b]$, where the interval $[a,b]$ can be bounded as well as unbounded. However, to simplify the exposure we focus on problems defined on a bounded interval, keeping in mind that similar results can be established also for problems defined on the semi-infinite interval, $[0,\infty[$, as well as the infinite interval, $]-\infty,\infty[$. Where special results are required we shall provide these but we shall generally omit the proofs.

6.1 Polynomial Approximation and Completeness

As remarked in Chapter 4, establishing completeness of a particular basis family is an issue of significant complexity. However, for the polynomial basis it turns out that completeness can be established assuming only that the trigonometric basis is L^2-complete. We have used this classic result in previous chapters also and a proof can be found in [??].

The first step towards a completeness proof for the polynomial basis involves a fundamental existence theorem for the polynomial approximation of a continuous functions on the interval

Theorem 25 (Weierstrass). *For any continuous function, $u(x) \in C^0[0,1]$ and an arbitrary $\varepsilon > 0$, there exists an N and a polynomial, $p_N(x) \in \mathsf{B}_N$, such that*

$$\|u(x) - p_N(x)\|_{L^\infty[0,1]} \leq \varepsilon \quad .$$

In other words, any continuous bounded function, $u(x)$, defined on $[0,1]$ can be uniformly approximated by an N'th order polynomial.

Proof: The Weierstrass approximation theorem is a theorem of existence. Consequently, we can prove it directly by constructing a pointwise convergent approximating polynomial.

Let us extend the function, $u(x) \in C^0[0,1]$, evenly to the interval $[-1,1]$ such that $u(x) \in C^0[-1,1]$ and $u(x) = u(-x)$. Let us also express $x = \cos\theta$ and introduce the function $v(\theta) = u(\cos\theta)$, where $v(\theta) \in C^0[-\pi,\pi]$ and $v(\theta) = v(-\theta)$. As the periodic extension of $v(\theta)$ itself is even, continuous and uniformly bounded, it can be expanded in a pointwise convergent cosine series as (cf. Theorem 2)

$$\left\| v(\theta) - \sum_{n=0}^{N} \hat{v}_n \cos n\theta \right\|_{L^\infty[-\pi,\pi]} \to 0 \quad \text{as} \quad N \to \infty \ ,$$

where $\hat{v}_n$ represents the cosine expansion coefficients, see Sec. 4.1. Introducing the substitution $\theta = \arccos x$ yields

$$\left\| u(x) - \sum_{n=0}^{N} \hat{v}_n \cos\left(n \arccos x\right) \right\|_{L^\infty[-1,1]} \to 0 \quad \text{as} \quad N \to \infty \ .$$

We complete the proof by showing that the basis function

$$\phi_n(x) = \cos\left(n \arccos x\right) \ ,$$

is a polynomial of order n. From the definition of $\phi_n(x)$, we have

$$\phi_0(x) = 1 \quad \text{and} \quad \phi_1(x) = x \ .$$

Using the identity

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta \ ,$$

we have a recurrence relation for $\phi_n(x)$ as

$$\phi_{n+1}(x) = 2x\phi_n(x) - \phi_{n-1}(x) \ .$$

Since ϕ_0 and ϕ_1 are polynomials, clearly $\phi_n(x)$ is a polynomial of order n. These polynomials, $\phi_n(x)$, are known as Chebyshev polynomials and play a central role in the context of spectral methods as we shall discuss

thoroughly later.

Establishing the existence of a pointwise convergent polynomial for the interval of $[-1,1]$ naturally covers the interval of $[0,1]$ also.

An alternative, and perhaps more classical, proof of the Weierstrass approximation theorem involves proving that the Bernstein polynomials

$$B_N(u,x) = \sum_{n=0}^{N} u\left(\frac{n}{N}\right) \binom{N}{n} x^n (1-x)^{N-n} \quad ,$$

converges uniformly to $u(x) \in C^0[0,1]$. The proof follows from the properties of the polynomials and the continuity of $u(x)$. A complete proof can be found in [??]. QED

Existence of the pointwise convergent polynomial sequence for any bounded continuous function on $[0,1]$ immediately yields convergence in any equivalent norm, including $L^2[0,1]$. Based on this, completeness of the polynomial basis in $L^2_w[0,1]$ can be established as

Theorem 26 (Completeness). *Any piecewise continuous function, $u(x) \in L^2_w[0,1]$, can be expanded in a polynomial series that is convergent in the mean*

$$\|u(x) - p_N(x)\|_{L^2_w[0,1]} \to 0 \quad as \quad N \to \infty \quad .$$

Hence, the polynomial basis in complete in $L^2_w[0,1]$.

Proof: We have already established $L^2[0,1]$-convergence provided only $u(x) \in C^0[0,1]$ using the unweighted inner product. However, since $w(x) \in L^1[0,1]$, and nonnegative, convergence in the weighted inner product follows directly from $L^2[0,1]$ convergence since

$$\|u\|^2_{L^2_w[0,1]} \leq \|u\|^2_{L^2[0,1]} \|w\|^2_{L^2[0,1]} \quad .$$

It thus suffices to discuss convergence for piecewise continuous functions in $L^2[0,1]$.

Let us introduce a function, $v(x) \in C^0[0,1]$, that is a continuous approximation to $u(x)$. We can always find a $v(x)$ such that

$$\|u(x) - v(x)\|_{L^2[0,1]} < \varepsilon \quad ,$$

for any $\varepsilon > 0$. Indeed, assume that $u(x)$ has a point of discontinuity at

x_0. We may construct $v(x)$ such that it equals $u(x)$ outside the interval of $[x_0-\delta, x_0+\delta]$. Inside this interval, we construct $v(x)$ by a line segment connecting the point of $(x_0-\delta, u(x_0-\delta))$ with $(x_0+\delta, u(x_0+\delta))$. From this it easily follows that

$$\lim_{\delta\to 0} \|u(x)-v(x)\|_{L^2[0,1]} = 0 \ , \tag{6.1}$$

provided only that $u(x) \in L^2[0,1]$. Recall also that $v(x) \in C^0[0,1]$, implying that there exists a pointwise convergent approximating polynomial, $p_N \in \mathsf{B}_N$, to $v(x)$.

Consider now the error

$$\begin{aligned}
\|u(x)-p_N(x)\|^2_{L^2[0,1]} &= \|(u(x)-v(x))+(v(x)-p_N(x))\|^2_{L^2[0,1]} \\
&= \|u(x)-v(x)\|^2_{L^2[0,1]} + \|v(x)-p_N(x)\|^2_{L^2[0,1]} \\
&\quad +2\int_0^1 (u(x)-v(x))\,(v(x)-p_N(x))\,dx \\
&\le \|u(x)-v(x)\|^2_{L^2[0,1]} + \|v(x)-p_N(x)\|^2_{L^2[0,1]} \\
&\quad +2\|u(x)-v(x)\|_{L^2[0,1]}\|v(x)-p_N(x)\|_{L^2[0,1]} \ .
\end{aligned}$$

Each term of this expression can be made arbitrarily small using Eq.(6.1) and Theorem 25, hence establishing convergence in the mean of the approximation to any piecewise continuous function, $u(x) \in L^2_w[0,1]$. QED

The result in the general bounded interval $[a,b]$ follows directly

Lemma 7. Any piecewise continuous function, $u(x) \in L^2_w[a,b]$, can be expanded in a polynomial series that is convergent in the mean

$$\|u(x)-p_N(x)\|_{L^2_w[a,b]} \to 0 \quad \text{as} \quad N\to\infty \ .$$

The combination of Theorem 26 and Lemma 7 supplies the proof of completeness for the polynomial basis on any finite interval.

Let us also consider the question of orthogonality of the polynomial basis

$$\psi_n(x) = x^n \ .$$

For simplicity and without loss of generality, we take the interval to be $[-1,1]$.

The monomial basis is clearly not orthogonal with respect to the unweighted inner product, $(\cdot,\cdot)_{L^2[-1,1]}$, since

$$(\psi_0,\psi_2)_{L^2[-1,1]} = \int_{-1}^{1} x^2\,dx = \frac{2}{3} \ .$$

However, as all element, $\psi_n(x)$, of the basis clearly are independent we may use a Gram-Schmidt orthogonalization to construct polynomials, $\phi_n(x)$, that are mutually orthogonal with respect to some specific weighted inner product. Indeed, choosing as an example the unweighted inner product we obtain the first basis element as

$$\phi_0(x) = \frac{\psi_0(x)}{\|\psi_0\|_{L^2[-1,1]}} = \frac{1}{\sqrt{2}} \ .$$

Likewise, the second basis element becomes

$$\tilde{\phi}_1(x) = \psi_1(x) - (\phi_0,\psi_1)_{L^2[-1,1]}\phi_0(x) \ , \ \phi_1(x) = \frac{\tilde{\phi}_1(x)}{\|\tilde{\phi}_1\|_{L^2[-1,1]}} = \sqrt{\frac{3}{2}}x \ ,$$

while the next is

$$\begin{aligned}
\tilde{\phi}_2(x) &= \psi_2(x) - (\phi_0,\psi_2)_{L^2[-1,1]}\phi_0(x) - (\phi_1,\psi_2)_{L^2[-1,1]}\phi_1(x) \ , \\
\phi_2(x) &= \frac{\tilde{\phi}_2(x)}{\|\tilde{\phi}_2\|_{L^2[-1,1]}} = \sqrt{\frac{5}{8}}(3x^2-1) \ ,
\end{aligned}$$

and so on. It is always possible to construct orthogonal polynomial basis families with respect to the general weighted inner product using the Gram-Schmidt orthogonalization.

In the above developments of the completeness result, Weierstrass approximation theorem plays a crucial role. However, as given in Theorem 25, it does not include polynomial approximation on the semi-infinite or infinite interval. It is, nevertheless, possible to state Weierstrass-like theorems also for these intervals provided certain bounds are put on the growth of the function, $u(x)$, being approximated as x approaches infinity. We shall only give the relevant theorems for reference as the proofs are rather involved and beyond the scope of the present text. [26]

Theorem 27. *Consider any continuous function, $u(x) \in C^0[0,\infty[$, for which there exists a constant, δ, such that*

$$\lim_{x\to\infty} u(x)\exp(-\delta x) \to 0 \quad .$$

Then for any $\varepsilon > 0$, there exists an N and a polynomial, $p_N(x) \in \mathsf{B}_N$, such that

$$\sup_{x\in[0,\infty[} |u(x) - p_N(x)|\exp(-\delta x) \leq \varepsilon \quad .$$

Theorem 28. *Consider any continuous function, $u(x) \in C^0]-\infty,\infty[$, for which there exists a constant, δ, such that*

$$\lim_{x\to\pm\infty} u(x)\exp(-\delta x^2) \to 0 \quad ,$$

Then for any $\varepsilon > 0$, there exists an N and a polynomial, $p_N(x) \in \mathsf{B}_N$, such that

$$\sup_{x\in]-\infty,\infty[} |u(x) - p_N(x)|\exp(-\delta x^2) \leq \varepsilon \quad .$$

6.2 Classical Orthogonal Polynomials

In the last section we established $L^2_w[a,b]$ completeness of the polynomial basis and realized that they can be expressed on orthogonal form. These two properties are important for analysis as well as for computational efficiency but they alone are not enough to establish that the polynomial basis is suitable for the construction of spectral methods. For this purpose we need to understand the rate of convergence of the polynomial approximation to functions defined on a finite interval and understand how the properties of the function impacts the convergence rate.

To address these central issues we shall find it convenient to recover the orthogonal polynomials through a different route. Indeed, we shall take a detour and consider the Sturm-Liouville problem and the eigensolutions to such problems.

6.2.1 Sturm-Liouville Eigensolutions.

Let us consider the general Sturm-Liouville problem

$$\mathcal{L}\phi(x) = -\frac{d}{dx}\left(p(x)\frac{d\phi(x)}{dx}\right) + q(x)\phi(x) = \lambda w(x)\phi(x) \quad , \tag{6.2}$$

subject to the boundary conditions

$$\begin{aligned} \alpha_-\phi(-1) + \beta_-\phi'(-1) = 0 \quad , \quad \alpha_-^2 + \beta_-^2 \neq 0 \quad , \\ \alpha_+\phi(1) + \beta_+\phi'(1) = 0 \quad , \quad \alpha_+^2 + \beta_+^2 \neq 0 \quad . \end{aligned} \tag{6.3}$$

Here and in the following we restrict the attention to the interval $[-1,1]$ for simplicity. In Eq. (6.2) we have the real functions, $p(x) \in C^1[-1,1]$ and strictly positive in $]-1,1[$, $q(x) \in C^0[-1,1]$ and non-negative and bounded, and the non-negative continuous weightfunction, $w(x) \in C^0[-1,1]$.

Assuming that $\alpha_-\beta_- \leq 0$ and $\alpha_+\beta_+ \geq 0$ one can show [?] that the solutions to the Sturm-Liouville eigenvalue problem are unique sets of eigenfunctions, $\phi_n(x)$, and eigenvalues, λ_n. The eigenfunctions, $\phi_n(x)$, form an $L^2[-1,1]$ complete basis while the nonnegative and unique eigenvalues form an unbounded sequence, $\lambda_0 < \lambda_1 < \lambda_2 \ldots$. Based on this, it is customary to order the eigensolutions in unique pairs as (λ_n, ϕ_n).

This leads to the first important observation

Theorem 29. *The eigensolutions to the Sturm-Liouville problem are mutually orthogonal in the* $\|\cdot\|_{L^2_w[-1,1]}$*-norm*

$$(\phi_n, \phi_m)_{L^2_w[-1,1]} = \int_{-1}^{1} \phi_n(x)\phi_m(x)w(x)\,dx = \gamma_n\delta_{nm} \quad ,$$

where $\gamma_n = (\phi_n, \phi_n)_{L^2_w[-1,1]}$.

Proof: Recall that the Sturm-Liouville operator, Eq.(6.2),

$$\mathcal{L}\phi_n(x) = \lambda_n w(x)\phi_n(x) \quad ,$$

is self-adjoint for solutions, $\phi_n(x)$, obeying the boundary conditions, Eq.(6.3). Multiplying with $\phi_m(x)$ and integrating over $[-1,1]$ yields

$$(\phi_m, \mathcal{L}\phi_n)_{L^2[-1,1]} = \lambda_n\,(\phi_m, \phi_n)_{L^2_w[-1,1]} \quad . \tag{6.4}$$

Interchanging indices we obtain an equivalent expression

$$(\phi_n, \mathcal{L}\phi_m)_{L^2[-1,1]} = \lambda_m\,(\phi_m, \phi_n)_{L^2_w[-1,1]} \quad . \tag{6.5}$$

However, as $\mathcal{L}$ is self-adjoint one recovers, by subtracting Eq.(6.5) from Eq.(6.4), that

$$(\lambda_n - \lambda_m)\,(\phi_m, \phi_n)_{L^2_w[-1,1]} = 0 \quad .$$

Since the eigenvalues, λ_n, are unique this implies

$$\int_{-1}^{1} \phi_n(x)\phi_m(x)w(x)\,dx = \gamma_n \delta_{nm} \quad ,$$

where $\gamma_n = (\phi_n, \phi_n)_{L^2_w[-1,1]}$. QED

Note that we have not yet identified the eigenfunctions as polynomials. Indeed, this may only be done for specific choices of the functions, $q(x)$, $p(x)$ and $w(x)$, as we shall discuss shortly. However, for the eigenfunctions to be useful for the expansion of general functions, they, i.e., $\phi_n(x)$, must form a complete family. The completeness of the basis set formed from the eigensolutions is a consequence of the corresponding eigenvalues, λ_n, being an unbounded sequence for n approaching infinity. The connection between completeness and the unbounded growth of λ_n may be understood heuristically by considering the Sturm-Liouville problem with $p(x) = q(x) = w(x) = 1$ subject to homogeneous boundary conditions as

$$\frac{d^2}{dx^2}\phi_n(x) + \lambda_n \phi_n(x) = 0 \quad .$$

The solution is given as

$$\phi_n(x) = A_n \sin\left(\sqrt{\lambda_n}\pi x\right) \quad .$$

Hence, $1/\sqrt{\lambda_n}$ takes the role of a typical spatial scale. Completeness requires that arbitrary functions be expandable in $\phi_n(x)$. Thus, λ_n must be an unbounded sequence in n to be able to catch arbitrarily small scales. Conversely, the fact that λ_n is unbounded for the eigenfunctions of the Sturm-Liouville problems also indicates that these eigenfunctions form a complete family. A rigorous proof of this can be found in [?].

Assuming that $u(x) \in L^2_w[-1,1]$, consider the expansion

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n \phi_n(x) \quad ,$$

with the truncated approximation

$$\mathcal{P}_N u(x) = \sum_{n=0}^{N} \hat{u}_n \phi_n(x) \quad .$$

The continuous expansion coefficients are obtained from orthogonality as

$$\hat{u}_n = \frac{1}{\gamma_n} (u, \phi_n)_{L^2_w[-1,1]} \quad , \quad \gamma_n = \|\phi_n\|^2_{L^2_w[-1,1]} \quad .$$

As discussed previously, convergence in the mean and orthogonality yields implies Parseval identity

$$\int_{-1}^{1} u^2(x) w(x)\, dx = (u,u)_{L^2_w[-1,1]} = \sum_{n=0}^{\infty} \gamma_n \hat{u}_n^2 \quad .$$

Hence, measured in $\|\cdot\|_{L^2_w[-1,1]}$ the truncation error is given as

$$\left\| u(x) - \sum_{n=0}^{N} \hat{u}_n \phi_n(x) \right\|^2_{L^2_w[-1,1]} = \left[\sum_{n=N+1}^{\infty} \gamma_n \hat{u}_n^2 \right] \quad .$$

Since γ_n are constants independent of $u(x)$ and bounded as n approaches infinity since $\phi_n(x) \in L^2_w[a,b]$, convergence will depend solely on the decay of the expansion coefficients, $\hat{u}_n$. This is a situation much as we found for expansions based on the trigonometric polynomials discussed previously.

The decay of the expansion coefficients, $\hat{u}_n$, can be estimated as

$$\begin{aligned}
\hat{u}_n &= \frac{1}{\gamma_n}(u,\phi_n)_{L^2_w[-1,1]} = \frac{1}{\gamma_n} \int_{-1}^{1} u(x)\phi_n(x) w(x)\, dx \\
&= \frac{1}{\gamma_n \lambda_n} \int_{-1}^{1} u(x) \mathcal{L}\phi_n(x)\, dx = \frac{1}{\gamma_n \lambda_n} \int_{-1}^{1} u\left[-(p\phi_n')' + q\phi_n\right] dx \\
&= \frac{1}{\gamma_n \lambda_n} \left[-up\phi_n'\right]_{-1}^{1} - \frac{1}{\gamma_n \lambda_n} \int_{-1}^{1} \left[-u' p \phi_n' + qu\phi_n\right] dx \\
&= \frac{1}{\gamma_n \lambda_n} \left[p\left(u'\phi_n - u\phi_n'\right)\right]_{-1}^{1} + \frac{1}{\gamma_n \lambda_n} \int_{-1}^{1} \left[\mathcal{L}u(x)\right] \phi_n(x)\, dx \\
&= \frac{1}{\gamma_n \lambda_n} \left[p\left(u'\phi_n - u\phi_n'\right)\right]_{-1}^{1} + \frac{1}{\gamma_n \lambda_n} \left(u_{(1)}, \phi_n\right)_{L^2_w[-1,1]} \quad , \qquad (6.6)
\end{aligned}$$

where we have introduced the symbol

$$u_{(m)}(x) = \frac{1}{w(x)}\mathcal{L}u_{(m-1)}(x) \quad ,$$

and $u_{(0)}(x) = u(x)$. This derivation is valid provided $u_{(1)}(x) \in L^2_w[-1,1]$, i.e., $u^{(2)}(x) \in L^2_w[-1,1]$ and $w(x)^{-1} \in L^1[-1,1]$.

The estimate of the expansion coefficients, $\hat{u}_n$, contains two terms. From the Cauchy-Schwarz inequality, the second term in Eq. (6.6) is bounded as

$$\begin{aligned}\frac{1}{\gamma_n\lambda_n}\left(u_{(1)},\phi_n\right)_{L^2_w[-1,1]} &\leq \frac{1}{\gamma_n\lambda_n}\left|\int_{-1}^{1} u_{(1)}(x)\phi_n(x)w(x)\,dx\right| \\ &\leq \frac{1}{\gamma_n\lambda_n}\left(\int_{-1}^{1} u^2_{(1)}(x)w(x)\,dx\right)^{1/2}\left(\int_{-1}^{1}\phi^2_n(x)w(x)\,dx\right)^{1/2} \\ &= \frac{1}{\lambda_n\sqrt{\gamma_n}}\|u_{(1)}\|_{L^2_w[-1,1]} \sim \mathcal{O}\left(\frac{1}{\lambda_n}\right) \quad ,\end{aligned}$$

since γ_n is bounded and strictly positive and $u_{(1)}(x) \in L^2_w[-1,1]$.

To proceed beyond this point it is convenient to split the treatment into two separate cases that leads to distinctly different results for the convergence rate.

6.2.1.1 The Regular Sturm-Liouville Problem.

Let us first consider the case where $p(x)$ and the weightfunction, $w(x)$, both are strictly positive, known as the regular Sturm-Liouville problem. Considering the result in Eq. (6.6) we realize that if $u(x)$ satisfies

$$\begin{aligned}u'(\pm1)\phi_n(\pm1) - \;& u(\pm1)\phi'_n(\pm1) = 0 \;\Rightarrow \\ & \alpha_\pm u(\pm1) + \beta_\pm u'(\pm1) = 0 \quad ,\end{aligned} \tag{6.7}$$

by using Eq.(6.3) then the expansion coefficients decay as

$$|\hat{u}_n| \simeq C\frac{1}{\lambda_n}\|u_{(1)}\|_{L^2_w[-1,1]} \quad .$$

Indeed, if the whole sequence of $u_{(l)}$, $l = 0,..,m$ satisfies the boundary conditions, Eq.(6.7) we obtain the spectral like decay as

$$|\hat{u}_n| \simeq C \frac{1}{(\lambda_n)^m} \|u_{(m)}\|_{L^2_w[-1,1]} \quad .$$

This situation is very similar to the case for Fourier series where we had to assume that $u(x)$ and its derivatives, $u^{(m)}(x)$, were periodic to ensure spectral decay of the expansion coefficients. Here the constraints are even harder to fulfill.

The eigensolutions to the regular Sturm-Liouville problem with homogeneous boundary conditions have the asymptotic behavior [**?**]

$$\lambda_n \simeq (n\pi)^2 \left(\int_{-1}^{1} \sqrt{\frac{w(x)}{p(x)}}\, dx \right)^{-2} \quad \text{as} \quad n \to \infty \ ,$$

and

$$\phi_n(x) \simeq A_n \sin \left[\sqrt{\lambda_n} \int_1^x \sqrt{\frac{w(x)}{p(x)}}\, dx \right] \quad \text{as} \quad n \to \infty \ .$$

Hence, if $u(x)$ does not satisfy the boundary conditions, Eq.(6.7), the convergence rate is dominated by the boundary term in Eq.(6.6) as

$$|\hat{u}_n| \propto \frac{1}{\sqrt{\lambda_n}} \propto \frac{1}{n} \quad .$$

If, on the other hand, $u(x)$ satisfies Eq.(6.7) we obtain

$$|\hat{u}_n| \simeq \frac{1}{n^2} \quad ,$$

as $u_{(1)}(x) \in L^2_w[-1,1]$.

This analysis allows us to conclude that we can only expect algebraic decay of the expansion coefficients the general function, $u(x)$, except in very special cases where the sequence $u_{(l)}$, $l = 0, .., m$ and therefore $u(x)$, satisfies a very special set of boundary conditions, Eq.(6.7). For general problems this is clearly the case.

6.2.1.2 The Singular Sturm-Liouville Problem.

Let us now consider the singular Sturm-Liouville problem by requiring that $p(\pm 1) = 0$, i.e., $p(x)$ vanishes at the boundary points but remains positive and continuous. This has the consequence that the boundary term in Eq.(6.6) vanishes provided only that $p(\pm 1)u'(\pm 1) = 0$. In this

case we immediately have

$$|\hat{u}_n| \simeq C\frac{1}{(\lambda_n)^m}\|u_{(m)}\|_{L^2_w} \ ,$$

for $u_{(m)}(x) \in L^2_w[a,b]$, implying that $u^{(2m)}(x) \in L^2_w[a,b]$. Consequently, in case the function $u(x) \in C^\infty[a,b]$ we recover spectral decay of the expansion coefficients, i.e., $|\hat{u}_n|$ decays faster than any algebraic order of λ_n. This result is valid independent of specific boundary conditions on $u(x)$ at the boundaries.

This suggests that the eigensolutions of the singular Sturm-Liouville problem are well suited for expanding arbitrary functions defined in the finite interval as the eigenfunctions form a complete, orthogonal basis family. Moreover, the convergence rate of the expansions depends solely on the regularity of the function being expanded and not on the boundary conditions of the basis or the behavior of the function at the boundary of the finite interval beyond the weak regularity constraint on $p(\pm 1)u'(\pm 1)$ mentioned above.

6.2.2 Jacobi Polynomials.

While the above discussion of the properties of the eigensolutions to the singular Sturm-Liouville problem strongly supports using these solutions to approximate arbitrary functions on the interval, we still need to actually identify these solutions as being polynomials of order n. This, on the other hand, depends on proper choices of the two functions, $p(x)$, $q(x)$ and the weightfunction, $w(x)$.

Let us first attend to the case where the interval is finite and seek polynomial solutions to the singular Sturm-Liouville problem defined on the interval $[-1,1]$. The definition of $p(x)$, $q(x)$, and the weightfunction, $w(x)$, leading to eigensolutions of polynomial form are

Theorem 30. *The eigensolutions, $\phi_n(x) \in L^2_w[-1,1]$, to the singular Sturm-Liouville problem are polynomials of order n if and only if the functions, $p(x)$, $q(x)$, and the weightfunction, $w(x)$, are given as*

$$p(x) = (1-x)^{\alpha+1}(1+x)^{\beta+1} \ , \quad w(x) = (1-x)^{\alpha}(1+x)^{\beta} \ , \quad q(x) = cw(x) \ ,$$

provided that $\alpha, \beta > -1$ and

$$\lambda_n = n(n+\alpha+\beta+1) - c \ .$$

Proof: We begin by assuming that the polynomial, $\phi_n(x) \in \mathsf{B}_N$, is a solution to Eq. (6.2)

$$\phi_n(x) = \frac{1}{\lambda_n w(x)} \mathcal{L}\phi_n(x) \quad .$$

Expanding the Sturm-Liouville operator, we obtain

$$\phi_n(x) = \frac{1}{\lambda_n} \left(-\frac{p'(x)}{w(x)} \phi_n'(x) - \frac{p(x)}{w(x)} \phi_n''(x) + \frac{q(x)}{w(x)} \phi_n(x) \right) \quad ,$$

which must hold for any n. By equating the orders of the polynomials on the two sides for $n = 0, 1, 2$, we recover that

$$\frac{p(x)}{w(x)} \in \mathsf{B}_2 \quad , \quad \frac{p'(x)}{w(x)} \in \mathsf{B}_1 \quad , \quad \frac{q(x)}{w(x)} \in \mathsf{B}_0 \quad ,$$

as the eigenvalue, λ_n, is a constant for fixed n. Thus, $q(x) = cw(x)$, while the only solution to the two remaining equations are

$$p(x) = c_1 (1-x)^{\alpha+1} (1+x)^{\beta+1} \quad ,$$

and

$$w(x) = c_2 (1-x)^{\alpha} (1+x)^{\beta} \quad ,$$

as $p(x)$ is required to vanish at the endpoints of the interval. Since the two constants, c_1 and c_2, act as normalization constants only we take them to be unity. However, we must ensure that $w(x) \in L^1[-1, 1]$ as it is used to define a norm. This is ensured provided

$$\int_{-1}^{1} (1-x)^{\alpha} (1+x)^{\beta} \, dx = 2^{\alpha+\beta+1} \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} \quad ,$$

is bounded, e.g., $\alpha, \beta > -1$.

To complete the proof we need to show that the particular choice of $p(x)$, $q(x)$, and $w(x)$ yields polynomial eigensolutions of the Sturm-Liouville problem with the eigenvalue

$$\lambda_n = n(n + \alpha + \beta + 1) - c \quad . \tag{6.8}$$

To see this, introduce $p(x)$ and $w(x)$ into the Sturm-Liouville equation, Eq.(6.2), to obtain

$$-(1-x^2)\phi_n''(x) + ((\alpha+\beta+2)x + \alpha + \beta)\phi_n'(x) = \lambda_n \phi_n(x) \quad . \tag{6.9}$$

We take $c = 0$ without loss of generality as it just reflects a shift of the eigenvalue. Assuming that

$$\phi_n(x) = \sum_{l=0}^{n} a_l x^l \quad ,$$

inserting this into Eq. (6.9) and equating the leading coefficient of the polynomial yields the unique result. QED

This result completes our search for orthogonal polynomials suitable for approximating piecewise continuous functions on the finite interval. Since the eigensolutions are polynomials we have all ready established completeness. Orthogonality follows from the association with eigensolutions to a Sturm-Liouville problem. This connection also ensures rapid convergence of the approximation to smooth solutions as the convergence rate depends solely on the regularity of the function being approximated.

The polynomial solutions to the singular Sturm-Liouville problem on the finite interval are known as Jacobi polynomials, $P_n^{(\alpha,\beta)}(x)$, and appear as solutions to Eq.(6.9). We note the unbounded growth of λ_n for increasing n as evidence of the completeness of the basis as discussed previously.

The Rodrigues' formula for the Jacobi polynomial is given as

$$(1-x)^\alpha (1+x)^\beta P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}\frac{d^n}{dx^n}\left[(1-x)^{\alpha+n}(1+x)^{\beta+n}\right] \quad , \tag{6.10}$$

for integer n. An explicit formula is given as

$$\begin{aligned} P_n^{(\alpha,\beta)}(x) &= \frac{1}{2^n}\sum_{k=0}^{n}\binom{n+\alpha}{k}\binom{n+\beta}{n-k}(x-1)^{n-k}(x+1)^k \qquad (6.11)\\ &= \frac{\Gamma(2n+\alpha+\beta+1)}{2^n n!\Gamma(n+\alpha+\beta+1)}\left[x^n + \frac{(\alpha-\beta)n}{2n+\alpha+\beta}x^{n-1} + \cdots\right] \quad , \end{aligned}$$

from which we can also directly recover the identity

$$\frac{d}{dx}P_n^{(\alpha,\beta)}(x) = \frac{1}{2}\left(n+\alpha+\beta+1\right)P_{n-1}^{(\alpha+1,\beta+1)}(x) \quad . \tag{6.12}$$

The Jacobi polynomial are normalized such that

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n} = \frac{\Gamma(n+\alpha+1)}{n!\Gamma(\alpha+1)} \quad . \tag{6.13}$$

An important consequence of the symmetry of the weight function, $w(x)$, and the orthogonality of the Jacobi polynomials is the symmetry relation

$$P_n^{(\alpha,\beta)}(x) = (-1)^n P_n^{(\beta,\alpha)}(-x) \tag{6.14}$$

i.e., the Jacobi polynomials are even and odd depending on the order of the polynomial.

The expansion of functions, $u(x) \in L_w^2[-1,1]$, using the Jacobi polynomial, $P_n^{(\alpha,\beta)}(x)$, takes the form

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n P_n^{(\alpha,\beta)}(x) \quad ,$$

where the continuous expansion coefficients, $\hat{u}_n$, are found through the weighted inner product as

$$\begin{aligned} \hat{u}_n &= \frac{1}{\gamma_n} \left(u, P_n^{(\alpha,\beta)} \right)_{L_w^2[-1,1]} \\ &= \frac{1}{\gamma_n} \int_{-1}^{1} u(x) P_n^{(\alpha,\beta)}(x) (1-x)^\alpha (1+x)^\beta \, dx \quad , \end{aligned} \tag{6.15}$$

with the normalizing constant, γ_n, being

$$\begin{aligned} \gamma_n &= \left\| P_n^{(\alpha,\beta)} \right\|_{L_w^2[-1,1]}^2 \\ &= \frac{2^{\alpha+\beta+1}}{(2n+\alpha+\beta+1)n!} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)} \quad . \end{aligned} \tag{6.16}$$

A central issue to address is how to actually evaluate the Jacobi polynomials at a given abcisse. Using Eq. (6.10) or Eq. (6.11) we obtain the two first members of the family as

$$P_0^{(\alpha,\beta)}(x) = 1 \quad , \quad P_1^{(\alpha,\beta)}(x) = \frac{1}{2}(\alpha+\beta+2)x + \frac{1}{2}(\alpha-\beta) \quad .$$

It is, however, cumbersome to derive the polynomials this way. The next result suggests an easier way.

Theorem 31. *All Jacobi polynomials, $P_n^{(\alpha,\beta)}(x)$, satisfy a three-term recurrence relation of the form*

$$xP_n^{(\alpha,\beta)}(x) = a_{n-1,n}^{(\alpha,\beta)} P_{n-1}^{(\alpha,\beta)}(x) + a_{n,n}^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) + a_{n+1,n}^{(\alpha,\beta)} P_{n+1}^{(\alpha,\beta)}(x) \ ,$$

where $a^{(\alpha,\beta)}$ depends only on α, β and n.

Proof: Let us first consider the case of $n = 1$ for which the recurrence relation reduces to

$$xP_1^{(\alpha,\beta)}(x) = a_{0,1}^{(\alpha,\beta)} P_0^{(\alpha,\beta)}(x) + a_{1,1}^{(\alpha,\beta)} P_1^{(\alpha,\beta)}(x) + a_{2,1}^{(\alpha,\beta)} P_2^{(\alpha,\beta)}(x) \ .$$

Since $P_n^{(\alpha,\beta)}(x)$ is a polynomial, this amounts to finding three constants in order to match the coefficients of a second order polynomial. This can clearly always be done.

Consider now the case of $n > 1$. Clearly, we may always choose $a_{n+1,n}^{(\alpha,\beta)}$ such that

$$p_n(x) = a_{n+1,n}^{(\alpha,\beta)} P_{n+1}^{(\alpha,\beta)}(x) - xP_n^{(\alpha,\beta)}(x) \in \mathsf{B}_n \ ,$$

i.e., $p_n(x)$ is a polynomial of at most order n. Using the orthogonality of the Jacobi polynomials, we obtain, by multiplying with $P_m^{(\alpha,\beta)}(x)$ for all $m \leq n-2$ and integrating over $[-1,1]$, the result

$$\begin{aligned}
&\left(p_n, P_m^{(\alpha,\beta)}\right)_{L_w^2[-1,1]} = \\
&a_{n+1,n}^{(\alpha,\beta)} \left(P_{n+1}^{(\alpha,\beta)}, P_m^{(\alpha,\beta)}\right)_{L_w^2[-1,1]} - \left(xP_n^{(\alpha,\beta)}, P_m^{(\alpha,\beta)}\right)_{L_w^2[-1,1]} = \\
&a_{n+1,n}^{(\alpha,\beta)} \left(P_{n+1}^{(\alpha,\beta)}, P_m^{(\alpha,\beta)}\right)_{L_w^2[-1,1]} - \left(P_n^{(\alpha,\beta)}, xP_m^{(\alpha,\beta)}\right)_{L_w^2[-1,1]} = 0 \ ,
\end{aligned}$$

where the last reduction follows from the observation that $xP_m^{(\alpha,\beta)}(x)$ is a polynomial of order $n-1$ at most and can always be normalized appropriately to become a Jacobi polynomial. Thus, since $p_n(x) \in \mathsf{B}_n$ but has no components from $n-2$ to 0 it must be expressible as a linear combination on the form

$$p_n(x) = -a_{n-1,n}^{(\alpha,\beta)} P_{n-1}^{(\alpha,\beta)}(x) - a_{n,n}^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) \ ,$$

and the result follows. QED

The derivation of the actual constants that enter into the three-term recurrence relation is straightforward, however lengthy, and is done by matching the leading coefficients of the polynomials on each side of Theorem 31, using the expression of the polynomial in Eq.(6.11). Following this procedure, we obtain for $n > 0$

$$\begin{aligned} a_{n-1,n}^{(\alpha,\beta)} &= \frac{2(n+\alpha)(n+\beta)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta)} \ , \\ a_{n,n}^{(\alpha,\beta)} &= -\frac{\alpha^2-\beta^2}{(2n+\alpha+\beta+2)(2n+\alpha+\beta)} \ , \\ a_{n+1,n}^{(\alpha,\beta)} &= \frac{2(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+2)(2n+\alpha+\beta+1)} \ , \end{aligned} \tag{6.17}$$

with the special case for $n = 0$ that $a_{-1,0}^{(\alpha,\beta)} = 0$. Using the recurrence relation, all Jacobi polynomials can be evaluated at any $x \in [-1, 1]$ and for any order of the polynomial.

Recurrence formulas relating Jacobi polynomials of different order can be established in several ways. Indeed, we shall find the following result very useful.

Theorem 32. *All Jacobi polynomials, $P_n^{(\alpha,\beta)}(x)$, satisfy a three-term recurrence relation of the form*

$$P_n^{(\alpha,\beta)}(x) = b_{n-1,n}^{(\alpha,\beta)} \frac{dP_{n-1}^{(\alpha,\beta)}(x)}{dx} + b_{n,n}^{(\alpha,\beta)} \frac{P_n^{(\alpha,\beta)}(x)}{dx} + b_{n+1,n}^{(\alpha,\beta)} \frac{P_{n+1}^{(\alpha,\beta)}(x)}{dx} \ ,$$

where $b^{(\alpha,\beta)}$ depends only on α, β and n.

Proof: We begin by taking a derivative of the three-term recurrence relation given in Theorem 31 to obtain

$$\begin{aligned} P_n^{(\alpha,\beta)}(x) + x\frac{d}{dx}P_n^{(\alpha,\beta)}(x) = a_{n-1,n}^{(\alpha,\beta)} \frac{d}{dx}P_{n-1}^{(\alpha,\beta)}(x) + \\ a_{n,n}^{(\alpha,\beta)} \frac{d}{dx}P_n^{(\alpha,\beta)}(x) + a_{n+1,n}^{(\alpha,\beta)} \frac{d}{dx}P_{n+1}^{(\alpha,\beta)}(x) \ . \end{aligned}$$

Equation Eq.(6.12) guarantees that also the derivative of $P_n^{(\alpha,\beta)}(x)$ is a Jacobi polynomial. Hence, recalling Theorem 31, the derivative also satisfies a recurrence relation on the form

$$x\frac{d}{dx}P_n^{(\alpha,\beta)}(x) = \tilde{a}_{n-1,n}^{(\alpha,\beta)}\frac{d}{dx}P_{n-1}^{(\alpha,\beta)}(x) + \tilde{a}_{n,n}^{(\alpha,\beta)}\frac{d}{dx}P_n^{(\alpha,\beta)}(x) + \tilde{a}_{n+1,n}^{(\alpha,\beta)}\frac{d}{dx}P_{n+1}^{(\alpha,\beta)}(x) \quad . \tag{6.18}$$

Combining these two recurrence relations yields

$$\begin{aligned} P_n^{(\alpha,\beta)}(x) = &\left(a_{n-1,n}^{(\alpha,\beta)} - \tilde{a}_{n-1,n}^{(\alpha,\beta)}\right)\frac{d}{dx}P_{n-1}^{(\alpha,\beta)}(x) + \\ &\left(a_{n,n}^{(\alpha,\beta)} - \tilde{a}_{n,n}^{(\alpha,\beta)}\right)\frac{d}{dx}P_n^{(\alpha,\beta)}(x) + \\ &\left(a_{n+1,n}^{(\alpha,\beta)} - \tilde{a}_{n+1,n}^{(\alpha,\beta)}\right)\frac{d}{dx}P_{n+1}^{(\alpha,\beta)}(x) \end{aligned}$$

hence establishing the existence of the new recurrence. QED

Combining Eq.(6.12) and Eq.(6.18) with the recurrence from Theorem 31 allows for expressing the coefficients in Theorem 32 as

$$\begin{aligned} b_{n-1,n}^{(\alpha,\beta)} &= -\frac{1}{n+\alpha+\beta}a_{n-1,n}^{(\alpha,\beta)} \quad , \\ b_{n,n}^{(\alpha,\beta)} &= -\frac{2}{\alpha+\beta}a_{n,n}^{(\alpha,\beta)} \quad , \\ b_{n+1,n}^{(\alpha,\beta)} &= \frac{1}{n+1}a_{n+1,n}^{(\alpha,\beta)} \quad , \end{aligned} \tag{6.19}$$

An important property of the Jacobi polynomials is stated as

Theorem 33 (Christoffel-Darboux). *For any Jacobi polynomial, $P_n^{(\alpha,\beta)}(x)$, we have*

$$\begin{aligned} \sum_{n=0}^{N} &\frac{1}{\gamma_n}P_n^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(y) \\ &= \frac{a_{N+1,N}^{(\alpha,\beta)}}{\gamma_N}\frac{P_{N+1}^{(\alpha,\beta)}(x)P_N^{(\alpha,\beta)}(y) - P_N^{(\alpha,\beta)}(x)P_{N+1}^{(\alpha,\beta)}(y)}{x-y} \quad , \end{aligned}$$

where

$$\frac{a_{N+1,N}^{(\alpha,\beta)}}{\gamma_N} = \frac{2^{-(\alpha+\beta)}}{2N+\alpha+\beta+2}\frac{\Gamma(N+2)\Gamma(N+\alpha+\beta+2)}{\Gamma(N+\alpha+1)\Gamma(N+\beta+1)} \quad ,$$

using Eq.(6.16) and Eq.(6.17).

Proof: Let us first, for convenience, orthonormalize the Jacobi polynomials and consider

$$\tilde{P}_n^{(\alpha,\beta)}(x) = \frac{1}{\sqrt{\gamma_n}} P_n^{(\alpha,\beta)}(x) \quad .$$

The recurrence relation in Theorem 31 takes the form

$$x\tilde{P}_n^{(\alpha,\beta)}(x) = \tilde{a}_{n-1,n}^{(\alpha,\beta)} \tilde{P}_{n-1}^{(\alpha,\beta)}(x) + \tilde{a}_{n,n}^{(\alpha,\beta)} \tilde{P}_n^{(\alpha,\beta)}(x) + \tilde{a}_{n+1,n}^{(\alpha,\beta)} \tilde{P}_{n+1}^{(\alpha,\beta)}(x) \quad ,$$

with the recurrence coefficients

$$\begin{aligned}
\tilde{a}_{n-1,n}^{(\alpha,\beta)} &= \sqrt{\frac{\gamma_{n-1}}{\gamma_n}} a_{n-1,n}^{(\alpha,\beta)} = \sqrt{\frac{n(n+2\alpha)}{(2n+2\alpha+1)(2n+2\alpha-1)}} \quad , \\
\tilde{a}_{n,n}^{(\alpha,\beta)} &= a_{n,n}^{(\alpha,\beta)} \quad , \\
\tilde{a}_{n+1,n}^{(\alpha,\beta)} &= \sqrt{\frac{\gamma_{n+1}}{\gamma_n}} a_{n+1,n}^{(\alpha,\beta)} = \sqrt{\frac{(n+1)(n+2\alpha+1)}{(2n+2\alpha+3)(2n+2\alpha+1)}} \quad .
\end{aligned}$$

A key observation to make is that $\tilde{a}_{n-1,n}^{(\alpha,\beta)} = \tilde{a}_{n,n-1}^{(\alpha,\beta)}$.

A direct application of this recurrence relation yields

$$\begin{aligned}
\tilde{a}_{N+1,N}^{(\alpha,\beta)} & \quad \frac{\tilde{P}_{N+1}^{(\alpha,\beta)}(x)\tilde{P}_N^{(\alpha,\beta)}(y) - \tilde{P}_N^{(\alpha,\beta)}(x)\tilde{P}_{N+1}^{(\alpha,\beta)}(y)}{x-y} \\
&= \tilde{P}_N(x)\tilde{P}_N(y) + \tilde{a}_{N-1,N}^{(\alpha,\beta)} \frac{\tilde{P}_N^{(\alpha,\beta)}(x)\tilde{P}_{N-1}^{(\alpha,\beta)}(y) - \tilde{P}_{N-1}^{(\alpha,\beta)}(x)\tilde{P}_N^{(\alpha,\beta)}(y)}{x-y} \\
&= \tilde{P}_N(x)\tilde{P}_N(y) + \tilde{a}_{N,N-1}^{(\alpha,\beta)} \frac{\tilde{P}_N^{(\alpha,\beta)}(x)\tilde{P}_{N-1}^{(\alpha,\beta)}(y) - \tilde{P}_{N-1}^{(\alpha,\beta)}(x)\tilde{P}_N^{(\alpha,\beta)}(y)}{x-y} \quad ,
\end{aligned}$$

where the last result follows from the symmetry of the recurrence coefficients. Clearly we can repeat the reduction to finally recover the sum of the N orthonormal Jacobi polynomials. The result is obtained by expressing the sum using the standard Jacobi polynomials. QED

A consequence of this result is the following

Theorem 34. *All Jacobi polynomials, $P_n^{(\alpha,\beta)}(x)$, satisfy a three-term*

recurrence relation of the form

$$(1-x^2)\frac{dP_n^{(\alpha,\beta)}(x)}{dx} = c_{n-1,n}^{(\alpha,\beta)}P_{n-1}^{(\alpha,\beta)}(x)+c_{n,n}^{(\alpha,\beta)}P_n^{(\alpha,\beta)}(x)+c_{n+1,n}^{(\alpha,\beta)}P_{n+1}^{(\alpha,\beta)}(x) \quad ,$$

where $c^{(\alpha,\beta)}$ *depends only on* α, β *and* n.

Proof: We shall just sketch the proof. Using the result of Theorem 33 for $x = 1$ we recover

$$\sum_{n=0}^{N} \frac{1}{\gamma_n}P_n^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(1) = \frac{a_{N+1,N}^{(\alpha,\beta)}}{\gamma_N}\frac{P_{N+1}^{(\alpha,\beta)}(1)P_N^{(\alpha,\beta)}(x) - P_N^{(\alpha,\beta)}(1)P_{N+1}^{(\alpha,\beta)}(x)}{1-x} \quad .$$

This N'th order polynomial must be orthogonal to a constant under the weight $w(x) = (1-x)^{\alpha+1}(1+x)^{\beta}$ and must therefore be proportional to $P_N^{(\alpha+1,\beta)}(x)$. Working out the constants yields

$$P_n^{(\alpha+1,\beta)}(x) = \frac{2}{2n+\alpha+\beta+2}\frac{(n+\alpha+1)P_n^{(\alpha,\beta)}(x) - (n+1)P_{n+1}^{(\alpha,\beta)}(x)}{1-x} \quad .$$

Equivalently, by taking $x = -1$ in Theorem 33, we recover

$$P_n^{(\alpha,\beta+1)}(x) = \frac{2}{2n+\alpha+\beta+2}\frac{(n+\beta+1)P_n^{(\alpha,\beta)}(x) + (n+1)P_{n+1}^{(\alpha,\beta)}(x)}{1+x} \quad .$$

The result follows by combining Eq.(6.12), the recurrence in Theorem 31 and the two above results. The details can be found in [?]. QED

Working out the details yields the coefficients for the recurrence of Theorem 34 as

$$\begin{aligned} c_{n-1,n}^{(\alpha,\beta)} &= \frac{2(n+\alpha)(n+\beta)(n+\alpha+\beta+1)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)} \quad , \\ c_{n,n}^{(\alpha,\beta)} &= \frac{2n(\alpha-\beta)(n+\alpha+\beta+1)}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)} \quad , \\ c_{n+1,n}^{(\alpha,\beta)} &= -\frac{2n(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)} \quad . \end{aligned} \tag{6.20}$$

Having discussed the general Jacobi polynomials it seems only natural to raise the question as to which specific polynomial family, among all the Jacobi polynomials, is best suited for the approximation of functions defined on the finite interval. As can only be expected, the answer to this question depends on how the error is measured.

6.2.2.1 Legendre Polynomials

Among the many possible measures, the most natural is perhaps the unweighted $L^2[-1,1]$. Let us therefore first identify of the particular expansion coefficients, $\hat{u}_n$, associated with the Jacobi polynomials that minimize the approximation error in $L^2[-1,1]$. Thus, we seek among all the Jacobi polynomials, $P^{(\alpha,\beta)}(x)$, that particular choice of α and β that satisfies

$$\min_{\hat{u}_n} \left\| u(x) - \sum_{i=0}^{N} \hat{u}_n P_n^{(\alpha,\beta)}(x) \right\|_{L^2[-1,1]} .$$

The answer to this question is an immediate consequence of the properties of expansions of functions in Hilbert spaces.

Theorem 35. *Assume that $p_0(x), p_1(x), \ldots, p_N(x)$ represents a sequence of polynomials which are mutually orthogonal with respect the weighted inner product as*

$$(p_n, p_k)_{L_w^2[-1,1]} = \gamma_n \delta_{nk} \quad ,$$

where $\|p_n\|^2_{L_w^2[-1,1]} = \gamma_n$. For any $u(x) \in L_w^2[-1,1]$ we have

$$\left\| u(x) - \sum_{n=0}^{N} \frac{1}{\gamma_n}(u,p_n)_w p_n(x) \right\|_{L_w^2[-1,1]} \leq \left\| u(x) - \sum_{n=0}^{N} \hat{v}_n p_n(x) \right\|_{L_w^2[-1,1]} ,$$

for any choice of $\hat{v}_n$.

Proof: The proof follows directly by expressing the right hand side as

$$(u,u)_{L_w^2[-1,1]} - \sum_{n=0}^{N} \frac{1}{\gamma_n}(u,p_n)^2_{L_w^2[-1,1]} + \sum_{n=0}^{N} \gamma_n \left(\hat{v}_n - \frac{1}{\gamma_n}(u,p_n)_{L_w^2[-1,1]} \right)^2 .$$

Clearly, the two first terms have nothing to do with how the expansion coefficients, $\hat{v}_n$, are computed. Hence, to minimize the last term the

best choice is

$$\hat{v}_n = \frac{1}{\gamma_n} (u, p_n)_{L^2_w[-1,1]} \quad .$$

QED

Theorem 35 shows that the best approximating polynomial in $L^2[-1,1]$ is that particular Jacobi polynomial that is orthogonal in $L^2[-1,1]$, i.e., with the weightfunction $w(x) = 1$ as is recovered for $\alpha = \beta = 0$. These polynomials are known as Legendre polynomials, $P_n(x)$, and appear directly as eigensolutions to the Sturm-Liouville problem

$$\frac{d}{dx}(1-x^2)\frac{dP_n(x)}{dx} + n(n+1)P_n(x) = 0 \quad ,$$

obtained from Eq.(6.2) with $\lambda_n = n(n+1)$ and $p(x) = 1 - x^2$, $q(x) = 0$, and $w(x) = 1$.

The Legendre polynomials are related to the Jacobi polynomials, $P_n^{(0,0)}(x)$, as

$$P_n(x) = P_n^{(0,0)}(x) \quad ,$$

and the Rodrigues formula for the Legendre polynomials is derived from Eq.(6.10) as

$$P_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n} \left\{ (1-x^2)^n \right\} \quad . \tag{6.21}$$

An explicit formula can be recovered from Eq.(6.11) on the form

$$P_n(x) = \frac{1}{2^n} \sum_{k=0}^{[n/2]} (-1)^k \binom{n}{k} \binom{2n-2k}{n} x^{n-2k} \quad , \tag{6.22}$$

where $[n/2]$ refers to the integer part of the fraction.

The Legendre polynomials are bounded as [?]

$$|P_n(x)| \leq 1 \quad , \quad |P_n'(x)| \leq \frac{1}{2} n(n+1) \quad ,$$

with the boundary values being

$$P_n(\pm 1) = (\pm 1)^n \quad , \quad P_n'(\pm 1) = \frac{(\pm 1)^{n+1}}{2} n(n+1) \quad . \tag{6.23}$$

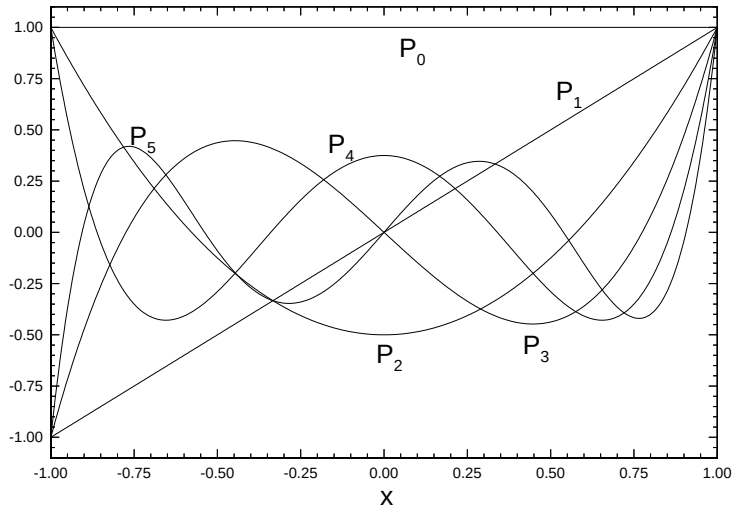

figure 6.1. Plot of the first 5 Legendre polynomials

Next we need to address the actual computation of the Legendre polynomials. Using Eqs. (6.21)-(6.22) we recover the first few Legendre polynomials as

$$P_0(x) = 1\ ,\quad P_1(x) = x\ ,\quad P_2(x) = \frac{1}{2}(3x^2 - 1)\ ,\quad P_3(x) = \frac{1}{2}(5x^3 - 3x)\ \ ,$$

and so on. The first 5 Legendre polynomials are shown in Fig. 6.1.

Note that the Legendre polynomials are identical up to a constant to the polynomials obtained by orthogonalization in $L^2[-1,1]$ of the monomial basis discussed in Section 6.1.

The three-term recurrence relation, Theorem 31, for the Legendre polynomials is

$$xP_n(x) = \frac{n}{2n+1}P_{n-1}(x) + \frac{n+1}{2n+1}P_{n+1}(x)\ \ , \tag{6.24}$$

using Eq.(6.17), yields a direct way of evaluate the polynomials to arbitrary order.

From Theorem 32 we recover the recurrence relation

$$P_n(x) = -\frac{1}{2n+1}P'_{n-1}(x) + \frac{1}{2n+1}P'_{n+1}(x)\ \ , \tag{6.25}$$

while Theorem 34 yields a recurrence relation on the form

$$(1-x^2)P_n'(x) = \frac{n(n+1)}{2n+1}P_{n-1}(x) - \frac{n(n+1)}{2n+1}P_{n+1}(x) \quad . \tag{6.26}$$

These and other properties of the Legendre polynomials are summarized in Appendix C.

6.2.2.2 Chebyshev Polynomials

Rather than considering the Legendre polynomial, we could seek to identify the polynomial family which minimizes the approximation error in $L^\infty[-1,1]$. This requires us to work a little harder but, as we shall soon see, the result is well worth the extra effort.

Let us first define the polynomial of best approximation

Definition 7. *Consider a function, $u(x) \in C^0[-1,1]$, and an approximating N'th order polynomial $p_N^*(x) \in \mathsf{B}_N$. The polynomial which minimizes the uniform approximation error*

$$\max_{x\in[-1,1]} |u(x) - p_N^*(x)| \quad ,$$

is termed the polynomial of best approximation.

While an explicit form of the polynomial of best approximation for an arbitrary smooth functions remains unknown, we can say something quite general about this special polynomial.

Theorem 36. *The polynomial of best approximation exists and is unique for any $u(x) \in C^0[-1,1]$.*

This is a classic result in approximation theory, guaranteeing existence and uniqueness of the polynomial of best approximation for C^0-functions on the finite interval. A proof can be found in, e.g., [**?**, **?**].

Theorem 37. *Assume that $u(x) \in C^0[-1,1]$ and that $p_N(x) \in \mathsf{B}_N$ is an N'th order polynomial. Introduce the error function*

$$e_N(x) = u(x) - p_N(x) \quad ,$$

and let

$$E_N = \max_{x\in[-1,1]} |e_N(x)| \quad .$$

If and only if $p_N(x)$ is the polynomial of best approximation are there $N+2$ points, $-1 \leq x_0 < \ldots < x_{N+1} \leq 1$, where $e_N(x)$ attains the value of E_N and with alternating signs, i.e. $e_N(x_i) = -e_N(x_{i+1})$.

This latter theorem is known as the Chebyshev Equioscillation Theorem and plays a key role in the theory of best approximation as it completely specifies the best approximating polynomial. The proof of this classic result can be found in, e.g. [**?**, **?**].

Assume now that we have an interpolating polynomial, $p_N(x) \in \mathsf{B}_N$, specified at $N+1$ points, x_j. Such a polynomial is clearly unique. Considering the error induced by using this polynomial for the approximation of a smooth function, $u(x) \in C^{N+1}[-1,1]$, then the Cauchy remainder is given as [**?**]

$$R_N(x) = u(x) - p_N(x) = \frac{\prod_{j=0}^{N}(x-x_j)}{(N+1)!} u^{(N+1)}(\xi) \quad ,$$

where ξ represents a value in the interior of the domain, $[-1,1]$. We can not in general estimate the value of $u^{(N+1)}(\xi)$ except in the special case where $u(x) = x^{N+1}$ in which case $u^{(N+1)}(\xi)$ is constant. In this special case, we have

$$R_N(x) = \prod_{j=0}^{N}(x-x_j) \in \mathsf{B}_{N+1} \quad ,$$

and we can attempt to specify x_j in order to minimize the remainder. In other words, we seek the polynomial, $R_N(x) \in \mathsf{B}_{N+1}$, which is the polynomial of best approximation to zero or, alternatively, the polynomial, $p_N(x) \in \mathsf{B}_N$, that uniformly minimizes the remainder

$$R_N(x) = x^{N+1} - p_N(x) \quad .$$

From Theorem 36 we know that such a polynomial of best approximation exists and is unique. Thus, following Theorem 37, we need only construct a polynomial of order $N+1$, i.e., the polynomial remainder, which attains its absolute maximum at $N+2$ points and with alternating sign.

Introducing the transformation $x = \cos\theta$, valid for $\theta \in [0,\pi]$, and

consider the function $v(x) = \cos(N+1)\theta$, it is clear that $v(x)$ attains its maximum absolute value of unity at exactly $N+2$ points given as

$$x_j = \cos\left(\frac{\pi}{N+1}j\right) \quad , \quad j = 0, \ldots, N+1 \quad .$$

The polynomial of best approximation is specified by the $N+1$ points at which the reminder vanishes, i.e., the roots of

$$v(x) = \cos\left((N+1)\arccos x\right) \quad ,$$

given as

$$x_j = \cos\left(\frac{\pi}{2}\frac{2j+1}{N+1}\right) \quad , \quad j = 0, \ldots, N \quad .$$

This is exactly the set, x_j, that minimizes the Cauchy remainder. Furthermore, if these grid points are used to construct $p_N(x)$ as an approximation to $u(x)$ it will have the least possible maximum error. In other words, $v(x)$, is the polynomial of best approximation to zero.

We proved in Theorem 25 that $v(x) \in \mathsf{B}_{N+1}$ in fact is a polynomial. Let us now introduce the definition

$$T_n(x) = \cos\left(n\arccos x\right) \quad ,$$

called the Chebyshev polynomials.

We still need, however, to study the general convergence and approximation properties of the Chebyshev polynomials. Fortunately, the following result comes to our rescue

Theorem 38. *The Chebyshev polynomial, $T_n(x)$, is a solution to the singular Sturm-Liouville problem with $p(x) = \sqrt{1-x^2}$, $q(x) = 0$ and the weightfunction, $w(x) = (\sqrt{1-x^2})^{-1}$, as*

$$\frac{d}{dx}\left(\sqrt{1-x^2}\frac{dT_n(x)}{dx}\right) + \frac{\lambda_n}{\sqrt{1-x^2}}T_n(x) = 0 \quad ,$$

with $\lambda_n = n^2$ and is related to the Jacobi polynomials with $\alpha = \beta = -\frac{1}{2}$ as

$$T_n(x) = \frac{(n!2^n)^2}{(2n)!}P_n^{(-\frac{1}{2},-\frac{1}{2})}(x) = \left[\binom{n-\frac{1}{2}}{n}\right]^{-1} P_n^{(-\frac{1}{2},-\frac{1}{2})}(x) \quad .$$

Proof: We first note that the singular Sturm-Liouville problem is equivalent to Eq.(6.2) with the parameters

$$\alpha = \beta = -\frac{1}{2} \quad .$$

Since we established in Theorem 30 that the only polynomial solutions to the singular Sturm-Liouville problem are the Jacobi polynomials, it suffices to prove that $T_n(x)$ satisfies the differential equation in the interior of $[-1,1]$ as we have all ready established that $T_n(x) \in \mathsf{B}_n$. That $T_n(x)$ satisfies the differential equation is derived trivially by introducing the substitution $x = \cos\theta$.

The normalization constant follows from the normalization chosen for the Jacobi polynomials as given in Eq.(6.13) by requiring that

$$T_n(\pm 1) = (\pm 1)^n \quad .$$

This concludes the proof. QED

Since the Chebyshev polynomials are nothing but a special member of the family of Jacobi polynomials we can immediately apply the theory for Jacobi polynomials, thereby assuring completeness and exponential convergence for the approximation of smooth functions on the interval.

The Rodrigues' formula for Chebyshev polynomials is obtained directly from Eq.(6.10) by normalizing appropriately as

$$T_n(x) = \frac{(-1)^n n! 2^n}{(2n)!}\sqrt{1-x^2}\frac{d^n}{dx^n}\left\{(1-x^2)^{n-\frac{1}{2}}\right\} \quad , \tag{6.27}$$

while the explicit equivalent of Eq.(6.11) becomes

$$T_n(x) = \frac{n}{2}\sum_{k=0}^{[n/2]}(-1)^k\frac{(n-k-1)!}{k!(n-2k)!}(2x)^{n-2k} = \cos\left(n\arccos x\right) \quad . \tag{6.28}$$

The definition of the Chebyshev polynomials yields the bounds

$$|T_n(x)| \leq 1 \quad , \quad |T_n'(x)| \leq n^2 \quad ,$$

with the boundary values

$$T_n(\pm 1) = (\pm 1)^n \quad , \quad T_n'(\pm 1) = (\pm 1)^{n+1} n^2 \quad , \tag{6.29}$$

due to the normalization employed.

Using Eqs. (6.27)-(6.28) we recover the first few Chebyshev polyno-

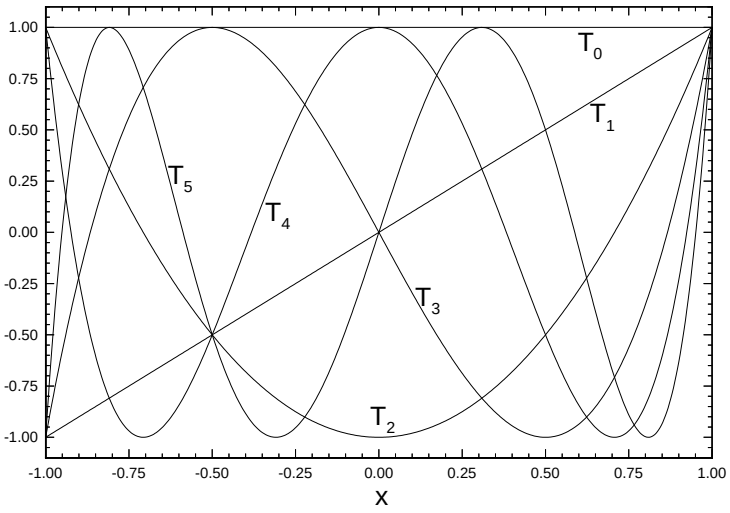

figure 6.2. Plot of the first 5 Chebyshev polynomials

mials on the form

$$T_0(x) = 1 \ , \ T_1(x) = x \ , \ T_2(x) = 2x^2 - 1 \ , \ T_3(x) = 4x^3 - 3x \ \ .$$

The first 5 Chebyshev polynomials are illustrated in Fig. 6.2.

For the Chebyshev polynomials, the three-term recurrence relation derived in Theorem 31 yields

$$xT_n(x) = \frac{1}{2}T_{n-1}(x) + \frac{1}{2}T_{n+1}(x) \ \ , \tag{6.30}$$

using the normalized recurrence coefficients in Eq.(6.17).

Likewise, we obtain the recurrence relation from Theorem 32 as

$$T_n(x) = -\frac{1}{2(n-1)}T'_{n-1}(x) + \frac{1}{2(n+1)}T'_{n+1}(x) \ \ , \tag{6.31}$$

while the recurrence of Theorem 34 yields a relation on the form

$$(1-x^2)T'_n(x) = \frac{n}{2}T_{n-1}(x) - \frac{n}{2}T_{n+1}(x) \ \ . \tag{6.32}$$

In Appendix C, we summarize the properties of Chebyshev polynomials in general, including several results not given here.

6.2.2.3 Ultraspherical Polynomials.

A special subclass of the Jacobi polynomials are known as the Ultraspherical polynomials, $P_n^{(\alpha)}(x)$, and are related to the Jacobi polynomials as

$$P_n^{(\alpha)}(x) = \frac{\Gamma(\alpha+1)\Gamma(n+2\alpha+1)}{\Gamma(2\alpha+1)\Gamma(n+\alpha+1)} P_n^{(\alpha,\alpha)}(x) \quad , \tag{6.33}$$

i.e., they represent the symmetric Jacobi polynomials normalized in a slightly different way. As realized in the two preceding sections, the symmetric Jacobi polynomials are particularly well suited for the approximation of functions on the interval and most of the remaining discussion of polynomial approximation of nonperiodic functions will be restricted to methods based on ultraspherical polynomials. Since the majority of spectral methods are based on either Chebyshev or Legendre polynomials, the development of the appropriate theory for the ultraspherical polynomials includes both cases. The relation between Legendre and Chebyshev polynomials and the ultraspherical polynomials are

$$P_n(x) = P_n^{(0)}(x) \quad , \quad T_n(x) = n \lim_{\alpha \to -\frac{1}{2}} \Gamma(2\alpha+1) P_n^{(\alpha)}(x) \quad , \tag{6.34}$$

where the limit is taken as $\Gamma(2\alpha+1)$ has a pole for $\alpha = -1/2$. A similar limit has to be taken in all subsequent formulas to recover the results for the Chebyshev polynomials. The ultraspherical polynomials, $P_n^{(\alpha)}(x)$, are also known as the Gegenbauer polynomials, $C_n^{(\lambda)}(x)$, on the form

$$C_n^{(\lambda)}(x) = P_n^{(\lambda-\frac{1}{2})}(x) \quad .$$

The ultraspherical polynomials, $P_n^{(\alpha)}(x)$, appear as the solution to the Sturm-Liouville problem

$$\frac{d}{dx}(1-x^2)^{\alpha+1}\frac{dP_n^{(\alpha)}(x)}{dx} + n(n+2\alpha+1)(1-x^2)^{\alpha} P_n^{(\alpha)}(x) = 0 \quad , \tag{6.35}$$

obtained directly from Eq.(6.2), with $\lambda_n = n(n+2\alpha+1)$ and $p(x) = (1-x^2)^{\alpha+1}$, $q(x) = 0$ and $w(x) = (1-x^2)^{\alpha}$.

The Rodrigues formula for the ultraspherical polynomials is obtained from Eq.(6.10) and Eq.(6.33) as

$$(1-x^2)P_n^{(\alpha)}(x) = \frac{\Gamma(\alpha+1)\Gamma(n+2\alpha+1)}{\Gamma(2\alpha+1)\Gamma(n+\alpha+1)} \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n}\left\{(1-x^2)^{n+\alpha}\right\} \quad , \tag{6.36}$$

while an explicit formula is recovered from Eq.(6.11) on the form

$$P_n^{(\alpha)}(x) = \frac{1}{\Gamma(\alpha+\frac{1}{2})}\sum_{k=0}^{[n/2]}(-1)^k\frac{\Gamma(\alpha+\frac{1}{2}+n-k)}{k!(n-2k)!}(2x)^{n-2k} \qquad (6.37)$$
$$= \frac{\Gamma(\alpha+1)\Gamma(2n+2\alpha+1)}{2^n n!\Gamma(2\alpha+1)\Gamma(n+\alpha+1)}\left[x^n+\cdots\right] \ ,$$

where $[n/2]$ refers to the integer part of the fraction.

The relation between different ultraspherical polynomials, Eq.(6.12), takes the form

$$\frac{d}{dx}P_n^{(\alpha)}(x) = (2\alpha+1)\,P_{n-1}^{(\alpha+1)}(x) \ , \qquad (6.38)$$

while the value of $P_n^{(\alpha)}(x)$ at the boundary is

$$P_n^{(\alpha)}(\pm 1) = (\pm 1)^n\begin{pmatrix} n+2\alpha \\ n \end{pmatrix} \ , \qquad (6.39)$$

from Eq.(6.13) and Eq.(6.33).

Using Eq.(6.36) or Eq.(6.37) we recover the first two polynomials as $P_0^{(\alpha)}(x)=1$ and $P_1^{(\alpha)}(x)=(2\alpha+1)x$ while the subsequent polynomials can be computed through the recurrence relation in Theorem 31 of the form

$$xP_n^{(\alpha)}(x) = a_{n-1,n}^{(\alpha)}P_{n-1}^{(\alpha)}(x) + a_{n+1,n}^{(\alpha)}P_{n+1}^{(\alpha)}(x) \ , \qquad (6.40)$$

where the recurrence coefficients, obtained by normalizing Eq.(6.17) appropriately, becomes

$$a_{n-1,n}^{(\alpha)} = \frac{n+2\alpha}{2n+2\alpha+1} \ , \quad a_{n+1,n}^{(\alpha)} = \frac{n+1}{2n+2\alpha+1} \ . \qquad (6.41)$$

The symmetry of the ultraspherical polynomials is emphasized by the relation

$$P_n^{(\alpha)}(x) = (-1)^n P_n^{(\alpha)}(-x) \ . \qquad (6.42)$$

The recurrence relation of the form given in Theorem 32 becomes

$$P_n^{(\alpha)}(x) = b_{n-1,n}^{(\alpha)}\frac{dP_{n-1}^{(\alpha)}(x)}{dx} + b_{n+1,n}^{(\alpha)}\frac{dP_{n+1}^{(\alpha)}(x)}{dx} \ , \qquad (6.43)$$

where the recurrence coefficients are obtained directly from Eq.(6.19) as

$$b_{n-1,n}^{(\alpha)} = -\frac{1}{2n+2\alpha+1} \quad , \quad b_{n+1,n}^{(\alpha)} = \frac{1}{2n+2\alpha+1} \quad . \tag{6.44}$$

Let us finally also given the result of Theorem 34 for the ultraspherical polynomials as

$$(1-x^2)\frac{dP_n^{(\alpha)}(x)}{dx} = c_{n-1,n}^{(\alpha)} P_{n-1}^{(\alpha)}(x) + c_{n+1,n}^{(\alpha)} P_{n+1}^{(\alpha)}(x) \quad , \tag{6.45}$$

with the coefficients being

$$c_{n-1,n}^{(\alpha)} = \frac{(n+2\alpha+1)(n+2\alpha)}{2n+2\alpha+1} \quad , \quad c_{n+1,n}^{(\alpha)} = -\frac{n(n+1)}{2n+2\alpha+1} \quad , \tag{6.46}$$

from Eq.(6.20).

In Chapter 2 we introduced the concept of points-per-wavelength required to accurately represent a wave, e.g., for the Fourier spectral method we found that we needed the minimum value of only two points to represent the wave exactly.

It is illustrative to consider this question also for polynomial expansions as it provides guidelines of how fine a grid, or how many modes, one needs to accurately represent a wave.

Example 22. Consider the plane wave

$$u(x) = \exp(ikx) \quad ,$$

and assume that approximate it using

$$u_N(x) = \sum_{n=0}^{N} \hat{u}_n P_n^{(\alpha)}(x) \quad .$$

One can show [?] that

$$\hat{u}_n = \frac{1}{2^{(\alpha+1)}} \frac{(2n+2\alpha+1)\Gamma(2\alpha+1)}{\Gamma(\alpha+1)} \frac{\sqrt{2\pi}}{\sqrt{k}k^{\alpha}} i^n J_{n+1/2+\alpha}(k) \quad ,$$

where $J_n(k)$ is the Bessel function of the first kind.

A reasonable of equivalent of points-per-wavelength is degrees-of-freedom per wavelength defined as

$$p = \frac{\lambda}{2/n} = \frac{n\lambda}{2} = \frac{\pi n}{k} \quad .$$

Assuming that n is large and using the asymptotic representation

$$\begin{aligned} J_{n+1/2+\alpha}(k) \simeq & \frac{1}{\sqrt{2\pi(n+1/2+\alpha)}} \left(\frac{e\pi}{2(n+1/2+\alpha)} \right)^{n+1/2+\alpha} \\ & \leq \frac{1}{\sqrt{2\pi n}} \left(\frac{e\pi}{2n} \right)^{n+1/2+\alpha} \quad , \end{aligned}$$

one obtains

$$|\hat{u}_n| \simeq \frac{1}{2^\alpha} \frac{\Gamma(2\alpha+1)}{\Gamma(\alpha+1)} \left(\frac{e}{2} \right)^{\alpha+1/2} \frac{1}{n^\alpha} \left(\frac{e\pi}{2p} \right)^n = C(\alpha) \frac{1}{n^\alpha} \left(\frac{e\pi}{2p} \right)^n \quad .$$

Assuming that α is held constant, a sufficient condition for exponential decay is

$$p > \frac{e\pi}{2} \simeq 4 \quad .$$

Thus, about 4 degrees-of-freedom, e.g., modes or grid points, per wavelength are required to experience the exponential convergence. While this is twice the number required for the Fourier case, we can now efficiently represent also nonperiodic functions with as little as 4 points/modes-per-wavelength which is still dramatically less than needed for the low order finite difference schemes discussed in Chapter 2.

6.2.3 Laguerre Polynomials.

In previous sections we have focused on the development of polynomials suitable for representing functions defined on a finite interval. An equivalent development can be undertaken for problems defined on the semi-infinite domain $x \in [0, \infty[$.

Rather than engaging in a thorough analysis, very similar in spirit to the one in the previous sections, we focus the attention on polynomial eigensolutions to the singular Sturm-Liouville problem defined on $[0, \infty[$ with $p(x) = x\exp(-x)$, $q(x) = 0$ and $w(x) = \exp(-x)$ as

$$\frac{d}{dx}x\exp(-x)\frac{dL_n(x)}{dx} + n\exp(-x)L_n(x) = 0 \quad ,$$

where $L_n(x)$, known as the Laguerre Polynomial, is defined on $[0,\infty[$. As $p(x)$ is singular at $x = 0$ we can expect exponential decay of the expansion in $L_n(x)$ independent of the boundary conditions of the approximated function. Note that the eigenvalue problem has polynomial solutions only for $\lambda_n = n$ as can be shown in a way similar to that in the proof of Theorem 30. The linear growth of λ_n is to be contrasted to the result for the Jacobi polynomials where $\lambda_n \sim \mathcal{O}(n^2)$, i.e., one could expect a significantly slower rate of convergence for expansions based on Laguerre polynomials.

The Rodrigues formula for the Laguerre polynomials is given as

$$\exp(-x)L_n(x) = \frac{1}{n!}\frac{d^n}{dx^n}\left\{x\exp(-x)\right\} \quad , \tag{6.47}$$

with an explicit expression of the form

$$L_n(x) = \sum_{k=0}^{n}\frac{(-1)^k}{k!}\begin{pmatrix} n \\ n-k \end{pmatrix}x^k \quad . \tag{6.48}$$

The polynomial expansion, based on the use of Laguerre polynomials, of $u(x) \in L^2_w[0,\infty]$ becomes

$$u(x) = \sum_{n=0}^{\infty}\hat{u}_n L_n(x) \quad ,$$

where the expansion coefficients are recovered through the inner product as

$$\hat{u}_n = \frac{1}{\gamma_n}\left(u, L_n\right)_{L^2_w[0,\infty]} = \int_0^\infty u(x)L_n(x)\exp(-x)\,dx \quad , \tag{6.49}$$

where $\gamma_n = 1$.

The Laguerre polynomials are normalized such that

$$L_n(0) = 1 \quad , \quad \frac{dL_n(0)}{dx} = -n \quad , \tag{6.50}$$

and, using Eq.(6.47) or Eq.(6.48), the first few polynomials can be computed as

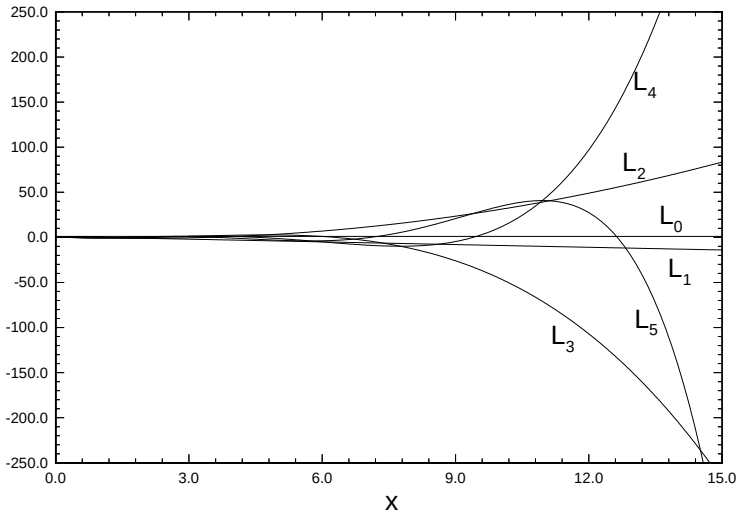

figure 6.3. Plot of the first 5 Laguerre polynomials

$$L_0(x) = 1 \ , \ L_1(x) = 1 - x \ , \ L_2(x) = \frac{1}{2}x^2 - 2x + 1 \ \ ,$$

and so on. For illustration, we plot in Fig. 6.3 the first few Laguerre polynomials.

The higher order Laguerre polynomials can be found using a recurrence like the one stated in Theorem 31 as

$$xL_n(x) = -nL_{n-1}(x) + (2n+1)L_n(x) - (n+1)L_{n+1}(x) \ \ . \qquad (6.51)$$

Also, one may obtain a relation of the form

$$L_n(x) = L'_n(x) - L'_{n+1}(x) \quad , \qquad (6.52)$$

similar to the expression in Theorem 32.

Example 23. Consider the plane wave

$$u(x) = \exp(ikx)\exp(-\delta x) \ \ , \ \ \delta \geq 0 \ \ ,$$

i.e., $\delta > 0$ implies a decay as x approaches infinity, for which we seek an approximation as

$$u_N(x) = \sum_{n=0}^{N} \hat{u}_n L_n(x) \quad .$$

One can show [?] that

$$\hat{u}_n = \left(\frac{\delta - ik}{1 + \delta - ik} \right)^n \quad .$$

Let us define degrees-of-freedom per wavelength as

$$p = \frac{\lambda}{x/n} \quad .$$

Again assuming that n is large, we have the asymptotic representation

$$L_n(x) \simeq \frac{1}{\sqrt{\pi}} \exp(x/2) \frac{1}{\sqrt[4]{nx}} \cos\left(2\sqrt{nx} - \frac{\pi}{4} \right) \quad ,$$

leading to

$$|\hat{u}_n| \leq \frac{1}{\sqrt{\pi}} \left(\frac{\delta^2 + k^2}{(1+\delta)^2 + k^2} \right)^{n/2} \exp\left(\frac{\pi n}{pk} \right) \left(\frac{2\pi n^2}{pk} \right)^{-1/4} \quad .$$

The requirement for exponential convergence clearly is

$$\left(\frac{(1+\delta)^2 + k^2}{\delta^2 + k^2} \right)^{n/2} \left(\frac{2\pi n^2}{pk} \right)^{1/4} > \exp\left(\frac{\pi n}{pk} \right) \quad .$$

Taking, as a first example, a non-decaying plane wave, i.e., $\delta = 0$, with the aim of resolving one wavelength, i.e., $k = np^{-1} = 1$, one recovers that $n = p > 6.4$, slightly higher than for the ultraspherical polynomials. Increasing the domain of interest, i.e., increasing np^{-1}, yields an increase in n also. However, the increase is linear, e.g., taking $np^{-1} = 2$ yields a requirement of $n = 2p > 14.4$, reflecting an almost constant value of p.

Inspecting the above condition for spectral convergence suggests that taking $\delta > 0$ would lower the requirement on p. The optimal value δ is seen to be

$$\delta = \frac{1}{2} \left(\sqrt{1 + 4k^2} - 1 \right) \quad .$$

Using this value, again with $k = np^{-1} = 1$, yields $n = p > 4.8$ which

is very close to the value for the ultraspherical polynomial. This result requires the function being approximation to decay approximately as $\exp(-x/2)$ which is reasonable considering the behavior of the polynomials.

In Appendix C, we summarize the properties of Laguerre polynomials in general, including several results not given here. We should also mention that an even more general family of Laguerre polynomials can derived from $p(x) = x^{\alpha+1}\exp(-x)$, $q(x) = 0$ and $w(x) = x^{\alpha}\exp(-x)$ in the Sturm-Liouville problem while the eigenvalue remains $\lambda_n = n$. The polynomials are known as generalized Laguerre polynomials, $L_n^{(\alpha)}(x)$, and have properties very similar to the special case of $\alpha = 0$ discussed above. However, as these generalized Laguerre polynomials are used even less for the construction of spectral methods for solving partial differential equations than the classical Laguerre polynomials we shall not discuss them further. A general introduction can be found in [26].

6.2.4 Hermite Polynomials.

Similar to the introduction of Laguerre polynomials for the polynomial approximations of functions defined on the semi-infinite interval, we may likewise seek polynomials suited for the approximation of functions defined on the infinite domain, $x \in]-\infty, \infty[$.

If we seek polynomial solutions to the non-singular Sturm-Liouville problem defined on the infinite domain, we recover these solutions for $p(x) = \exp(-x^2)$, $q(x) = 0$, and $w(x) = \exp(-x^2)$ as

$$\frac{d}{dx}\exp(-x^2)\frac{dH_n(x)}{dx} + 2n\exp(-x^2)H_n(x) = 0 \ ,$$

where the Hermite polynomial, $H_n(x)$, is defined for $x \in]-\infty, \infty[$. Note that even though the Hermite polynomials appears as solutions to a regular Sturm-Liouville problem, spectral convergence can be maintained as no boundary conditions need to be enforced or, rather, the function being approximated is assumed to vanish as x approaches infinity. As for the Laguerre polynomials we recover a linear growth in n of the associated eigenvalue since $\lambda_n = 2n$.

The Hermite polynomial has a Rodrigues formula of the form

$$\exp(-x^2)H_n(x) = (-1)^n \frac{d^n}{dx^n}\left\{\exp(-x^2)\right\} \quad , \tag{6.53}$$

with an explicit expression as

$$H_n(x) = n! \sum_{k=0}^{[n/2]} \frac{(-1)^k}{k!(n-2k)!}(2x)^{n-2k} \quad . \tag{6.54}$$

Polynomial approximation, using the Hermite polynomials, of functions $u(x) \in L^2_w[-\infty,\infty]$ is given as

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n H_n(x) \quad ,$$

where the expansion coefficients are given through the inner product as

$$\hat{u}_n = \frac{1}{\gamma_n}\,(u, H_n)_{L^2_w[-\infty,\infty]} = \frac{1}{\gamma_n}\int_{-\infty}^{\infty} u(x)H_n(x)\exp(-x^2)\,dx \quad , \tag{6.55}$$

where

$$\gamma_n = (H_n, H_n)_{L^2_w[-\infty,\infty]} = \sqrt{\pi}2^n n! \quad .$$

The Hermite polynomials are normalized such that

$$H_n(0) = (-1)^{n/2}\frac{n!}{(n/2)!} \quad ,$$

for n being even and $H_n(0) = 0$ for n odd.

Using Eq.(6.53) or Eq.(6.54), the first few polynomials can be expressed as

$$H_0(x) = 1 \ , \ H_1(x) = 2x \ , \ H_2(x) = 4x^2 - 2 \ , \ H_3(x) = 8x^3 - 12x \quad ,$$

and in Fig. 6.4 we show the first few Hermite polynomials.

The higher order Hermite polynomials are most easily obtained using the recurrence relation

$$xH_n(x) = nH_{n-1}(x) + \frac{1}{2}H_{n+1}(x) \quad . \tag{6.56}$$

Also, one may obtain a relation of the form

$$H_n(x) = \frac{1}{2(n+1)}H'_{n+1}(x) \quad , \tag{6.57}$$

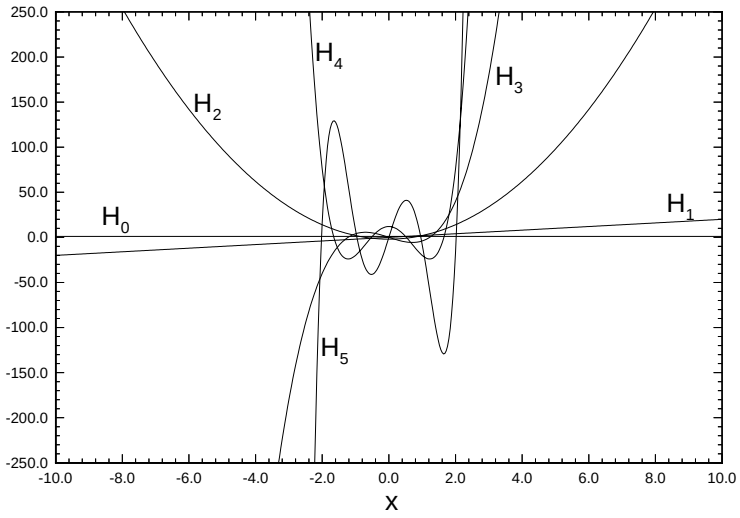

figure 6.4. Plot of the first 5 Hermite polynomials

similar to the expression appearing from Theorem 32. Further properties of the Hermite polynomials are summarized in Appendix C.

Example 24. Consider the plane wave

$$u(x) = \exp(ikx) \quad ,$$

and seek an approximation as

$$u_N(x) = \sum_{n=0}^{N} \hat{u}_n H_n(x) \quad .$$

One can show [**?**] that

$$\hat{u}_n = \frac{i^n k^n}{2^n n!} \exp\left(-\frac{k^2}{4}\right) \quad .$$

Let us again define degrees-of-freedom per wavelength as

$$p = \frac{\lambda}{2x/n} \quad .$$

For n being large, we have the asymptotic representation [?]

$$H_n(x) \simeq \frac{\Gamma(n+1)}{\Gamma(n/2+1)} \exp\left(\frac{x^2}{2}\right) \cos\left(\sqrt{(}2n+1)x - n\frac{\pi}{2}\right) \quad ,$$

leading to

$$|\hat{u}_n| \leq \frac{k^n}{2^n\Gamma(n/2+1)} \exp\left(-\frac{k^2}{4}\right) \exp\left(\frac{1}{2}\left(\frac{n\pi}{kp}\right)^2\right)$$

One easily finds the requirement for exponential convergence as

$$\frac{2^n(n/2)!}{k^n} \exp\left(\frac{k^2}{4}\right) > \exp\left(\frac{1}{2}\left(\frac{n\pi}{kp}\right)^2\right) \quad .$$

Consider, as a first example, that one aims at resolving one wave, i.e., $k = np^{-1} = 1$, one finds the requirement to to be $n = p > 6$, suggesting a behavior very similar to that of the previously considered polynomials.

One problem, however, with the Hermite expansion is exposed by considering the case where one wishes to consider a bigger domain, i.e., increasing np^{-1}. Taking $np^{-1} = 2$ yields $n = 2p > 15$ while $np^{-1} = 4$ results in $n = 4p > 44$. Thus, p is a function of the size of the domain. This makes the Hermite expansion less attractive as p becomes very large for large domains.

As for the Laguerre expansion, the situation improves when considering decaying waves. However, the improvement is only quantitative in lowering p for specific choices of k while it maintains the nonlinear growth of p with the size of the domain.

6.3 Approximation by Ultraspherical Polynomials

In previous sections we identified orthogonal and complete polynomial families suitable for the approximation of functions defined on finite and infinite domains. This sets the stage for the continued development of spectral methods based on these orthogonal polynomials.

In this section we focus on the approximation of functions defined on a bounded interval and we will, without loss of generality, restrict the attention to the interval $[-1, 1]$. Other intervals can be handled by a linear variable transformation as discussed in Lemma 7. Additionally,

we restrict the attention to series expansions based on the ultraspherical polynomials of which the Legendre and Chebyshev polynomials both appear as special cases. Only in very rare cases are spectral methods based on expansions of polynomials that are not ultraspherical. Hence, it is with little loss of generality that we restrict the attention to this specific class of polynomials rather than dealing. One should keep in mind, however, that spectrally accurate approximations to smooth problems can be formulated using the general Jacobi polynomials if a particular application suggest advantages of doing so.

The subsequent discussion includes the definition of continuous and discrete expansion coefficients, emphasizing the key issues related to the evaluation of derivatives of functions expressed in terms of the ultraspherical polynomials. Due to the importance of expansions in Legendre and Chebyshev polynomials, these cases will be given special attention during the development. A thorough discussion of the approximation theory for truncated expansions using ultraspherical polynomials is postponed to Sec. 6.6.1.

6.3.1 The Continuous Expansion.

Consider the continuous expansion of functions, $u(x) \in L_w^2[-1,1]$, in ultraspherical polynomials on the form

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n P_n^{(\alpha)}(x) \quad . \tag{6.58}$$

The expansion coefficients are found as

$$\hat{u}_n = \frac{1}{\gamma_n} \left(u, P_n^{(\alpha)} \right)_{L_w^2[-1,1]} = \frac{1}{\gamma_n} \int_{-1}^{1} u(x) P_n^{(\alpha)}(x) (1-x^2)^{\alpha} \, dx \quad , \tag{6.59}$$

where

$$\gamma_n = \left\| P_n^{(\alpha)} \right\|_{L_w^2[-1,1]}^2 = 2^{2\alpha+1} \frac{\Gamma^2(\alpha+1)\Gamma(n+2\alpha+1)}{n!(2n+2\alpha+1)\Gamma^2(2\alpha+1)} \quad , \tag{6.60}$$

If we assume that $u(x) \in L_w^2[-1,1]$ is expressed as in Eq.(6.58), we need to consider whether it is possible to recover the expansion coefficients, $\hat{u}_n^{(q)}$, for

$$\frac{d^q u(x)}{dx^q} = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} P_n^{(\alpha)}(x) \quad ,$$

i.e., given the expansion for $u(x)$ how does one recover the expansion for $u^{(q)}$. Solving partial differential equations, this is a key operation and it needs to be performed accurately and efficiently.

The following result establishes the required connection between the expansion coefficients.

Theorem 39. *Assume that the q'th derivative, $u^{(q)}(x) \in L^2_w[-1,1]$, is expanded in ultraspherical polynomials as*

$$u^{(q)}(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} P_n^{(\alpha)}(x) \quad .$$

Then the approximation of $u^{(q-1)}(x)$,

$$u^{(q-1)}(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q-1)} P_n^{(\alpha,\beta)}(x) \quad ,$$

can be recovered up to a constant through the relation ($n > 0$)

$$\hat{u}_n^{(q-1)} = b_{n,n-1}^{(\alpha)} \hat{u}_{n-1}^{(q)} + b_{n,n+1}^{(\alpha)} \hat{u}_{n+1}^{(q)} \quad ,$$

where $b_{n,n-1}^{(\alpha)}$ and $b_{n,n+1}^{(\alpha)}$ are defined in Eq.(6.44).

Proof: We establish the result by recovering the expansion coefficients, $\hat{u}_n$, of a function from the expansion coefficients, $\hat{u}_n^{(1)}$, of the derivative of the function up to a constant. Generalization to the general case follows directly.

Recall the recurrence relation, Eq.(6.43), and continue as

$$\begin{aligned}
\frac{d}{dx}u(x) &= \sum_{n=0}^{\infty} \hat{u}_n^{(1)} P_n^{(\alpha)}(x) \\
&= \sum_{n=0}^{\infty} \hat{u}_n^{(1)} \left(b_{n-1,n}^{(\alpha)} \frac{dP_{n-1}^{(\alpha)}(x)}{dx} + b_{n+1,n}^{(\alpha)} \frac{P_{n+1}^{(\alpha)}(x)}{dx} \right) \\
&= \sum_{m=-1}^{\infty} b_{m,m+1}^{(\alpha)} \hat{u}_{m+1}^{(1)} \frac{dP_m^{(\alpha)}(x)}{dx} + \sum_{m=1}^{\infty} b_{m,m-1}^{(\alpha)} \hat{u}_{m-1}^{(1)} \frac{dP_m^{(\alpha)}(x)}{dx} \\
&= \sum_{m=1}^{\infty} \left[b_{m,m-1}^{(\alpha)} \hat{u}_{m-1}^{(1)} + b_{m,m+1}^{(\alpha)} \hat{u}_{m+1}^{(1)} \right] \frac{dP_m^{(\alpha)}(x)}{dx}
\end{aligned}$$

$$= \sum_{n=0}^{\infty} \hat{u}_n \frac{dP_n^{(\alpha)}(x)}{dx} \quad ,$$

where we have used that $P_0^{(\alpha)}(x)$ is constant and defined that the polynomial $P_{-1}^{(\alpha)}(x)$ to be zero.

The equality clearly has to be valid for each polynomial as they are independent and the theorem follows.

Note that $\hat{u}_0$ remains undetermined as the operation essentially is an integration, hence leaving a constant undetermined. QED

Inverting the tridiagonal integration operator derived in Theorem 39 one obtains an operator to recover $\hat{u}_n^{(q)}$ from $\hat{u}_n^{(q-1)}$, i.e., it expresses the expansion coefficients for the differentiated function in terms of the expansion of the original function. The operator is given as ($\alpha \neq -1/2$)

$$\hat{u}_n^{(q)} = (2n + 2\alpha + 1) \sum_{\substack{p=n+1 \\ n+p \text{ odd}}}^{\infty} \hat{u}_p^{(q-1)} \quad , \tag{6.61}$$

and forms an upper triangular matrix with zeros along the diagonal and in the first column. In the case of a finite expansion, it also has zeros in the last row. Note that in contrast to the Fourier series, where differentiation corresponds to a local operation in spectral space, computing the polynomial expansion coefficients for the derivative generally involves all the expansion coefficients, $\hat{u}_n$. Hence, the computation of $\hat{u}_n^{(q)}$ from $\hat{u}_n^{(q-1)}$ using Eq.(6.61) is in general an $\mathcal{O}(N^2)$ operation.

Fortunately, Theorem 39 suggests a more efficient way to compute derivatives of functions expanded in ultraspherical polynomials if a finite expansion is used. Assume we have the expansion

$$\mathcal{P}_N \frac{d^{(q-1)}u(x)}{dx^{(q-1)}} = \sum_{n=0}^{N} \hat{u}_n^{(q-1)} P_n^{(\alpha)}(x) \quad ,$$

and we seek an approximation of the derivative as

$$\mathcal{P}_N \frac{d^{(q)}u(x)}{dx^{(q)}} = \sum_{n=0}^{N} \hat{u}_n^{(q)} P_n^{(\alpha)}(x) \quad .$$

Considering a finite expansion only, we realize that $\hat{u}_N^{(q)} = \hat{u}_{N+1}^{(q)} = 0$. Theorem 39 then suggests the backward recursion formula for $n \in [1, N]$

as

$$\hat{u}_{n-1}^{(q)} = (2n + 2\alpha - 1)\left[\frac{1}{2n + 2\alpha + 3}\hat{u}_{n+1}^{(q)} + \hat{u}_n^{(q-1)}\right] \quad . \tag{6.62}$$

Applying the backward recursion, Eq.(6.62), for the computation of $\hat{u}_n^{(q)}$, the computational workload is reduced to $\mathcal{O}(N)$.

Contrary to the case of a continuous Fourier series, truncation and differentiation does not in general commute, i.e.,

$$\mathcal{P}_{N-1}\frac{du}{dx} \neq \frac{d}{dx}\mathcal{P}_N u \quad .$$

This is a natural consequence of the global nature of the differentiation process in spectral space, Eq.(6.61). Hence, the exact way in which the infinite dimensional operator is terminated does make a difference and inhibits the commutation.

6.3.1.1 The Continuous Legendre Expansion.

The Legendre expansion of a function, $u(x) \in L^2[-1, 1]$, is given as

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n P_n(x) \quad , \tag{6.63}$$

with the expansion coefficients being

$$\hat{u}_n = \frac{1}{\gamma_n}(u, P_n)_{L^2[-1,1]} = \frac{2n+1}{2}\int_{-1}^{1} u(x)P_n(x)\,dx \quad ,$$

since $\gamma_n = 2/(2n + 1)$ using Eq.(6.16).

Connecting Eq. (6.25) to Theorem 39 provides a relation between the expansion coefficients, $\hat{u}_n^{(q-1)}$, and those of its derivative, $\hat{u}_n^{(q)}$, as

$$\hat{u}_n^{(q-1)} = \frac{1}{2n-1}\hat{u}_{n-1}^{(q)} - \frac{1}{2n+3}\hat{u}_{n+1}^{(q)} \quad . \tag{6.64}$$

The inversion of this tridiagonal integration operator yields a differentiation operator on the form

$$\forall n: \quad \hat{u}_n^{(q)} = (2n+1) \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{\infty} \hat{u}_p^{(q-1)} \quad . \tag{6.65}$$

As discussed in relation to Eq. (6.62), we may also compute, $\hat{u}_n^{(q)}$ through a backward recurrence for the truncated expansion by using

$$\forall n \in [1, N]: \quad \hat{u}_{n-1}^{(q)} = \frac{2n-1}{2n+3}\hat{u}_{n+1}^{(q)} + (2n-1)\hat{u}_n^{(q-1)} \quad , \tag{6.66}$$

provided only that we are dealing with a finite approximation to Eq.(6.63) such that $\hat{u}_N^{(q)} = \hat{u}_{N+1}^{(q)} = 0$.

6.3.1.2 The Continuous Chebyshev Expansion.

The continuous Chebyshev expansion of a function, $u(x) \in L_w^2[-1,1]$, becomes

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n T_n(x) \quad ,$$

with the expansion coefficients being

$$\hat{u}_n = \frac{1}{\gamma_n}(u, T_n)_{L_w^2[-1,1]} \quad ,$$

with

$$\gamma_n = \int_{-1}^{1} T_n(x)T_n(x)\frac{1}{\sqrt{1-x^2}}\,dx = c_n\frac{\pi}{2} \quad ,$$

and

$$c_n = \begin{cases} 2 & n = 0 \\ 1 & n > 0 \end{cases} \quad . \tag{6.67}$$

This additional constant is a consequence of the particular normalization we have chosen, i.e., $T_n(\pm 1) = (\pm 1)^n$. Thus, we obtain that

$$\hat{u}_n = \frac{1}{\gamma_n}(u, T_n)_{L_w^2[-1,1]} = \frac{2}{c_n \pi}\int_{-1}^{1} u(x)T_n(x)\frac{1}{\sqrt{1-x^2}}\,dx \quad .$$

Through Eq.(6.31), we obtain, using Theorem 39, a connection between the expansion coefficients, $\hat{u}_n^{(q-1)}$ and those of its derivative, $\hat{u}_n^{(q)}$, on the form $(n > 0)$

$$\hat{u}_n^{(q-1)} = \frac{c_{n-1}}{2n}\hat{u}_{n-1}^{(q)} - \frac{1}{2n}\hat{u}_{n+1}^{(q)} \quad , \tag{6.68}$$

where c_{n-1} enters due to the normalization. Inverting this tridiagonal

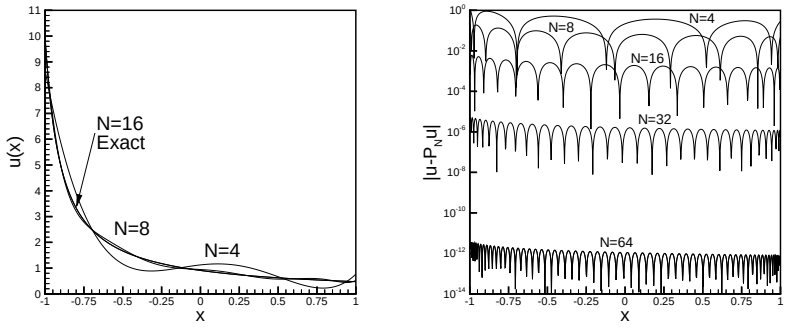

figure 6.5. a) Chebyshev series approximation of Example 25 for increasing resolution. b) Pointwise error of approximation for increasing resolution

integration operator yields a differentiation operator as

$$\forall n: \quad \hat{u}_n^{(q)} = \frac{2}{c_n} \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{\infty} p\hat{u}_p^{(q-1)} \quad . \tag{6.69}$$

To arrive at an $\mathcal{O}(N)$ method to compute the expansion coefficients for the approximate derivative we employ the backward recurrence provided by Eq. (6.68) as

$$\forall n \in [1, N]: \quad c_{n-1}\hat{u}_{n-1}^{(q)} = \hat{u}_{n+1}^{(q)} + 2n\hat{u}_n^{(q-1)} \quad , \tag{6.70}$$

where $\hat{u}_{N+1}^{(q)} = \hat{u}_N^{(q)} = 0$.

Prior to continuing the development of the discrete expansions let us consider an example to illustrate the resolution power of the Chebyshev expansion.

Example 25. Consider $u(x) \in C^\infty[-1, 1]$, as

$$u(x) = \frac{1}{x+a} \quad , \quad a > 1 \quad ,$$

for which the continuous expansion coefficients are given as

$$\hat{u}_n = \frac{2}{c_n} \frac{1}{\sqrt{a^2-1}} (\sqrt{a^2-1} - a)^n \quad .$$

As the function is smooth we find, as expected, that the expansion coefficients decay exponentially fast in n. Note that when a approaches 1 the function develops a strong gradient at $x = -1$ and becomes singular in the limit. In this example we used $a = 1.1$.

In Fig. 6.5 we plot the Chebyshev series approximation of $u(x)$ and the pointwise error for increasing N. We clearly observe the expected spectral convergence of the Chebyshev series and also that the convergence is uniform. Note that there is nothing special about the behavior of $u(x)$ at the boundaries, illustrating the resolution power of orthogonal polynomials for general functions defined on the bounded interval.

6.3.2 Gauss Quadrature for Ultraspherical Polynomials.

While the ultraspherical polynomials seem ideally suited for the approximation of functions defined on finite intervals we are facing a familiar problem when trying to use them. As for the continuous Fourier series expansion, using the ultraspherical polynomials requires the evaluation of an integral to recover the expansion coefficients, $\hat{u}_n$. For practical problems this is clearly not feasible.

Guided by the successful use of discrete approximations to the Fourier integral, leading to the discrete Fourier series and its dual formulation, we shall seek to identify similar discrete approximation to the integrals associated with the ultraspherical polynomials. As we shall learn shortly, classical Gauss quadratures provides the key step that enables the practical use of polynomial methods for the approximation of general functions.

Let us first recall the general polynomial expansion to approximate $u(x) \in L^2_w[-1,1]$, as

$$\mathcal{P}_N u(x) = \sum_{n=0}^{N} \hat{u}_n P_n^{(\alpha)}(x) \quad , \quad \hat{u}_n = \frac{1}{\gamma_n} \int_{-1}^{1} u(x) P_n^{(\alpha)}(x)(1-x^2)^{\alpha}\, dx \quad ,$$

where the normalizing factor, γ_n, is given in Eq.(6.6).

As for the Fourier series, Chap. 4, we seek to approximate the integrals as

$$\int_{-1}^{1} p(x)w(x)\, dx = \sum_{j=1}^{M} \tilde{w}_j p(\tilde{x}_j) + \sum_{j=0}^{N-M} w_j p(x_j) \quad ,$$

as well as possible. Here $\tilde{x}_j$ represents M preassigned grid points. To minimize the error associated with using the discrete sum rather than the integral we must determine the $2(N+1)-M$ unknowns, $\tilde{w}_j, w_j$ and x_j, that maximizes the order of the polynomial, $p(x)$, for which the summation is exact. The vector of $(\tilde{x}_j, x_j)$ then takes the role of the grid points on which $p(x)$ is to be evaluated, and $(\tilde{\omega}_j, \omega_j)$ the associated nodal weights.

One should note that provided only that the grid points, $(\tilde{x}_j, x_j)$, are distinct, one can always find a set of weights that integrate any polynomial, $p(x) \in \mathsf{B}_N$, exactly. The remarkable thing is that by carefully choosing the weights and the grid points, one can do much better than that.

Theorem 40. *Assume that a distinct set of $N+1$ collocation points, x_i, is given and construct the polynomial, $q(x) \in \mathsf{B}_{N+1}$, as*

$$q(x) = \prod_{j=1}^{M} (x - \tilde{x}_j) \prod_{j=0}^{N-M} (x - x_j) \ ,$$

where $\tilde{x}_j$ refers to M specific collocation points.

The quadrature rule

$$\int_{-1}^{1} p(x) w(x)\, dx = \sum_{j=1}^{M} \tilde{w}_j p(\tilde{x}_j) + \sum_{j=0}^{N-M} w_j p(x_j) \ ,$$

is exact for $p(x) \in \mathsf{B}_{2N+1-M}$ if and only if

a) It is exact for all $p(x) \in \mathsf{B}_N$.

b) For all $p(x) \in \mathsf{B}_{N-M}$, $p(x)$ is orthogonal to $q(x)$ in the weighted inner product as

$$(p, q)_{L^2_w[-1,1]} = \int_{-1}^{1} p(x) q(x) \omega(x)\, dx = 0 \ .$$

Proof: Let us first assume that the summation is exact for all $p(x) \in \mathsf{B}_{2N+1-M}$. The validity of assumption a) is thus trivial. Furthermore assume that $p(x) \in \mathsf{B}_{N-M}$. Then clearly $q(x)p(x) \in \mathsf{B}_{2N+1-M}$ for which the summation is exact. This results in

$$\int_{-1}^{1} q(x) p(x) w(x)\, dx = \sum_{j=1}^{M} \tilde{w}_j q(\tilde{x}_j) p(\tilde{x}_j) + \sum_{j=0}^{N-M} w_j q(x_j) p(x_j) = 0 \ ,$$

due to the construction of $q(x)$. This implies orthogonality of p and q as in assumption b).

Suppose conversely that a) and b) hold and that $p(x) \in \mathsf{B}_{2N+1-M}$. Then, by matching the coefficients, we can always find two polynomials, $r(x) \in \mathsf{B}_{N-M}$ and $s(x) \in \mathsf{B}_N$ such that

$$p(x) = q(x)r(x) + s(x) \quad .$$

Recalling condition b) implies

$$\begin{aligned} \int_{-1}^{1} p(x)w(x)\,dx &= \int_{-1}^{1} \left[q(x)r(x) + s(x)\right] w(x)\,dx \\ &= \int_{-1}^{1} s(x)w(x)\,dx \end{aligned}$$

as $r(x) \in \mathsf{B}_{N-M}$. However, the remaining term can be integrated exactly under assumption a), hence completing the proof that the two assumptions are both necessary and sufficient to guarantee the accuracy of the summation. QED

Theorem 40 is a remarkable result. It establishes the existence of summation rules, known as Gauss quadrature rules, that are exact for the integration of polynomials up to order $2N+1-M$ using only $N+1$ integration points. Note that increasing the number of specified collocation points results in a decreased maximum accuracy of the summation as it essentially removes degrees of freedom available to construct the summation rule.

We are still faced with the open problem of computing the quadrature points, x_j, and the weights required to form the quadrature rule. Let us, for a minute, assume that the collocation points are given. Then the $N+1$ weights, w_j and $\tilde{w}_j$, are recovered by using condition a) in Theorem 40 on the form

$$\forall k \in [0,\ldots,N]: \quad \int_{-1}^{1} x^k\, w(x)\,dx = \sum_{j=1}^{M} \tilde{w}_j(\tilde{x}_j)^k + \sum_{j=0}^{N-M} w_j(x_j)^k \quad . \tag{6.71}$$

As the summation is required to be exact for any $p(x) \in \mathsf{B}_N$ we can express this using the monomial basis to obtain $N+1$ equations for the $N+1$ unknown weights.

Computing the quadrature points, x_i, is a bit more involved and requires that assumption b) in Theorem 40 be used. An immediate

consequence of the required orthogonality between $q(x) \in \mathsf{B}_{N+1}$ and any ω-orthogonal $p(x) \in \mathsf{B}_{N-M}$ is that we can express $q(x)$ as

$$q(x) = p_{N+1}(x) + a_N p_N(x) + \ldots + a_{N-M+1} p_{N-M+1}(x) \quad . \tag{6.72}$$

The M constants, $a_N, \ldots, a_{N-M+1}$, must be found such that $q(\tilde{x}_j) = 0$, consistent with predefined grid points and the definition of $q(x)$ given in Theorem 40. The remaining quadrature points, x_j, can be found as the $N+1-M$ roots of the orthogonal polynomial, $q(x)$, in Eq.(6.72), provided only that the following holds

Theorem 41. *Consider the orthogonal polynomial, $p_N(x) \in \mathsf{B}_N$ with $N \geq 1$. Then the N roots of the polynomial are real, distinct, and located in the interior of $[-1, 1]$.*

Proof: The theorem follows from the orthogonality. Indeed, we have

$$(p_0, p_N)_{L^2_w[-1,1]} = \int_{-1}^{1} p_N(x) w(x)\, dx = 0 \quad ,$$

for $N \geq 1$. Since $w(x) > 0$ we know that $p_N(x)$ must change sign at least once in the interior of $[-1, 1]$. Assume that $x_0, \ldots, x_L$ represent the $L+1$ interior points at which $p_N(x)$ changes sign and construct the polynomial

$$p_L(x) = \prod_{j=0}^{L} (x - x_j) \quad .$$

As $p_L(x)$ changes sign at the same x_j as does $p_N(x)$, it is clear that the product $p_N(x)p_L(x)$ does not change sign. Thus, we recover that have $|(p_N, p_L)_w| > 0$, which is a contradiction of orthogonality except if $L = N$, i.e., $p_N(x)$ has exactly N real distinct roots in the interior of the domain. QED

This completes the required development of the quadrature rules for which we have identified the $N+1$ distinct quadrature points, x_j, as the roots of an orthogonal polynomial, $q(x) \in \mathsf{B}_{N+1}$, in Eq.(6.72), while the associated weights, w_j can be found by solving the linear system appearing from Eq.(6.71). To proceed beyond this point it is illustrative to restrict the attention to three special cases, distinguished by the number, M, of specified grid points in the quadrature rule.

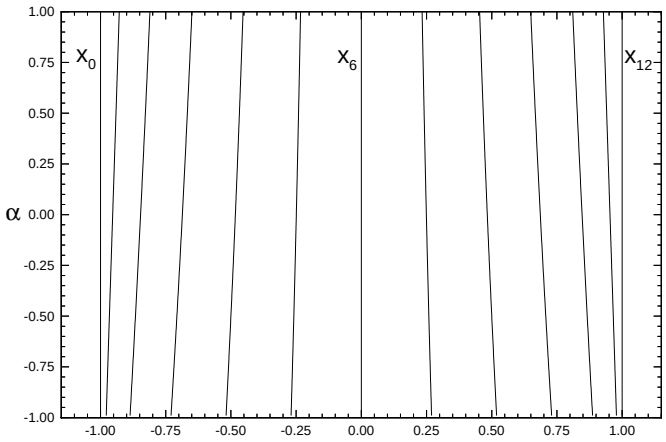

figure 6.6. The Ultraspherical Gauss-Lobatto collocation points, x_i, for $N = 12$ for the polynomial, $P_{12}^{(\alpha)}(x)$, as a function of α.

6.3.2.1 Gauss-Lobatto Quadrature.

The Gauss-Lobatto quadrature for the ultraspherical polynomial, $P_N^{(\alpha)}(x)$, is defined for $M = 2$, i.e., we define two collocation points, in this case the edge points, $\tilde{x}_1 = -1$ and $\tilde{x}_2 = 1$ while, following the result of Eq. (6.72), the remaining grid points are found as the roots of the polynomial

$$q(x) = P_{N+1}^{(\alpha)}(x) + a_N P_N^{(\alpha)}(x) + a_{N-1} P_{N-1}^{(\alpha)}(x) \quad , \tag{6.73}$$

where a_N and a_{N-1} are chosen such that $q(\pm 1) = 0$ and the weights, w_i, are found using Eq.(6.71). Hence, assume that the $N+1$ collocation points, x_i, are given as $-1 = x_0, x_1, \ldots, x_{N-1}, x_N = 1$ and the $N+1$ weights, w_i, ordered as $\tilde{w}_1 = w_0, w_1, \ldots, w_{N-1}, w_N = \tilde{w}_2$ are found as solutions to Eq.(6.71) it follows directly from Theorem 40 that

$$\int_{-1}^{1} p(x) w(x)\, dx = \sum_{j=0}^{N} p(x_j) w_j \quad ,$$

is exact for all $p(x) \in \mathsf{B}_{2N-1}$.

The weights, w_j, can be given explicitly as [?]

$$w_j = \begin{cases} (\alpha+1)\Pi_{N,j}^{(\alpha)} & j = 0, N \\ \Pi_{N,j}^{(\alpha)} & j \in [1, N-1] \end{cases} \quad , \tag{6.74}$$

where

$$\Pi_{N,j}^{(\alpha)} = 2^{2\alpha+1} \frac{\Gamma^2(\alpha+1)\Gamma(N+2\alpha+1)}{NN!(N+2\alpha+1)\Gamma^2(2\alpha+1)} \left[P_N^{(\alpha)}(x_j)\right]^{-2} \quad .$$

The quadrature nodes, x_j, on the other hand, can not in general be given in explicit form and must be obtained by numerical means. In Fig. 6.6 we plot the position of the ultraspherical Gauss-Lobatto nodes for $\alpha \in]-1, 1[$. One observation well worth making is that the nodes cluster close to the boundaries with the amount of clustering decreasing as α increases.

For the ultraspherical polynomials, there is a more convenient representation of $q(x)$, Eq.(6.73), as

$$q(x) = (1-x^2)\frac{d}{dx}P_N^{(\alpha)}(x) \quad . \tag{6.75}$$

This follows directly from Theorem 34 for Jacobi polynomials (Eq.(6.45) for ultraspherical polynomials) since

$$\begin{aligned}(p,q)_{L_w^2[-1,1]} &= \int_{-1}^{1} p(x)(1-x^2)\frac{d}{dx}P_N^{(\alpha)}(x)w(x)\,dx \\ &= \int_{-1}^{1} p(x)\left(c_{N-1,N}P_{N-1}^{(\alpha)}(x) + c_{N+1,N}P_{N+1}^{(\alpha)}(x)\right)w(x)\,dx = 0 \quad ,\end{aligned}$$

which we recognize as the condition from Theorem 40 specifying $q(x)$. The last reduction follows since $p(x) \in \mathsf{B}_{N-2}$.

6.3.2.2 Gauss-Radau Quadrature.

Rather than specifying two collocation points, one choose to include only one of the endpoints of the interval, $[-1, 1]$, in the summation. Such rules are known as Gauss-Radau quadrature methods.

If we chose to include the point $\tilde{y}_1 = -1$, the remaining quadrature points are found as the roots of the polynomial

$$q(y) = P_{N+1}^{(\alpha)}(y) + a_N P_N^{(\alpha)}(y) \quad , \tag{6.76}$$

with a_N being chosen such that $q(-1) = 0$ and the weights, v_j, are given as [?]

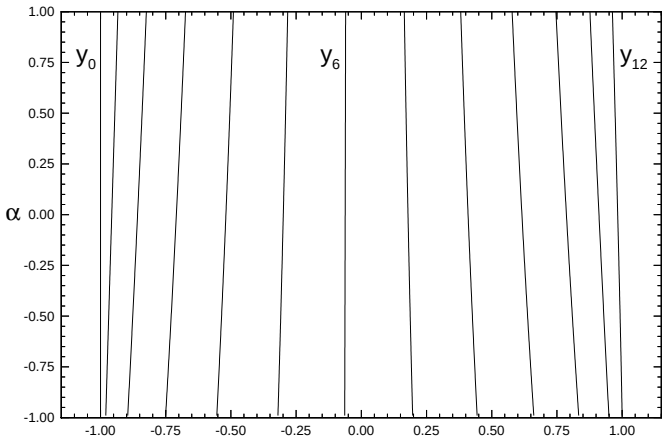

figure 6.7. The Gauss-Radau collocation points, y_i, for $N = 12$ for the ultraspherical, $P_{12}^{(\alpha)}(x)$, as a function of α with the node $y_0 = -1$ being fixed.

$$v_j = \begin{cases} (\alpha+1)\Pi_{N,0}^{(\alpha)} & j = 0 \\ \Pi_{N,j}^{(\alpha)} & j \in [1, N] \end{cases} \quad , \tag{6.77}$$

where

$$\Pi_{N,j}^{(\alpha)} = 2^{2\alpha} \frac{\Gamma^2(\alpha+1)\Gamma(N+2\alpha+1)}{N!(N+\alpha+1)(N+2\alpha+1)\Gamma^2(2\alpha+1)} (1-y_j) \left[P_N^{(\alpha)}(y_j)\right]^{-2} \quad .$$

Following Theorem 40 we recover that

$$\int_{-1}^{1} p(y) w(y)\, dy = \sum_{j=0}^{N} p(y_j) v_j \quad ,$$

is exact for all $p(y) \in \mathsf{B}_{2N}$.

The quadrature nodes, y_j, must be obtained by numerical means. In Fig. 6.7 we plot the position of the ultraspherical Gauss-Radau quadrature nodes for $N = 12$ and $\alpha \in]-1, 1[$. As in the case of the Gauss-Lobatto quadrature the nodes cluster close to the boundaries with the amount of clustering decreasing as α increasing and only the left boundary is included in the nodal set.

The formulation of a Gauss-Radau quadrature that includes the right endpoint follows directly from the above by mirroring the weights, v_j,

as well as the quadrature nodes, y_j, around the center of the interval.

6.3.2.3 Gauss Quadrature

Let us finally consider the case where no quadrature points are specified a priori, i.e. $M = 0$, in which case all quadrature points are in the interior of the domain $[-1, 1]$ following Theorem 41. The quadrature points, z_j, appear as the roots of the polynomial,

$$q(z) = P_{N+1}^{(\alpha)}(z) \quad , \tag{6.78}$$

while the weights, u_j, are obtained from Eq.(6.71) on the form [?]

$$u_j = 2^{2\alpha+1} \frac{\Gamma^2(\alpha+1)\Gamma(2\alpha+N+2)}{\Gamma(N+2)\Gamma^2(2\alpha+1)(1-z_j^2)} \left[\frac{d}{dx} P_{N+1}^{(\alpha)}(z_j)\right]^{-2} , \tag{6.79}$$

for all $j \in [0, N]$. We obtain directly from Theorem 40 that

$$\int_{-1}^{1} p(z)w(z)\,dz = \sum_{j=0}^{N} p(z_j)u_j \quad ,$$

is exact for all $p(x) \in \mathsf{B}_{2N+1}$. This is recognized as the classic Gauss quadrature, providing the rule of exact integration of a polynomial of maximum order.

The quadrature points, z_j, are generally not given on analytic form and in Fig. 6.8 we plot for illustration the position of the nodes in the case of the ultraspherical polynomials for $N = 12$ and $\alpha \in]-1, 1[$. We emphasize, as is also evident from Fig. 6.8, that the nodal set associated with the Gauss quadrature does not include any of the endpoints of the interval.

6.3.2.4 Quadrature for Legendre Polynomials

The quadrature formulas for the integration of polynomials specified at the Legendre quadrature points can be obtained directly from the formulas derived above by setting $\alpha = 0$. Due to the extensive use of the Legendre polynomials we summarize the expressions in the following.

Legendre Gauss-Lobatto Quadrature. The Legendre Gauss-Lobatto quadrature points, x_j, appear as the roots of the polynomial

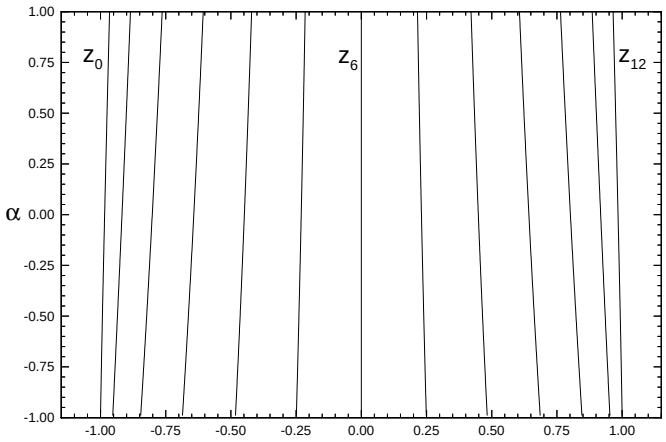

figure 6.8. The Gauss collocation points, z_i, for $N = 12$ for the ultraspherical polynomial, $P_{12}^{(\alpha)}(x)$, as a function of α.

$$q(x) = (1 - x^2)\frac{d}{dx}P_N(x) \quad , \tag{6.80}$$

following Eq.(6.75). Unfortunately, no explicit formula is known for these. The weights, w_j, can be recovered directly from Eq.(6.74) on the form

$$w_j = \frac{2}{N(N+1)}[P_N(x_j)]^{-2} \quad . \tag{6.81}$$

Legendre Gauss-Radau Quadrature. The Legendre Gauss-Radau quadrature points, y_j, appear as the roots of the polynomial

$$q(y) = P_{N+1}(y) + P_N(y) \quad , \tag{6.82}$$

by using Eq.(6.76) assuming that $y = -1$ is included in the set of quadrature points for which no explicit formula is known. The weights, v_j, are given as

$$v_j = \frac{1}{(N+1)^2}\frac{1 - y_j}{[P_N(y_j)]^2} \quad . \tag{6.83}$$

Legendre Gauss Quadrature. From Eq.(6.78) we recover the Legendre Gauss quadrature points, z_j, as the $N + 1$ roots of the polynomial

$$q(z) = P_{N+1}(z) \quad . \tag{6.84}$$

The weights, u_j, are obtained directly from Eq.(6.79) for $\alpha = 0$ on the form

$$\forall j \in [0, N]: \quad u_j = \frac{2}{[1-(z_j)^2][P'_{N+1}(z_j)]^2} \quad . \tag{6.85}$$

6.3.2.5 Quadrature for Chebyshev Polynomials

The quadrature formulas, i.e., the quadrature points and the corresponding weights, for the Chebyshev polynomials can likewise be obtained from the general results for the ultraspherical polynomials, albeit some care is needed due to the required normalization, Eq.(6.34).

The Chebyshev quadrature distinguishes itself from the previous cases by allowing for explicit and simple expressions for the quadrature points as well as the corresponding weights. This supplies a compelling motivation for the use of these polynomials, apart from the fact that they are well suited for the approximation of general functions as discussed in Sec. 6.2.2.2.

Chebyshev Gauss-Lobatto Quadrature. The Chebyshev Gauss-Lobatto quadrature points, x_i, are given explicitly as

$$x_j = -\cos\left(\frac{\pi}{N} j\right) \quad , \quad j \in [0, \ldots, N] \quad . \tag{6.86}$$

as the roots of

$$q(x) = (1-x^2)\frac{d}{dx} T_N(x) \quad , \tag{6.87}$$

from Eq.(6.75). The corresponding weights, w_j, appear directly from Eq.(6.74)

$$w_j = \frac{\pi}{c_j N} \quad , \quad c_j = \begin{cases} 2 & j = 0, N \\ 1 & j \in [1, N-1] \end{cases} \quad . \tag{6.88}$$

Note the reduction in the expressions of the weights as a consequence of the equioscillatory property of the Chebyshev polynomials

$$T_N(x_j) = (-1)^{N+j} \quad .$$

Chebyshev Gauss-Radau Quadrature. The Chebyshev Gauss-Radau quadrature points, y_j, appear as the roots of the polynomial

$$q(y) = T_{N+1}(y) + T_N(y) \quad , \tag{6.89}$$

from Eq.(6.76) with the explicit expression

$$y_j = -\cos\left(\frac{2\pi}{2N+1}j\right) \quad , \quad j \in [0,\ldots,N] \quad , \tag{6.90}$$

assuming only that the left endpoint is included in the set of quadrature points. The weights, v_j, are given as

$$v_j = \frac{\pi}{c_j}\frac{1}{2N+1} \tag{6.91}$$

directly from Eq.(6.77). Here c_n takes the usual meaning, Eq.(6.67).

Chebyshev Gauss Quadrature. Equation (6.78) defines the Chebyshev Gauss quadrature points, z_j, as the $N+1$ roots of the polynomial

$$q(z) = T_{N+1}(z) \quad , \tag{6.92}$$

i.e., the quadrature points are

$$z_j = -\cos\left(\frac{(2j+1)\pi}{2N+2}\right) \quad , \quad j \in [0,\ldots,N] \quad . \tag{6.93}$$

We recognize this set of grid points as those derived in Sec. 6.2.2.2 which specifies the best approximating polynomial to zero, i.e., these points are particularly well suited for polynomial interpolation and, through an entirely different procedure, appear also as the Chebyshev Gauss quadrature points.

The weights, u_j, are obtained directly from Eq.(6.79) as

$$\forall j \in [0,N]: \quad u_j = \frac{\pi}{N+1} \quad , \tag{6.94}$$

i.e., they are constant for all the quadrature points. The Chebyshev Gauss quadrature is the only case among the quadrature for the Jacobi polynomials for which this is the case.

6.3.3 Discrete Inner Products and Norms.

The identification of the quadrature formulas enables the introduction of discrete versions of the inner product and the corresponding L^2_w-norm. We recall the continuous case which take the form

$$(f,g)_{L^2_w[-1,1]} = \int_{-1}^{1} f(x)g(x)w(x)\,dx \quad , \quad \|f\|_{L^2_w[-1,1]} = (f,f)^{1/2}_{L^2_w[-1,1]} \quad ,$$

for $f, g \in L^2_w[-1,1]$. Using the quadrature formulas it seems natural to define the corresponding discrete inner product

$$[f,g]_w = \sum_{j=0}^{N} f(x_j)g(x_j)w_j \quad , \quad \|f\|_{N,w} = [f,f]^{1/2}_w \quad ,$$

where $x_j = (x_j, y_j, z_j)$ can be any of the Gauss quadrature points with the corresponding weights, $w_j = (w_j, v_j, u_j)$, and $f, g \in \mathsf{B}_N$. We note that in the event that $f, g \in \mathsf{B}_N$ the discrete inner product and norm based on the Gauss-Radau and Gauss quadratures are identical to the continuous ones due to the accuracy of the quadratures. This, however, ceases to be true for the Gauss-Lobatto quadrature as $f(x)g(x) \in \mathsf{B}_{2N}$ and the quadrature is no longer exact.

Let us first compute the norm, $\tilde{\gamma}_n$, of $P_n^{(\alpha)}(x)$ using the three types of quadrature. Clearly, using Gauss or Gauss-Radau quadrature, we immediately recover

$$\tilde{\gamma}_n = \left(P_n^{(\alpha)}, P_n^{(\alpha)}\right)_{L^2_w[-1,1]} = 2^{2\alpha+1}\frac{\Gamma^2(\alpha+1)\Gamma(n+2\alpha+1)}{n!(2n+2\alpha+1)\Gamma^2(2\alpha+1)} \quad , \quad (6.95)$$

using Eq.(6.59).

For the Gauss-Lobatto quadrature, one obtains a slightly different conclusion as the quadrature is inexact for $n = N$. However, the following result allows us to evaluate the inner product.

Lemma 8. The ultraspherical polynomials, $P_N^{(\alpha)}(x)$, satisfy

$$\forall j \in [1, N-1]: \quad \frac{d}{dx}P_{N-1}^{(\alpha)}(x_j) = -NP_N^{(\alpha)}(x_j) \quad ,$$

provided x_j represents the interior Gauss-Lobatto quadrature points.

Using this result, the proof of which is left as an exercise, the weights together with Eq.(6.90) yields the norm for $P_N^{(\alpha)}(x)$ using the Gauss-Lobatto quadrature

$$\tilde{\gamma}_N = \|P_N^{(\alpha)}\|^2_{N,w} = 2^{2\alpha+1}\frac{\Gamma^2(\alpha+1)\Gamma(N+2\alpha+1)}{NN!\Gamma^2(2\alpha+1)} \quad . \quad (6.96)$$

While the discrete Gauss-Lobatto norm is slightly different from the continuous norm, it follows immediately from the above that they are equivalent for any polynomial, $u_N \in \mathsf{P}_N$, since

$$\|u_N\|_{L^2_w[-1,1]} \leq \|u_N\|_{N,\omega} \leq \sqrt{2 + \frac{2\alpha+1}{N}} \|u_N\|_{L^2_w[-1,1]} \ ,$$

as $\tilde{\gamma}_N > \gamma_N$ for all values of N and $\alpha > -1$.

Legendre Polynomials To summarize the results for Legendre polynomials, $P_n(x)$, we recover the discrete norms, $\tilde{\gamma}_n$, obtained for $\alpha = 0$ from Eqs.(6.95)-(6.96) as

$$\tilde{\gamma}_n = \begin{cases} \frac{2}{2n+1} & n < N \\ \frac{2}{2N+1} & n = N \text{ for Gauss and Gauss-Radau quadrature} \\ \frac{2}{N} & n = N \text{ for Gauss-Lobatto quadrature} \end{cases} . \quad (6.97)$$

Chebyshev Polynomials The results for the Chebyshev polynomials, $T_n(x)$, can likewise be summarized as

$$\tilde{\gamma}_n = \begin{cases} c_n \frac{\pi}{2} & n < N \\ \frac{\pi}{2} & n = N \text{ for Gauss and Gauss-Radau quadrature} \\ \pi & n = N \text{ for Gauss-Lobatto quadrature} \end{cases} , \quad (6.98)$$

where c_n takes the usual meaning defined in Eq.(6.67).

6.3.4 The Discrete Expansion

With the development of the quadrature rules, we have the tools in place to formulate accurate methods to commutate the expansion coefficients based on summations rather than integrations. As we identified several different ways of accurately approximating the integrals of polynomials using Gauss quadratures, it comes as no surprise that we can also define several discrete expansions with slightly different properties.

Let us first recall the definition of the continuous expansion as

$$\mathcal{P}_N u(x) = \sum_{n=0}^{N} \hat{u}_n P_n^{(\alpha)}(x) \ , \quad \hat{u}_n = \frac{1}{\gamma_n} \int_{-1}^{1} u(x) P_n^{(\alpha)}(x) (1-x^2)^\alpha \, dx \ ,$$

where the normalizing factor, γ_n, is given in Eq.(6.60).

Using the Gauss-Lobatto quadrature it is natural to define a discrete

approximation to the expansion on the form

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n P_n^{(\alpha)}(x) \quad , \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) P_n^{(\alpha)}(x_j) w_j \quad , \tag{6.99}$$

where the collocation points, x_j, are found as the roots of the polynomial, Eq.(6.75), while the corresponding weights, w_j, are given in Eq.(6.74) and the normalizing factor, $\tilde{\gamma}_n$, in Eqs.(6.95)-(6.96).

Likewise, we can base the definition of the discrete expansion coefficients on the use of the Gauss quadrature as

$$\mathcal{I}_N u(z) = \sum_{n=0}^{N} \tilde{u}_n P_n^{(\alpha)}(z) \quad , \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(z_j) P_n^{(\alpha)}(z_j) u_j \quad , \tag{6.100}$$

with the collocation points, z_j, obtained as the roots of the polynomial, Eq.(6.78), and the corresponding weights, u_j, from Eq.(6.79) with the normalization, $\tilde{\gamma}_n$, given in Eq.(6.95).

We could naturally also define discrete expansion coefficients based on the Gauss-Radau quadrature points. However, due to their little use we shall not pursue this approach in detail but simply state the central results.

The approximation of the discrete expansion coefficients using the Gauss quadratures has a striking consequence.

Theorem 42. *Let the discrete expansion coefficients, $\tilde{u}_n$, be an approximation to the continuous expansion coefficients, $\hat{u}_n$, obtained by using a Gauss quadrature.*

For any function, $u(x) \in L_w^2[-1,1]$, we then have

$$\forall x_j \ : \ \mathcal{I}_N u(x_j) = u(x_j) \quad ,$$

where x_j are the quadrature points associated with the Gauss quadrature.

Proof: Let us first demonstrate the result in case a Gauss quadrature, Eq.(6.100), is used to approximate the inner product. Introducing the discrete expansion coefficients into the polynomial approximation we recover

$$\begin{aligned}\mathcal{I}_N u(z) &= \sum_{n=0}^{N} \tilde{u}_n P_n^{(\alpha)}(z) \\ &= \sum_{n=0}^{N} \left(\frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(z_j) P_n^{(\alpha)}(z_j) u_j \right) P_n^{(\alpha)}(z) \\ &= \sum_{j=0}^{N} u(z_j) \left(u_j \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} P_n^{(\alpha)}(z) P_n^{(\alpha)}(z_j) \right) \\ &= \sum_{j=0}^{N} u(z_j) \tilde{l}_j(z) \quad ,\end{aligned}$$

where we have defined the polynomial

$$l_j(z) = u_j \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} P_n^{(\alpha)}(z) P_n^{(\alpha)}(z_j) \quad . \tag{6.101}$$

We note that $l_j(z) \in \mathsf{B}_N$ as expected. To establish the theorem we need to show that $l_j(z)$ is the Lagrange interpolation polynomial based on the Gauss quadrature nodes, i.e., $l_j(z_i) = \delta_{ij}$.

This follows by realizing that the sum in Eq.(6.101) can be evaluated explicitly by the Christoffel-Darboux, Theorem 33. Recalling that $\tilde{\gamma}_n = \gamma_n$ in the Gauss quadrature, Eq.(6.95), we recover

$$l_j(z) = u_j 2^{-(2\alpha+1)} \frac{\Gamma(N+2)\Gamma^2(2\alpha+1)}{\Gamma(N+2\alpha+1)\Gamma^2(\alpha+1)} \frac{P_{N+1}^{(\alpha)}(z) P_N^{(\alpha)}(z_j)}{z - z_j} \quad ,$$

since $P_{N+1}^{(\alpha)}(z_j) = 0$ defines the Gauss points. Clearly $l_j(z_i) = 0$ for $i \neq j$. Using l'Hospital rule we obtain for $i = j$ that

$$l_j(z_j) = u_j 2^{-(2\alpha+1)} \frac{\Gamma(N+2)\Gamma^2(2\alpha+1)}{\Gamma(N+2\alpha+1)\Gamma^2(\alpha+1)} P_N^{(\alpha)}(z_j) \frac{d}{dx} P_{N+1}^{(\alpha)}(z_j) \quad .$$

Introducing the Gauss weights, Eq.(6.79), yields

$$l_j(z_j) = (N+2\alpha+1) P_N^{(\alpha)}(z_j) \left[(1-z_j^2) \frac{d}{dx} P_{N+1}^{(\alpha)}(z_j) \right]^{-1} = 1 \quad ,$$

where the last reduction follows by combining Eq.(6.40) and Eq.(6.45) and using the definition of the Gauss quadrature points.

While the proof of the interpolation property of the discrete expansion based on the Gauss-Radau quadrature follows the above exactly, the case for the Gauss-Lobatto quadrature is more involved since $\tilde{\gamma}_N \neq \gamma_N$, Eqs.(6.95)-(6.96).

However, since the polynomial is given as

$$l_j(x) = w_j \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} P_n^{(\alpha)}(x) P_n^{(\alpha)}(x_j) \quad ,$$

the result can be recovered by separating out the last term, utilizing the Christoffel-Darboux identity and the Gauss-Lobatto weights, given in Eq.(6.74), together with Lemma 8 for the interior nodes. The result follows by considering the internal and edge nodes separately. QED

Similar to the case of the Fourier expansion, Chap. 4, we have thus established that there are two mathematically equivalent but computationally different ways of dealing with the discrete expansion. Indeed, we can use the discrete expansion coefficients, Eqs.(6.99)-(6.100), directly or we can make use of the Lagrange interpolation polynomial, $l_j(x)$, identified in Theorem 42.

To proceed along this latter line of thinking, we need to look further into the expressions for the Lagrange polynomials.

For interpolating at the Gauss-Lobatto nodes we have

Theorem 43. *The Lagrange interpolation polynomial, $l_j(x)$, based on the ultraspherical Gauss-Lobatto quadrature points, x_j, is given as*

$$l_j(x) = \begin{cases} (\alpha+1)\Pi_{N,j}^{(\alpha)}(x) & j = 0, N \\ \Pi_{N,j}^{(\alpha)}(x) & else \end{cases} \quad , \qquad (6.102)$$

where

$$\Pi_{N,j}^{(\alpha)}(x) = -\frac{1}{N(N+2\alpha+1)} \frac{(1-x^2)\left(P_N^{(\alpha)}\right)'(x)}{(x-x_j)P_N^{(\alpha)}(x_j)} \quad .$$

Proof: Using the Christoffel-Darboux identity the Lagrange interpolation polynomial can be expressed directly by following the proof of Theorem 42. However, as $l_j(x) \in \mathsf{B}_N$ and $l_j(x_i) = \delta_{ij}$ it suffices to construct a polynomial in B_N with this property to recover the unique Lagrange polynomial.

If we first look at the case of $j = 0$ it is clear that the polynomial

$$g(x) = c(x-1)\frac{d}{dx}P_N^{(\alpha)}(x) \quad ,$$

vanishes at all quadrature points with the exception of $x = -1$. Utilizing Eq.(6.38) and Eq.(6.39) one can straightforwardly obtain the missing constant c such that $g(-1) = 1$. The same approach can be used to recover the expression for $j = N$.

For the interior nodes, we realize that

$$g(x) = c\frac{1}{x - x_j}(1-x^2)\frac{d}{dx}P_N^{(\alpha)}(x) \quad ,$$

vanishes at all Gauss-Lobatto quadrature points and c is to be determined such that $g(x_j) = 1$. This is readily achieved by combining Eq.(6.40) and Eq.(6.45) to express the denominator of $g(x)$ in terms of $P_N^{(\alpha)}$ and $P_{N-1}^{(\alpha)}$, and apply l'Hospitals rule in combination with Lemma 8 and the definition of the Gauss-Lobatto quadrature points. Alternatively, one can first utilize l'Hospitals rule on $g(x)$ and then exploit that $P_N^{(\alpha)}$ satisfies a Sturm-Liouville equation, Eq.(6.35), to recover the same result. QED

Theorem 44. *The Lagrange interpolation polynomial, $l_j(y)$, based on the ultraspherical Gauss-Radau quadrature points, y_j, is given as*

$$l_j(y) = \begin{cases} (\alpha+1)\Pi_{N,j}^{(\alpha)}(y) & j = 0, N \\ \Pi_{N,j}^{(\alpha)}(y) & else \end{cases} \quad , \qquad (6.103)$$

where

$$\Pi_{N,j}^{(\alpha)}(y) = \frac{1}{2(N+\alpha+1)(N+2\alpha+1)}\frac{(1-y_j)}{P_N^{(\alpha)}(y_j)} \frac{(N+1)P_{N+1}^{(\alpha)}(y) + (N+2\alpha+1)P_N^{(\alpha)}}{y - y_j} \quad .$$

Theorem 45. *The Lagrange interpolation polynomial, $l_j(z)$, based on the ultraspherical Gauss quadrature points, z_j, is given as*

$$l_j(z) = \frac{P_{N+1}^{(\alpha)}(z)}{(z - z_j)\left(P_{N+1}^{(\alpha)}\right)'(z)} \quad . \tag{6.104}$$

Both results follow directly from the application of Theorem 33, the properties of the ultraspherical polynomials, Eqs.(6.40)-(6.45), the associated Sturm-Liouville equation, Eq.(6.35), and the definition of the quadrature.

Before we continue with the development of the methods based on discrete expansion coefficients and the equivalent formulation using the Lagrange interpolation polynomials, let us briefly touch on the issue of the aliasing errors associated with the use of the discrete expansion coefficients.

To obtain an estimate for the aliasing error, we consider the continuous function, $u(x)$, and relate the discrete expansion coefficients to the continuous expansion coefficients to recover

$$\begin{aligned}
\tilde{u}_n &= \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) P_n^{(\alpha)}(x_j) w_j \\
&= \sum_{l=0}^{\infty} \hat{u}_l \left(\frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} P_l^{(\alpha)}(x_j) P_n^{(\alpha)}(x_j) w_j \right) \\
&= \hat{u}_n + \sum_{l \geq N}^{\infty} \hat{u}_l \left[P_l^{(\alpha)}, P_n^{(\alpha)} \right]_w \quad ,
\end{aligned}$$

where the last term, representing the aliasing error, remains since the Gauss quadrature ceases to be exact for $l \geq N$. Here l is strictly larger than N for Gauss and Gauss-Radau integration only. Hence, the aliasing error takes the form

$$\|\mathcal{R}_N u\|_w^2 = \left\| \sum_{n=0}^{N} \sum_{l \geq N}^{\infty} \hat{u}_l \left[P_l^{(\alpha)}, P_n^{(\alpha)} \right]_w P_n^{(\alpha)}(x) \right\|_w^2 \quad .$$

Contrary to the situation for the trigonometric polynomials, we have no simple expression for this since $\left[P_l^{(\alpha)}, P_n^{(\alpha)} \right]_w \neq 0$ for any $l \geq N$ with a few notable exceptions for special choices of α.

As we shall discuss further in Sec. 6.6.1, only crude bounds on the aliasing error is known although sharper results are available for the im-

portant cases of Legendre and Chebyshev methods based on the Gauss and Gauss-Lobatto nodes. Nevertheless, it is safe to state that if the function, $u(x)$, is sufficiently smooth the global convergence rate remains spectral, i.e., while the aliasing error may have a quantitative impact it does not change the qualitative behavior of the approximation as compared to the continuous expansion. We shall return to these questions in more detail in Sec. 6.6.1.

Let us now return to the issue of how to obtain an approximation to the derivative of a function once the discrete approximation, expressed by using the discrete expansion coefficients or the Lagrange polynomials, is known.

We first focus on the use of the discrete expansion coefficients in which case the approximations to derivatives is obtained in a similar way as if the continuous expansion coefficients were being used. If we consider the Gauss-Lobatto approximation

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n P_n^{(\alpha)}(x) \quad , \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) P_n^{(\alpha)}(x_j) w_j \quad ,$$

we obtain an approximation to the derivative of $u(x)$ as

$$\mathcal{I}_N \frac{d}{dx} u(x) = \sum_{n=0}^{N} \tilde{u}_n^{(1)} P_n^{(\alpha)}(x) \quad ,$$

where the new expansion coefficients, $\tilde{u}_n^{(1)}$, are obtained by using the backward recursion

$$\tilde{u}_n^{(1)} = (2n + 2\alpha + 1) \left[\frac{1}{2n + 2\alpha + 5} \tilde{u}_{n+2}^{(1)} + \tilde{u}_{n+1} \right] \quad ,$$

from Eq.(6.62) and initialized by $\tilde{u}_{N+1}^{(1)} = \tilde{u}_N^{(1)} = 0$ as $\mathcal{I}_N u(x) \in \mathsf{B}_N$. We note that

$$\frac{d}{dx} \mathcal{I}_N u(x) \neq \mathcal{I}_{N-1} \frac{d}{dx} u(x) \quad ,$$

as is expected since even the continuous expansion lacks this property. Furthermore, in the discrete case we have additional errors caused by aliasing. Although the computation of the derivative here is exemplified using the Gauss-Lobatto quadrature the same holds if one is using discrete expansion coefficients based on any of the Gauss quadrature points. Higher derivatives is computed by applying the recurrence rela-

tion repeatedly.

In the equivalent formulation of the discrete expansion, utilizing the Lagrange interpolating polynomial as

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad ,$$

the derivative of $u(x)$ at the collocation points is approximated simply by differentiating the global polynomial, $l_j(x)$. In particular, if one evaluates it at the quadrature points one obtains

$$\mathcal{I}_N \frac{d}{dx} \mathcal{I}_N u(x_i) = \sum_{j=0}^{N} u(x_j) \left. \frac{dl_j(x)}{dx} \right|_{x_i} = \sum_{j=0}^{N} \mathrm{D}_{ij} u(x_j) \quad .$$

where the $(N+1) \times (N+1)$ differentiation matrix, D, has been introduced.

Considering first the Lagrange interpolation polynomial based on the Gauss-Lobatto quadrature points, Theorem 43, the corresponding differentiation matrix, D, has the entries

$$\mathrm{D}_{ij} = \begin{cases} \frac{\alpha - N(N+2\alpha+1)}{2(\alpha+2)} & i = j = 0 \\ \frac{\alpha x_j}{1-(x_i)^2} & i = j \in [1, N-1] \\ \frac{\alpha+1}{x_i - x_j} \frac{P_N^{(\alpha)}(x_i)}{P_N^{(\alpha)}(x_j)} & i \neq j, j = 0, N \\ \frac{1}{x_i - x_j} \frac{P_N^{(\alpha)}(x_i)}{P_N^{(\alpha)}(x_j)} & i \neq j, j \in [1, N-1] \\ -\mathrm{D}_{00} & i = j = N \end{cases} \tag{6.105}$$

Although the derivation, relying heavily on the properties of the ultraspherical polynomials summarized in Sec. 6.2.2.3, of these entries is somewhat lengthy it is nevertheless straightforward and the details are left as an exercise.

A similar result is obtained for the approximation based on the Gauss quadrature nodes and the associated Lagrange polynomial in Theorem 45. In this case the differentiation matrix, D, has the entries

$$\mathrm{D}_{ij} = \begin{cases} \frac{(\alpha+1) z_i}{1-(z_i)^2} & i = j \\ \frac{\left(P_{N+1}^{(\alpha)}\right)'(z_i)}{(z_i - z_j)\left(P_{N+1}^{(\alpha)}\right)'(z_j)} & i \neq j \end{cases} \quad . \tag{6.106}$$

For an approximation based on the use of the Gauss-Radau quadrature

points one can recover an equivalent expression by using the associated Lagrange interpolation polynomial given in Theorem 44.

Some or all of the differentiation matrices share a number of properties that we shall find it useful to be aware of. In particular, we have

Theorem 46. *The differentiation matrix,* D*, derived from the Lagrange interpolation polynomial based on any of the Gauss quadrature points, is nilpotent.*

Proof: We establish the result for the differentiation matrix based on the Gauss-Lobatto points. If we recall that the interpolation polynomial takes the form

$$l_j(x) = w_j \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} P_n^{(\alpha)}(x_j) P_n^{(\alpha)}(x) \quad ,$$

we can express the entries to the differentiation matrix as

$$\mathrm{D}_{ij} = w_j \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} P_n^{(\alpha)}(x_j) \frac{dP_n^{(\alpha)}(x_i)}{dx} = w_j \left(\boldsymbol{P}(x_j)\right)^T \left(\boldsymbol{P}_x(x_i)\right) \quad ,$$

where

$$\boldsymbol{P}(x_j) = \left[\frac{P_0^{(\alpha)}(x_j)}{\sqrt{\tilde{\gamma}_0}}, , \ldots , \frac{P_N^{(\alpha)}(x_j)}{\sqrt{\tilde{\gamma}_N}} \right]^T \quad ,$$

and

$$\boldsymbol{P}_x(x_j) = \left[\frac{1}{\sqrt{\tilde{\gamma}_0}} \frac{dP_0^{(\alpha)}(x_j)}{dx}, \ldots , \frac{1}{\sqrt{\tilde{\gamma}_N}} \frac{dP_N^{(\alpha)}(x_j)}{dx} \right]^T \quad .$$

Using Eq.(6.43), we can relate these two vectors by a linear transformation as

$$\boldsymbol{P}(x_j) = \mathrm{B}\boldsymbol{P}_x(x_j) \quad ,$$

where B is a bi-diagonal matrix with the entries given in Eq.(6.44). Introducing the (pseudo)-inverse of B we have

$$\mathrm{B}^{-1}\boldsymbol{P}(x_j) = \boldsymbol{P}_x(x_j) \quad ,$$

where B^{-1} is strictly upper triangular due to the special structure of B.

From this we recover

$$\mathrm{D}_{ij} = w_j \left(\boldsymbol{P}(x_j)\right)^T \mathrm{B}^{-1} \boldsymbol{P}(x_i) \ .$$

However, since $\boldsymbol{P}(x_j)^T \boldsymbol{P}(x_i) = w_j^{-1}\delta_{ij}$, we have that the differentiation matrix is uniformly similar to B^{-1}, which is strictly upper triangular. Hence

$$\mathrm{D}^{N+1} = \left(\mathrm{B}^{-1}\right)^{N+1} = 0 \ ,$$

confirming that D is indeed nilpotent.

As the proof relies entirely on the properties of the polynomials it applies also to the approximations based on the Gauss and the Gauss-Radau quadrature points. QED

This property is hardly a surprise, i.e., by differentiating a polynomial the order of the polynomial is reduced by one order and after $N+1$ differentiations the polynomial vanishes identically.

Theorem 47. *The differentiation matrix,* D*, based on the Gauss or the Gauss-Lobatto quadrature points is centro-antisymmetric*

$$\mathrm{D}_{ij} = -\mathrm{D}_{N-i,N-j} \ .$$

Proof: This property follows immediately from the expressions for the entries of D in Eq.(6.105) and Eq.(6.106), the even-odd symmetry of the ultraspherical polynomials, Eq.(6.42), and, as a reflection of this, the the symmetry of the quadrature points around $x = 0$. QED

It is worth emphasizing that the differentiation matrix based on Gauss-Radau quadrature points does not possess the centro-antisymmetric property due to the lack of symmetry in the grid points. As we shall return to in Chap. 9, this subtle symmetry enables a factorization of the differentiation matrices that ultimately allows for the computation of the derivatives at a reduced cost.

The computation of higher derivatives follows the approach for the computation of the first derivative. One may compute entries of the q'th order differentiation matrix, $\mathrm{D}^{(q)}$ by evaluating the q'th derivative of Lagrange interpolation polynomial at the quadrature points. Alterna-

tively, one may compute the q'th order differentiation matrix by simply multiplying the first order differentiation matrices, i.e.,

$$\mathrm{D}^{(q)} = (\mathrm{D})^q \quad ,$$

where $q \leq N$. Although this latter approach certainly is appealing in terms of simplicity, we shall experience in Chap. 9 that care is warranted in defining the entries of the differentiation matrices. Hence, whenever possible, one should strive to use the exact expressions for the entries rather than those obtained through matrix multiplications to lessen the impact of finite precision arithmetics.

6.3.4.1 The Discrete Legendre Expansion

Based on the theory developed above let us summarize the results for methods using discrete Legendre expansions, recovered for $\alpha = 0$.

Legendre Gauss Lobatto. In this case we consider

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n P_n(x) \quad , \quad \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) P_n(x_j) w_j \quad , \tag{6.107}$$

where $\tilde{\gamma}_n$ is given in Eq.(6.97) and the quadrature points, x_j, and the weights, w_j, are given as the solution to Eq.(6.80) and in Eq.(6.81), respectively.

Computation of the expansion coefficients for the derivatives is done using the backward recurrence relation given in Eq.(6.66).

Since $\mathcal{I}_N u(x)$ is the interpolant of $u(x)$ at the Legendre Gauss Lobatto quadrature points, as stated in Theorem 42, we may express the approximation as

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad , \tag{6.108}$$

where the Lagrange interpolation polynomial, obtained directly from Eq.(6.102) with $\alpha = 0$, takes the form

$$l_j(x) = \frac{-1}{N(N+1)} \frac{(1-x^2) P_N'(x)}{(x - x_j) P_N(x_j)} \quad . \tag{6.109}$$

Examples of the Lagrange polynomials based on the Legendre Gauss Lobatto points are shown in Fig. 6.9.

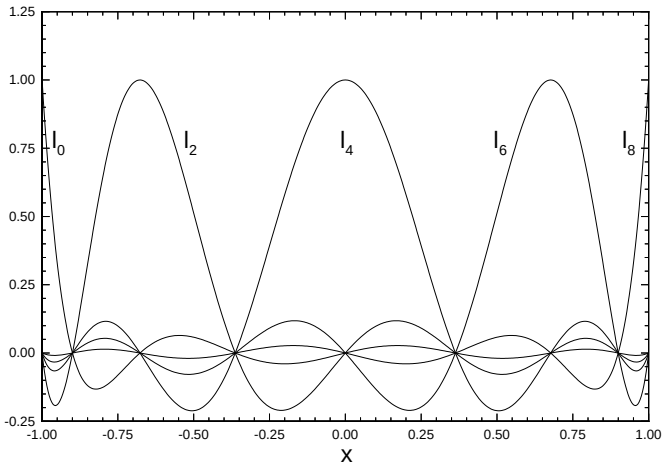

figure 6.9. The interpolating Lagrange polynomial, $l_j(x)$, based on the Legendre Gauss-Lobatto quadrature points with $N = 8$.

From Eq.(6.105) we recover the entries of the differentiation matrix, D, as

$$\mathrm{D}_{ij} = \begin{cases} -\frac{N(N+1)}{4} & i = j = 0 \\ 0 & i = j \in [1, N-1] \\ \frac{P_N(x_i)}{P_N(x_j)} \frac{1}{x_i - x_j} & i \neq j \\ \frac{N(N+1)}{4} & i = j = N \end{cases} \quad . \tag{6.110}$$

Legendre Gauss-Radau. As an alternative to the Gauss-Lobatto based interpolation, one can utilize the Gauss-Radau quadrature points, leading to the approximation

$$\mathcal{I}_N u(y) = \sum_{n=0}^{N} \tilde{u}_n P_n(y) \quad , \quad \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(y_j) P_n(y_j) v_j \quad . \tag{6.111}$$

Here the normalization constant, $\tilde{\gamma}_n$ can be found in Eq.(6.97), the weights are given in Eq.(6.82) and the quadrature points are found as the solution to Eq.(6.83).

Expressing the Gauss-Radau interpolation using the Lagrange polynomial yields

$$\mathcal{I}_N u(y) = \sum_{j=0}^{N} u(y_j) l_j(y) \quad , \tag{6.112}$$

where the Lagrange polynomial is obtained from Eq.(6.103) as

$$l_j(y) = \frac{1}{2}\frac{(1-y_j)}{N+1}\frac{P_{N+1}(y)+P_N(y)}{P_N(y_j)(y-y_j)} \quad . \tag{6.113}$$

The associated differentiation matrix can be derived by standard techniques, i.e., by differentiating the interpolation polynomials and evaluating it at the grid points.

Legendre Gauss. Consider finally the discrete Legendre expansion based on the Legendre Gauss approximation to the continuous expansion coefficients as

$$\mathcal{I}_N u(z) = \sum_{n=0}^{N} \tilde{u}_n P_n(z) \quad , \quad \tilde{u}_n = \frac{1}{\tilde{\gamma}_n}\sum_{j=0}^{N} u(z_j) P_n(z_j) u_j \quad . \tag{6.114}$$

The normalization constant, $\tilde{\gamma}_n$, being given in Eq.(6.97), the quadrature points, z_j, are found as the roots of the polynomial, Eq.(6.84), and the weights, u_j, from Eq.(6.85).

We express the interpolation as

$$\mathcal{I}_N u(z) = \sum_{j=0}^{N} u(z_j) l_j(z) \quad , \tag{6.115}$$

with the Lagrange interpolation polynomial derived from Eq.(6.104) on the form

$$l_j(z) = \frac{P_{N+1}(z)}{(z-z_j)P'_{N+1}(z_j)} \quad . \tag{6.116}$$

The differentiation matrix, D_{ij}, is obtained from Eq.(6.106) with the entries

$$\mathrm{D}_{ij} = \begin{cases} \frac{z_i}{1-z_i^2} & i = j \\ \frac{P'_{N+1}(z_i)}{(z_i-z_j)P'_{N+1}(z_j)} & i \neq j \end{cases} \quad . \tag{6.117}$$

6.3.4.2 The Discrete Chebyshev Expansion.

Methods based on Chebyshev polynomials continue to play a key role in the context of spectral methods. Their widespread use can be traced to a number of reasons. Not only are the polynomials given on a simple form but all the Gauss quadrature nodes and the associated weights are also, as shown in Sec. 6.3.2.5, given on closed form. Furthermore, as will be discussed in more detail in Chap. 9, the close relation ship between Chebyshev expansions and Fourier series allows for the fast evaluation of derivatives and interpolations.

Chebyshev Gauss Lobatto. We obtain the discrete expansion from Eq.(6.99) as

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n T_n(x) \quad , \quad \tilde{u}_n = \frac{2}{N\overline{c}_n} \sum_{j=0}^{N} \frac{1}{\overline{c}_j} u(x_j) T_n(x_j) \quad , \qquad (6.118)$$

utilizing the result of Eq.(6.98) and the weights in Eq.(6.88). In Eq.(6.118) we have introduced the parameter

$$\overline{c}_n = \begin{cases} 2 & n = 0, N \\ 1 & n \in [1, N-1] \end{cases} \quad .$$

The Chebyshev Gauss Lobatto quadrature points are given as

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

which allows us to express the computation of the interpolating polynomial at the quadrature points as

$$\mathcal{I}_N u(x_j) = \sum_{n=0}^{N} \tilde{u}_n \cos\left(\frac{\pi}{N} nj\right) \quad , \tilde{u}_n = \frac{2}{N\overline{c}_n} \sum_{j=0}^{N} \frac{1}{\overline{c}_j} u(x_j) \cos\left(\frac{\pi}{N} nj\right) \quad .$$

Hence, the discrete Chebyshev Gauss-Lobatto expansion is nothing more than a Cosine series in disguise and the expansion coefficients and the interpolation can be computed using the Fast Fourier Transform.

Using the discrete expansion, approximations to derivatives are obtained through the backward recurrence relation in Eq.(6.70). However, the equivalence between the discrete expansion and the interpolation at the quadrature points enables us to express the approximation as

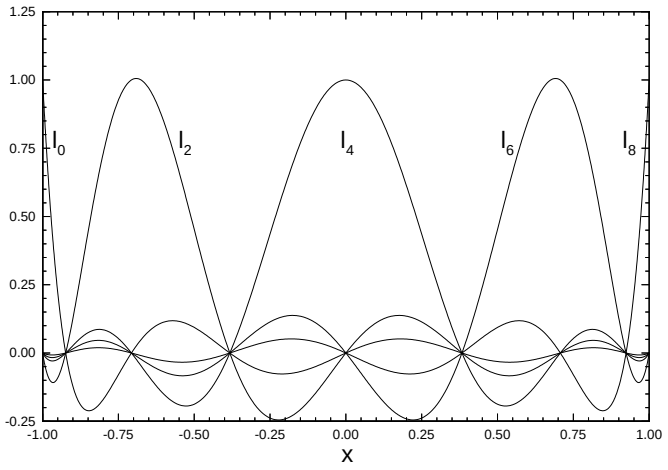

figure 6.10. The Lagrange interpolation polynomial, $l_j(x)$, based on the Chebyshev Gauss-Lobatto quadrature points with $N = 8$.

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad , \tag{6.119}$$

where the Lagrange interpolation polynomial is obtained directly from Eq.(6.102) with $\alpha = -1/2$ as

$$l_j(x) = \frac{(-1)^{N+j+1}(1-x^2)T'_N(x)}{\overline{c}_j N^2 (x - x_j)} \quad . \tag{6.120}$$

In Fig. 6.10 we show examples of the Lagrange polynomials, $l_j(x)$, based on the Chebyshev Gauss-Lobatto nodes. Comparing with the polynomials based on the Legendre Gauss-Lobatto nodes in Fig. 6.9 we observe only small differences as is a natural consequence of the grid points being qualitatively the same.

Associated with the interpolation polynomial is the differentiation matrix, D, obtained from Eq.(6.105), with the entries

$$\mathrm{D}_{ij} = \begin{cases} -\frac{2N^2+1}{6} & i = j = 0 \\ \frac{\overline{c}_i}{\overline{c}_j} \frac{(-1)^{i+j+N}}{x_i - x_j} & i \neq j \\ -\frac{x_i}{2(1-x_i^2)} & i = j \in [1, N-1] \\ \frac{2N^2+1}{6} & i = j = N \end{cases} \quad . \tag{6.121}$$

Chebyshev Gauss Radau. Using the Gauss-Radau quadrature points leads to an approximation as

$$\mathcal{I}_N u(y) = \sum_{n=0}^{N} \tilde{u}_n T_n(y) \quad , \quad \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(y_j) T_n(y_j) v_j \quad . \tag{6.122}$$

Here the normalization constant, $\tilde{\gamma}_n$, is given in Eq.(6.98), the weights are obtained from Eq.(6.91) and the quadrature points are given as

$$y_j = -\cos\left(\frac{2\pi}{2N+1} j\right) \quad , \quad j \in [0, \ldots, N] \quad .$$

Expressing the Gauss-Radau interpolation through the associated Lagrange polynomial yields

$$\mathcal{I}_N u(y) = \sum_{j=0}^{N} u(y_j) l_j(y) \quad , \tag{6.123}$$

with the Lagrange polynomial from Eq.(6.103) as

$$l_j(y) = \frac{1 - y_j}{N(2N+1)} \frac{(N+1) T_{N+1}(y) + N T_N(y)}{T_N(y_j)(y - y_j)} \tag{6.124}$$

The associated differentiation matrix can be derived by standard techniques.

Chebyshev Gauss. Let us finally also summarize the formulas for the application of the Chebyshev Gauss method. Indeed, the discrete expansion is recovered directly from Eq.(6.100) as

$$\mathcal{I}_N u(z) = \sum_{n=0}^{N} \tilde{u}_n T_n(z) \quad , \quad \tilde{u}_n = \frac{2}{c_n (N+1)} \sum_{j=0}^{N} u(z_j) T_n(z_j) \quad , \tag{6.125}$$

using Eq.(6.97), the weights given in Eq.(6.94) and c_n as defined in Eq.(6.67). Recall from Eq.(6.93) that the Chebyshev Gauss quadrature points are

$$z_j = -\cos\left(\frac{(2j+1)\pi}{2N+2}\right) \quad ,$$

indicating that also the Chebyshev Gauss discrete expansion coefficients may be obtained using a modified Fast Fourier Transform as the expansion is little more than a Cosine series.

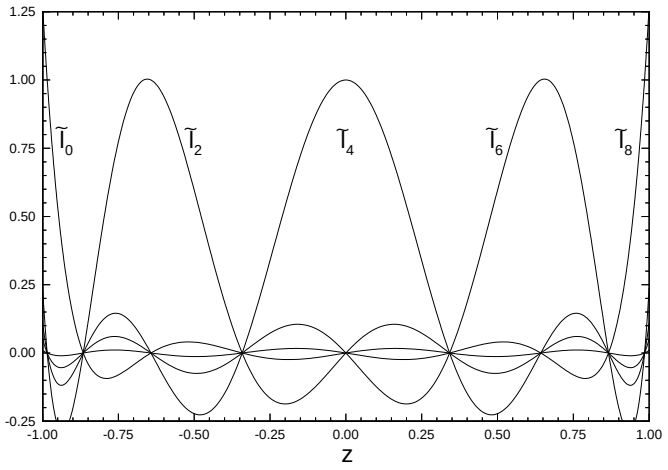

figure 6.11. The Lagrange interpolation polynomial, $l_j(z)$, based on the Chebyshev Gauss quadrature points with $N = 8$.

Using the Lagrange interpolation polynomials, we can express the Gauss interpolation as

$$\mathcal{I}_N u(z) = \sum_{j=0}^{N} u(z_j) l_j(z) \quad , \tag{6.126}$$

with the Lagrange polynomial from Eq.(6.104) as

$$l_j(z) = \frac{T_{N+1}(z)}{(z - z_j) T'_{N+1}(z_j)} \quad , \tag{6.127}$$

and the differentiation matrix, D_{ij}, from Eq.(6.106) as

$$\mathrm{D}_{ij} = \begin{cases} \frac{z_i}{2(1-z_i^2)} & i = j \\ \frac{T'_{N+1}(z_i)}{(z_i - z_j) T'_{N+1}(z_j)} & i \neq j \end{cases} \quad . \tag{6.128}$$

For the purpose of illustration we plot in Fig. 6.11 the Lagrange polynomials based on the Gauss quadrature points. We note in particular the different behavior at the boundaries of the domain as compared to polynomials based on the Gauss-Lobatto points and shown in Fig. 6.10.

6.3.5 On Lagrange Interpolation, Electrostatics, and the Lebesgue Constant.

As just discussed, one can approximate functions and their derivatives using discrete expansions as well as Lagrange interpolation polynomials. Indeed, for certain special choices of grid points and quadrature rules, these representations are identical.

However, if one is willing to leave the advantages of the dual formulation it appears only natural to consider approximations based solely on the Lagrange interpolation polynomials, i.e.,

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad ,$$

where the Lagrange interpolation polynomial, $l_j(x)$, based on the grid points, x_j, is given as

$$l_j(x) = \frac{q_N(x)}{(x - x_j) q'_N(x_j)} \quad , \quad q_N(x) = \prod_{j=0}^{N} (x - x_j) \quad . \tag{6.129}$$

Following the approach in previous sections, we recover the entries of the differentiation matrix as

$$\mathrm{D}_{ij} = \frac{1}{q'_N(x_j)} \begin{cases} q'_N(x_i)(x_i - x_j)^{-1} & i \neq j \\ \frac{1}{2} q''_N(x_i) & i = j \end{cases} \quad .$$

The only issue that requires attention to complete this approach is the specification of the grid points, x_j, which, it appears, we are completely free to choose.

To realize that care has to be exercised in this choice, let of consider a classical example.

Example 26.

Consider the analytic function, $u(x) \in C^\infty[-1, 1]$,

$$u(x) = \frac{1}{1 + 16x^2} \quad , \quad x \in [-1, 1] \quad ,$$

for which we shall seek an interpolation of the form

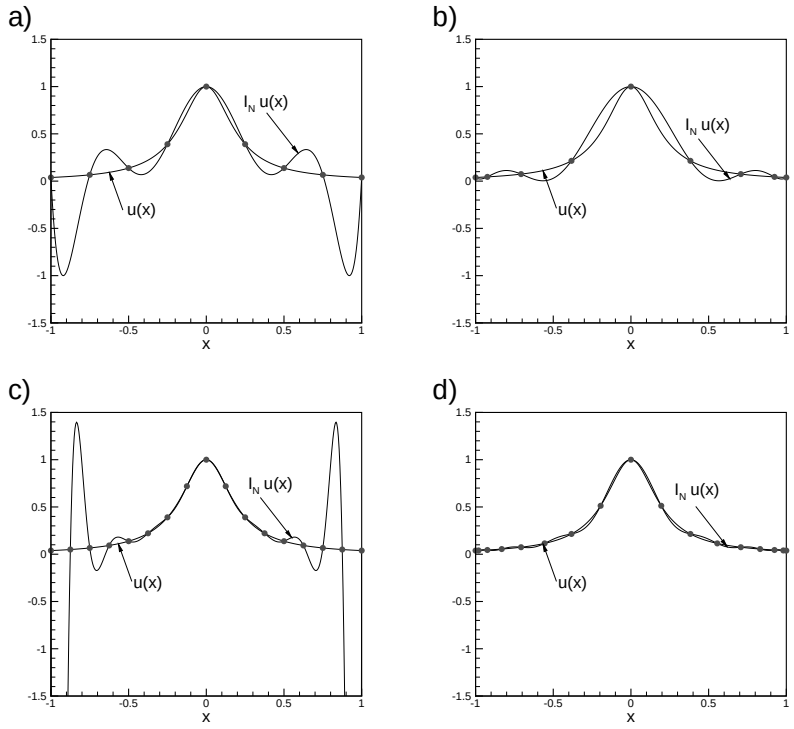

figure 6.12. a) Interpolation of $u(x)$ using $N = 8$ equidistant grid points. b) Interpolation of $u(x)$ using $N = 8$ Chebyshev-Gauss-Lobatto distributed grid points. c) Interpolation of $u(x)$ using $N = 16$ equidistant grid points. d) Interpolation of $u(x)$ using $N = 16$ Chebyshev-Gauss-Lobatto distributed grid points.

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad .$$

To illustrate the impact of choosing different grid points, x_j, to base the interpolation polynomials on, let us compare the interpolation using the equidistant points

$$x_j = \frac{2}{N} j - 1 \quad , \quad j \in [0..N] \quad ,$$

with that based on the Chebyshev-Gauss-Lobatto quadrature points

$$x_j = -\cos\left(\frac{\pi}{N}j\right) \quad , \quad j \in [0..N] \quad .$$

In Fig. 6.12 we illustrate the dramatic differences between the resulting interpolation polynomials of $u(x)$. We note that while the one based on the Chebyshev-Gauss-Lobatto grid points seems to converge as expected, the interpolation polynomials based on the equidistant grid is divergent as N increases. Clearly, the choice of the grid points matters when considering the quality of the global interpolation.

This wildly oscillatory and divergent behavior close to the limits of the domain is known as the Runge-phenomenon.

As the example illustrates, the choice of the grid points can severely impact the quality of the interpolation polynomials. Keeping in mind that we wish to use these polynomials to obtain good approximations to the spatial derivatives of $u(x)$ it is critical that we have an understanding of the properties of the grid points required to ensure a well behaved polynomial approximation.

A useful measure of the quality of the interpolation is introduced as

Theorem 48. *Assume that $u(x) \in C^0[-1,1]$ with $\mathcal{I}_N u(x)$ being the corresponding N'th order polynomial interpolation based on the grid points, x_j. Then*

$$\|u - \mathcal{I}_N u\|_\infty \leq [1 + \Lambda_N] \, \|u - p^*\|_\infty \quad ,$$

where p^* signifies the best approximating N'th order polynomial, and

$$\Lambda_N = \max_{x \in [-1,1]} \lambda_N(x) \quad , \quad \lambda_N(x) = \sum_{j=0}^{N} |l_j(x)| \quad ,$$

represents the Lebesque constant and the Lebesque function, respectively.

Proof: As $u(x) \in C^0[-1,1]$, the best approximating polynomial, p^*, exists and we immediately have

$$\|u - \mathcal{I}_N u\|_\infty \leq \|u - p^*\|_\infty + \|p^* - \mathcal{I}_N u\|_\infty \quad .$$

However, the uniqueness of p^* and $\mathcal{I}_N u$ implies that

$$\|p^* - \mathcal{I}_N u\|_\infty \leq \Lambda_N \|u - p^*\|_\infty \quad ,$$

where

$$\Lambda_N = \max_{x \in [-1,1]} \lambda_N(x) \quad , \quad \lambda_N = \sum_{j=0}^{N} |l_j(x)| \quad .$$

QED

A number of properties of the Lebesque function and the Lebesque constant are worth while emphasizing. In particular, we note that both depend only on the choice of x_j as they uniquely define the interpolation polynomials and, hence, these measures. Furthermore, it is clear from Theorem 48 that when choosing the grid points one should strive to minimize the Lebesque constant as that provides a direct measure between the actual interpolation and the best possible polynomial approximation.

As a more practical matter one can also use knowledge about the Lebesque function to come to understand how computational issues such as rounding errors can impact the accuracy of the interpolation. As an example, assume that $u_\varepsilon(x)$ represents a perturbed version of $u(x)$

$$\|u - u_\varepsilon\|_\infty \leq \varepsilon \quad .$$

The difference between the two polynomial representations are then given as

$$\|\mathcal{I}_N u - \mathcal{I}_N u_\varepsilon\|_\infty \leq \varepsilon \Lambda_N \quad .$$

Clearly, if the Lebesque constant, Λ_N, is large such that $\varepsilon \Lambda_N \ll 1$ is violated the interpolation is illposed and the impact of the rounding is very severe.

It is thus worth while looking at the behavior of the Lebesque function and the value of Λ_N for various choices of grid points. Indeed, one could hope to identify families of grid points, x_j, for which Λ_N remains a constant. A seminal result in approximation theory, however, rules the existence of such a set of grid point out [**?**, **?**]

Theorem 49. *For all sets of $N+1$ distinct grid points, $x_j \in [-1,1]$, and all values of N, the Lebesque constant is bounded as*

$$\Lambda_N \geq \frac{2}{\pi}\log(N+1) + A \quad ,$$

where

$$A = \frac{2}{\pi}\left(\gamma + \log\frac{4}{\pi}\right) \quad ,$$

in the limit of large N. *Here* $\gamma = 0.577221566..$ *represents Euler's constant.*

In other words, the Lebesque constant grows at least logarithmicly with N. This has the unfortunate consequence that for any given set of grid points there exists continuous functions for which the polynomial representations will exhibit nonuniform convergence [**?**]. On the other hand, one can also show that for any given continuous function one can always construct a set of grid points that will result in a uniformly convergent polynomial representation [**?**].

Thus, we can not in general seek one set of grid points, x_j, that will exhibit optimal behavior for all possible interpolation problems. However, the behavior of the Lebesque constant can serve as a guideline to understand whether certain families of grid points are likely to result in well behaved interpolation polynomials.

Computing the Lebesque constant for various specific choices of the grid points is an interesting, and in general, complex task and we shall not attempt to do so. It is, however, illustrative to consider some of the known results, in particular in view of the observations made in Fig. 6.12.

If we first consider interpolation based on the equidistant set of points we have [**?**]

Theorem 50. *Assume that the interpolation is based on the equidistributed set of grid points*

$$x_j = -1 + \frac{2j}{N} \quad , \quad j \in [0..N] \quad .$$

Then the corresponding Lebesque constant, Λ_N^{eq} *is bounded for* $N \geq 1$ *as*

$$\frac{2^{N-2}}{N^2} \leq \Lambda_N^{\text{eq}} \leq \frac{2^{N+3}}{N} \quad ,$$

with the asymptotic behavior given by

$$\Lambda_N^{\rm eq} \simeq \frac{2^{N+1}}{eN(\log N + \gamma)} \quad .$$

This is clearly far from optimal and for large values of N one can not expect anything meaningful from the interpolation based on the equidistant set of grid points, i.e., for $N \geq 65$, $\Lambda_N^{\rm eq} \sim 10^{16}$ and the illposedness of the interpolation is catastrophic.

Having realized that an equidistribution of grid points is far from optimal, the question arises as to whether we can identify common qualities that grid points, leading to well behaved interpolations, must share. To understand that, consider the Cauchy remainder for the interpolation

$$u(z) - \mathcal{I}_N u(z) = R_N(z) = \frac{u^{(N+1)}(\xi)}{(N+1)!} q(z) \quad ,$$

where ξ refers to some position in $[-1,1]$, $q(z)$ is defined in Eq.(6.129) although z is taken as the complex extension of x. Note that the grid points, x_j, remain real. Considering $q(z)$ it follows directly that

$$\log|q(z)| = \sum_{j=0}^{N} \log|z - x_j| = -(N+1)\phi_N(z) \quad , \tag{6.130}$$

where $\phi_N(z)$ can be interpreted as the electrostatic energy associated with $N+1$ unit mass, unit charge particles interacting according to a logarithmic potential. In the limit of $N \to \infty$ it is natural to model this as

$$\phi_\infty(z) = \lim_{N\to\infty} \frac{1}{N+1} \sum_{j=0}^{N} \log|z - x_j| = \int_{-1}^{1} \rho(x) \log|z - x|\, dx \quad ,$$

where $\rho(x)$ represents a normalized charge density, reflecting an asymptotic measure of the grid point distribution. This implies that

$$\lim_{N\to\infty} |q(z)|^{1/N} = \exp(-\phi_\infty(z)) \quad ,$$

for large values of N. Understanding the second part of the remainder, associated with the particular function being interpolated, is a bit more difficult. Using complex analysis allows one to derive that [?] that

$$\lim_{N\to\infty}\left|\frac{u^{(N+1)}(\xi)}{(N+1)!}\right|^{1/N} = \exp(\phi_\infty(z_0)) \quad ,$$

where z_0 represents the radius of the largest circle within which $u(z)$ is analytic. Hence, we recover that

$$\lim_{N\to\infty}|R(z)|^{1/N} = \exp(\phi_\infty(z_0) - \phi_\infty(z)) \quad .$$

Clearly, if $\phi_\infty(z_0) - \phi_\infty(z)$ is less than zero throughout the interval $[-1,1]$, we can expect exponential convergence in N. On the other hand, if $\phi_\infty(z_0) - \phi_\infty(z)$ exceeds zero anywhere in the unit interval we expect exponential divergence of the remainder and thus of the error.

Let us first return to the equidistant charge distribution, in which case $\rho(x) = \frac{1}{2}$ and the electrostatic energy becomes

$$\phi_\infty(z) = 1 + \frac{1}{2}\mathrm{Re}\left[|z-1|\log|z-1| - |z+1|\log|z+1|\right] \quad .$$

We first of all note that while $\phi_\infty(0) = 1$, we have that $\phi_\infty(\pm 1) = 1 - \log 2$, i.e., we should expect to see the most severe problems of divergence closer to the limits of the domain, as observed in Fig. 6.12. In fact, if we consider the u(x) in Ex. 26, it has a pole at $z_0 = \pm i/4$ and

$$\phi_\infty(\pm i/4) = 1 - \frac{1}{2}\left[\log\frac{17}{16} + \tfrac{1}{2}\arctan 4\right] \simeq 0.63823327.. \quad .$$

Thus, the remainder, and hence the polynomial representation of $u(x)$ will diverge in regions close to $[-1,1]$ as $\phi_\infty(\pm i/4) - \phi_\infty(x)$ will exceed zero. Finding the exact point of divergence yields $x \simeq \pm 0.794226...$ A close inspection of the results in Fig. 6.12 confirms this.

An equidistant distribution of grid points is evidently not a good choice for high-order polynomial interpolation of analytic functions defined on the interval. On the other hand, there is evidence in the above that the main problems close to the limits of the domain and that clustering the grid points relieves these problems as illustrated in Ex. 26.

To study this further let us begin by realizing that the continuous charge distribution leading to the Gauss-Lobatto-Chebyshev nodes used in Ex. 26 takes the form

$$\rho(x) = \frac{1}{\pi\sqrt{1-x^2}} \quad , \tag{6.131}$$

since we have

$$j = N \int_{-1}^{x_j} \frac{1}{\pi\sqrt{1-x^2}} \;\; \Rightarrow \;\; x_j = -\cos\left(\frac{\pi}{N} j\right) \;\; ,$$

where $j \in [0..N]$ is the charge number. With this, we have the corresponding electrostatic energy on the form

$$\phi_\infty(z) = -\log \frac{|z - \sqrt{z^2-1}|}{2} \;\; .$$

Inspection reveals that $\phi_\infty(z)$ are level curves associated with an ellipsoid with foci at ± 1. Furthermore, as z becomes purely real, this level curve collapses to a ridge spanning the domain of $[-1, 1]$ along which $\phi_\infty(z)$ takes on its maximum value. Hence, there are no restrictions on the position of the poles of the function, $u(x)$, being approximated as we can guarantee that $\phi_\infty(z_0) - \phi_\infty(x)$ is negative for any value of z_0. This collapse of the level curve of $\phi_\infty(z)$ to a perfectly flat ridge clearly represents the optimal choice when considering the electrostatic analysis.

Having identified grid point distributions which lead to well behaved Lagrange interpolations, let us now return to the evaluation of these interpolations in terms of the Lebesque constant.

For the symmetric Chebyshev-Gauss and Chebyshev-Gauss-Lobatto, both having the same asymptotic distribution given in Eq.(6.131), we have the Lebesque constant, Λ_N^{CG}, of the former as

$$\Lambda_N^{\mathrm{CG}} \leq \frac{2}{\pi} \log(N+1) + A + \frac{2}{\pi} \log 2 \;\; .$$

The constant A is given in Theorem 49. The Lebesque constant, Λ_N^{CGL} of the latter set of grid points is bounded as

$$\Lambda_N^{\mathrm{CGL}} \leq \Lambda_{N-1}^{\mathrm{CG}} \;\; ,$$

i.e., the Gauss-Lobatto points are, when measured by the growth of the Lebesque constant, superior to the Gauss points and very close to the theoretical optimum given in Theorem 49.

The particular characteristic of the Chebyshev distributed grid points that gives the well behaved Lagrange polynomials is the quadratic clustering of the grid points close to the ends of the domain. This quality is, however, shared among the zeros of all the ultraspherical polynomials as they all have a minimum grid size of the kind

$$\Delta x_{\min} = 1 - cN^2 \quad ,$$

where the constant c depends on the particular polynomial. This difference, however, vanishes as N approaches infinity as the grid distributions of the zeros of the ultraspherical polynomials share the limiting continuous charge distribution, Eq.(6.131).

With this result, it becomes clear that in choosing grid points well suited for polynomial interpolation, we need only impose structural conditions on the position of the grid points, e.g., close to the boundaries of the domain the grid points must cluster quadratically. This means that in terms of interpolation, the exact position of the grid points is immaterial, e.g., Legendre-Gauss-Lobatto points are as good as the Chebyshev-Gauss-Lobatto points as the basis of the Lagrange polynomials. This is also reflected in the associated Lebesque constant of the form [**?**]

$$\Lambda_N^{\mathrm{LGL}} \leq \frac{2}{\pi} \log(N+1) + 0.685.. \quad .$$

Having realized, however, that it is the quadratic behavior of the grid points close to the end of the interval that is the key to high-order convergence, there is nothing that prohibits us from seeking grid point distributions with this particular quality. Indeed, the closest to optimal grid point distribution as measured through the Lebesque constant, and for which a simple formula is known, is defined as

$$x_j^{\mathrm{ECG}} = -\frac{\cos\left(\frac{2j+1}{2N+2}\pi\right)}{\cos\left(\frac{\pi}{2N+2}\right)} \quad ,$$

known as the extended Chebyshev-Gauss grid points. These are not zeros of any ultraspherical polynomial, yet they have a Lebesque constant, Λ_N^{ECG}, bounded as

$$\Lambda_N^{\mathrm{ECG}} = \frac{2}{\pi} \log(N+1) + A + \frac{2}{\pi}\left(\log 2 - \frac{2}{3}\right) \quad ,$$

which is very close to the optimal set of grid points and for all practical purposes can serve as that.

The above discussion of the Lebesque constant and the electrostatic approach evolves, to a large extend, around the behavior of the interpolation as it depends on the choice of the grid points in the limit where N

is very large. It is, however, illustrative and useful to realize that there is a close connection between the zeros of the ultraspherical polynomials and the solution to a slightly modified version of the finite dimensional electrostatic problem, Eq.(6.130).

Let us define the electrostatic energy, $E(x_0, .., x_N)$, as

$$E(x_0, .., x_N) = -\frac{1}{2} \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \log |x_i - x_j| \ ,$$

for the $N+1$ unit mass, unit charge particles interacting according to a logarithmic potential and consider the problem as an N-body problem for which we seek the steady state, minimum energy solution if it exists. For the one given above, however, the dynamics of the problem is such that all charges would move to infinity as that would be the minimum energy solution. Let us therefore consider the slightly changed problem

$$E(p, x_0, .., x_N) = -\sum_{i=0}^{N} \left[p \log(1 - x_i^2) + \frac{1}{2} \sum_{\substack{j=0 \\ j \neq i}}^{N} \log |x_i - x_j| \right] \ . \quad (6.132)$$

This corresponds to forcing the $N+1$ charges with an exterior field corresponding to two charges, positioned at ± 1, of strength $p > 0$. If we now assume that all charges initially are positioned in the interior of $[-1, 1]$ they are confined there and nontrivial steady-state minimum energy solutions can be sought.

Considering the gradient of E, we find that a condition for minimum energy is

$$\frac{\partial E}{\partial x_i} = \frac{1}{2} \sum_{\substack{j=0 \\ j \neq i}}^{N} \frac{1}{x_i - x_j} - \frac{2x_i p}{1 - x_i^2} = 0 \ .$$

Using $q_N(x)$ as defined in Eq.(6.129) we recover

$$\frac{1}{2} \frac{q_N''(x_i)}{q_N'(x_i)} - \frac{2x_i p}{1 - x_i^2} = 0 \ ,$$

or equivalently

$$(1 - x_i^2) q_N''(x_i) - 4p x_i q_N'(x_i) = 0 \ .$$

Since this is a polynomial of order $N+1$ with $N+1$ point constraints

where it vanishes, it must, due to the definition of $q_N(x)$, be proportional to $q_N(x)$ it self. By matching coefficients we recover

$$(1-x^2)q_N''(x) - 4xpq_N'(x) + (N+1)(N+4p)q_N(x) = 0 \ . \qquad (6.133)$$

The polynomial solution, $q_N \in \mathsf{B}_{N+1}$, of Eq.(6.133) has the optimal solution to the electrostatic problem, Eq.(6.132), as it $N+1$ roots. If, however, we take $\alpha = 2p-1$ and multiply Eq.(6.132) by $(1-x^2)^\alpha$ we recover

$$\frac{d}{dx}(1-x^2)^{\alpha+1}\frac{dq_N}{dx} + (N+1)(N+2\alpha+2)(1-x^2)^\alpha q_N = 0 \ ,$$

which we may recognize as the Sturm-Liouville problem defining the general ultraspherical polynomial, $P_{N+1}^{(\alpha)}(x)$, i.e., $q_N(x) = P_{N+1}^{(\alpha)}(x)$.

Hence, a further manifestation of the close relation between grid points, well suited for interpolation, and the solution to problems of electrostatics is realized by observing that the minimum energy steady state charge distribution to the N-body problem stated in Eq.(6.132) is exactly the Gauss quadrature points of the ultraspherical polynomial, $P_N^{(2p-1)}(x)$. Using the simple relation

$$2\frac{dP_N^{(\alpha)}}{dx} = (N+1+2\alpha)P_{N-1}^{(\alpha+1)}(x) \ ,$$

we see that also the interior part of the Gauss-Lobatto points can be found as a solution to an electrostatic problem by taking $\alpha = 2(p-1)$, i.e., the Chebyshev-Gauss-Lobatto grid appears as the steady state solution for $p = 3/4$.

It should be noted that by allowing an asymmetric exterior field in Eq.(6.132) one can recover the Gauss quadrature nodes for all the Jacobi polynomials, $P_N^{(\alpha,\beta)}(x)$. See [?] for a discussion of this result.

6.4 Approximation by Laguerre Polynomials

Let us, albeit in much less detail, also discuss the use of Laguerre polynomials, introduced in Sec. 6.2.3, for the approximation of functions defined on the semi-infinite interval.

6.4.1 The Continuous Expansion.

Consider the continuous Laguerre expansion of a function, $u(x) \in L_w^2[0,\infty]$,

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n L_n(x) \quad , \tag{6.134}$$

with the expansion coefficients being

$$\hat{u}_n = (u, L_n)_{L_w^2[0,\infty]} = \int_0^{\infty} u(x) L_n(x) \exp(-x)\, dx \quad , \tag{6.135}$$

utilizing the orthonormality of the polynomials.

As for the expansions based on the ultraspherical polynomials, the first thing to address is how one utilizes the expansion of $u(x)$ on the form given in Eq.(6.134) to recover expressions for the expansion coefficients of the derivatives of $u(x)$. The required connection is established from a result similar to that of Theorem 39

Theorem 51. *Assume that the q'th derivative, $u^{(q)}(x) \in L_w^2[0,\infty]$, is expanded in Laguerre polynomials as*

$$u^{(q)}(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} L_n(x) \quad .$$

Then the representation of $u^{(q-1)}(x)$,

$$u^{(q-1)}(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q-1)} L_n(x) \quad ,$$

can be recovered up to a constant through the relation ($n > 0$)

$$\hat{u}_n^{(q-1)} = \hat{u}_n^{(q)} - \hat{u}_{n-1}^{(q)} \quad .$$

As for the ultraspherical basis, we can invert the tridiagonal operator to obtain

$$\hat{u}_n^{(q)} = - \sum_{p=n+1}^{\infty} \hat{u}_p^{(q-1)} \quad ,$$

resulting in a direct relation between $\hat{u}_n^{(q-1)}$ and $\hat{u}_n^{(q)}$.

As for the ultraspherical expansions, it is, however, more natural to consider the issue of representing the derivative from a truncated expansion. Hence, if we consider the finite expansion expansion

$$\mathcal{P}_N u^{(q-1)}(x) \simeq \frac{d^{q-1}}{dx^{q-1}} \mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n^{(q-1)} L_n(x) \quad ,$$

we recover the approximate expansion

$$\mathcal{P}_N u^{(q)}(x) \simeq \frac{d^{q}}{dx^{q}} \mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n^{(q)} L_n(x) \quad ,$$

through the backward recurrence

$$\hat{u}_{n-1}^{(q)} = \hat{u}_n^{(q)} - \hat{u}_n^{(q-1)} \quad , \tag{6.136}$$

using that $\hat{u}_N^{(q)} = 0$.

6.4.2 The Discrete Expansion.

Discussing the discrete Laguerre expansion

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n L_n(x) \quad , \quad \tilde{u}_n = \sum_{j=0}^{N} u(x_j) L_n(x_j) w_j \quad , \tag{6.137}$$

we shall first need to introduce a suitable quadrature rule as an approximation to the continuous inner product.

Among several alternatives we focus the attention on grid points, x_j, defined as the zeros of $L'_{N+1}(x)$ as well as $x_0 = 0$, i.e., in the terminology of the Sec. 6.3.2 it represents a mix between Gauss-Radau and Gauss-Lobatto points. For this one can show [?, ?, 89] that

$$\sum_{j=0}^{N} u(x_j) w_j = \int_0^\infty u(x) \exp(-x)\, dx \quad , \tag{6.138}$$

provided $u(x) \in \mathsf{B}_{2N}$, i.e., for all polynomials of order $2N$. We shall thus refer to this quadrature as a Gauss-Radau quadrature with the weights being

$$w_j = \frac{1}{N+1} \begin{cases} 1 & j = 0 \\ \left[L_{N+1}(x_j) \frac{d}{dx} L_N(x_j) \right]^{-1} & j = 1..N \end{cases} .$$

Alternative quadratures are discussed in [?, 89].

We can use the three-term recurrence relation of the Laguerre-polynomials, Eq.(6.51), to derive a Christoffel-Darboux identity for the Laguerre polynomials, hence enabling the proof of the key result

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n L_n(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad .$$

Thus, the discrete expansion represents an N'th order interpolation polynomial based on the Gauss-Radau points, x_j, which is given on closed form as

$$l_j(x) = \frac{-xL'_{N+1}(x)}{(N+1)L_{N+1}(x)(x-x_j)} \quad . \tag{6.139}$$

With this result, we have established the duality of the discrete expansion and the Lagrange interpolation polynomials based on the Laguerre-Gauss-Radau points, hence pawing the way for the approximation of derivatives using the recurrence, Eq.(6.136), as for the continues expansion, or by the definition of the differentiation matrix

$$\mathcal{I}_N \frac{du}{dx}\bigg|_{x_i} \simeq \frac{d}{dx}\mathcal{I}_N u\bigg|_{x_i} = \sum_{j=0}^{N} u(x_j) \mathrm{D}_{ij} \quad .$$

The entries of the differentiation matrix, D, are

$$\mathrm{D}_{ij} = \frac{dl_j}{dx}\bigg|_{x_i} = \frac{1}{2} \begin{cases} (N+2) & i=j=0 \\ 1 & i=j\neq 0 \\ \frac{L_{N+1}(x_i)}{L_{N+1}(x_j)} \frac{2}{x_i - x_j} & i \neq j \end{cases} \quad . \tag{6.140}$$

As for the differentiation matrices based on the ultraspherical Gauss-Radau points, there are no symmetries in the operator differentiation matrix.

Higher order derivatives can be computed by using backward recurrence repeatedly, by defining higher-order differentiation operators by repeated differentiation of Eq.(6.139) at the grid points, or by repeated application of the differentiation matrix, Eq.(6.140).

Alternative choices of grid points, leading to slightly different Lagrange polynomials and differentiation matrices, are discussed in [89, 26, 91].

6.5 Approximation by Hermite Polynomials

Let us, following the approach of the last few sections, finally discuss the use of Hermite polynomials, introduced in Sec. 6.2.4, for the representation of functions defined on the doubly infinite interval as well as the approximation of derivatives.

6.5.1 The Continuous Expansion.

Consider the continuous expansion of a function, $u(x) \in L^2_w[-\infty,\infty]$, using Hermite polynomials as

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n H_n(x) \quad , \tag{6.141}$$

where the expansion coefficients are given as

$$\hat{u}_n = \frac{1}{\gamma_n} (u, H_n)_{L^2_w[-\infty,\infty]} = \frac{1}{\gamma_n} \int_{-\infty}^{\infty} u(x) L_n(x) \exp(-x^2)\, dx \quad , \tag{6.142}$$

by the orthogonality of the polynomials. The normalization is

$$\gamma_n = \|H_n(x)\|^2_{L^2_w[-\infty,\infty]} = \sqrt{\pi} 2^n n! \quad .$$

Similar to previously discussed polynomial expansions the first thing to address in the context of spectral methods is how to use the expansion of $u(x)$, Eq.(6.141), to recover expressions for the expansion coefficients of the derivatives of $u(x)$. The required connection can be established from a result similar to that of Theorem 39 as

Theorem 52. *Assume that the q'th derivative, $u^{(q)}(x) \in L^2_w[-\infty,\infty]$, is expanded in Hermite polynomials as*

$$u^{(q)}(x) = \sum_{n=0}^{\infty} \hat{u}^{(q)}_n H_n(x) \quad .$$

Then the representation of $u^{(q-1)}(x)$,

$$u^{(q-1)}(x) = \sum_{n=0}^{\infty} \hat{u}^{(q-1)}_n H_n(x) \quad ,$$

can be recovered up to a constant using the relation ($n > 0$)

$$\hat{u}_n^{(q-1)} = \frac{1}{2n}\hat{u}_{n-1}^{(q)} \quad .$$

We note that the differentiation operator, as for the Fourier basis but contrary to the expansions based on ultraspherical or Laguerre polynomials, is diagonal. Hence, if we consider the finite expansion expansion

$$\mathcal{P}_N u^{(q-1)}(x) = \frac{d^{q-1}}{dx^{q-1}}\mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n^{(q-1)} H_n(x) \quad ,$$

we recover the approximate expansion

$$\mathcal{P}_N u^{(q)}(x) = \frac{d^q}{dx^q}\mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n^{(q)} H_n(x) \quad ,$$

through the backward recurrence

$$\hat{u}_{n-1}^{(q)} = 2n\hat{u}_n^{(q)} \quad , \tag{6.143}$$

using that $\hat{u}_N^{(q)} = 0$ due to the nature of the expansion. Note also, that the diagonality of the differentiation operator implies that truncation and differentiation commute as for the continuous Fourier expansion.

6.5.2 The Discrete Expansion.

Consider the discrete Hermite expansion

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n H_n(x) \quad , \quad \tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) H_n(x_j) w_j \quad .$$

We shall first consider a suitable quadrature rule. Among several alternatives we shall focus the attention on grid points, x_j, defined as the zeros of $H_{N+1}(x)$, i.e., in the terminology of the Sec. 6.3.2 it represents the Gauss points. For this one can show [?, 26] that

$$\sum_{j=0}^{N} u(x_j) w_j = \int_{\infty}^{\infty} u(x) \exp(-x^2)\, dx \quad , \tag{6.144}$$

provided $u(x) \in \mathsf{B}_{2N+1}$, i.e., for all polynomials of order $2N+1$. We shall thus refer to this quadrature as a Gauss quadrature with the weights

$$w_j = \sqrt{\pi} 2^{N+2} (N+1)! \left[H'_{N+1}(x_j) \right]^{-2} \quad .$$

As for ultraspherical expansion we may use the three-term recurrence relation of the Hermite-polynomials, Eq.(6.57), to derive a Christoffel-Darboux identity for the Hermite polynomials, hence establishing that

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n H_n(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad ,$$

i.e., the discrete expansion represents an N'th order interpolation polynomial based on the Gauss- points, x_j. The Lagrange interpolation polynomial is given on explicit form as

$$l_j(x) = \frac{H_{N+1}(x)}{H'_{N+1}(x_j)(x - x_j)} \quad . \tag{6.145}$$

With this result, we have recovered the duality of the discrete expansion and the Lagrange interpolation polynomials based on the Hermite-Gauss points, hence pawing the way for the approximation of derivatives using the direct connection, Eq.(6.143), as for the continuous expansion or by the definition of the differentiation matrix

$$\mathcal{I}_N \frac{du}{dx}\bigg|_{x_i} \simeq \frac{d}{dx} \mathcal{I}_N u \bigg|_{x_i} = \sum_{j=0}^{N} u(x_j) \mathrm{D}_{ij} \quad ,$$

emphasizing the introduction of the aliasing error, and with

$$\mathrm{D}_{ij} = \frac{dl_j}{dx}\bigg|_{x_i} = \begin{cases} x_i & i = j \\ \frac{H'_{N+1}(x_i)}{H'_{N+1}(x_j)} \frac{1}{x_i - x_j} & i \neq j \end{cases} \quad . \tag{6.146}$$

It follows directly from Eq.(6.53) that the Hermite polynomials are endowed with a symmetry as $H_n(x) = (-1)^n H_n(-x)$. This immediately implies that D is centro-antisymmetric, i.e., $\mathrm{D}_{ij} = -\mathrm{D}_{N-i,N-j}$ as was the case for the differentiation matrices based on the ultraspherical Gauss- or Gauss-Lobatto grid points.

Higher order derivatives can be computed by using the simple relation, Eq.(6.143), between expansion coefficients repeatedly. Alternatively, one can define higher-order differentiation operators by repeated differentiation of Eq.(6.145) at the grid points or by repeated application of the differentiation matrix, Eq.(6.146).

6.6 Approximation Theory for Smooth Functions.

In Sec. 6.2.1 we established that since the orthogonal polynomials appear as solutions to a singular Sturm-Liouville problem, we expect the polynomial expansion of smooth functions to converge at a rate depending only on the smoothness of the function being approximated. Indeed, for C^∞-functions we expect the convergence rate to be faster than any algebraic order of N, the maximum polynomial order in a way similar to that of Fourier series representations of analytic periodic functions.

In this section we shall discuss this in more detail for function of finite regularity. We aim at obtaining approximation results to confirm the accuracy in a quantitative manner. While the convergence behavior for the continuous expansion is closely related to the orthogonal polynomials themselves we need, as for the discrete Fourier expansion discussed in Sec. 4.3.2, to pay particular attention to the behavior of the discrete expansions due to the aliasing error.

The literature on approximation theory using orthogonal polynomials is vast and we will not attempt to survey all results in a rigorous manner. Rather we focus on the results of central importance in the present context of spectral methods and state more peripheral results without proof only.

6.6.1 Approximation by Ultraspherical Polynomial Expansions.

The ultraspherical polynomials holds a central position in the theory of spectral methods and it is also for expansions based on these polynomials, and in particular for expansions using Legendre and Chebyshev polynomials, that the theory is most complete. Hence, while we shall quote results of a general nature we shall pay special attention to the properties of approximations based on Legendre and Chebyshev polynomials when specific results are available for these expansions only.

6.6.1.1 The Continuous Expansion

Recall the spectral expansion

$$\mathcal{P}_N u(x) = \sum_{n=0}^{N} \hat{u}_n P_n^{(\alpha)}(x) \quad , \quad \hat{u}_n = \frac{1}{\gamma_n} \left(u, P_n^{(\alpha)} \right)_w \quad .$$

The aim is to estimate the distance between $u(x)$ and $\mathcal{P}_N u$ in the

weighted Sobolev norm, $\|\cdot\|_{H_w^m[-1,1]}$, where $w(x)$ is the weight under which $P_n^{(\alpha)}(x)$ is orthogonal.

To establishing the basic approximation results for expansions utilizing ultraspherical polynomials we rely on the approach first proposed in [12], and here extended to include the general ultraspherical polynomial basis.

Theorem 53. *For any $u(x) \in H_w^p[-1,1]$ and $p \geq 0$ there exists a constant C, independent of N, such that*

$$\|u - \mathcal{P}_N u\|_{L_w^2[-1,1]} \leq CN^{-p} \|u\|_{H_w^p[-1,1]} \quad .$$

Proof: We shall first recall Parsevals identity to obtain

$$\|u - \mathcal{P}_N u\|^2_{L_w^2[-1,1]} = \sum_{n=N+1}^{\infty} \gamma_n |\hat{u}_n|^2 \quad .$$

The expansion coefficients, $\hat{u}_n$, are given as

$$\hat{u}_n = \frac{1}{\gamma_n} \int_{-1}^{1} u(x) P_n^{(\alpha)}(x)(1-x^2)^{\alpha}\, dx \quad ,$$

where $P_n^{(\alpha)}(x)$ satisfies the Sturm-Liouville equation

$$(1-x^2)^{-\alpha} \frac{d}{dx}(1-x^2)^{\alpha+1} \frac{dP_n^{(\alpha)}}{dx} + \lambda_n P_n^{(\alpha)} = \left[\mathcal{Q} + \lambda_n\right] P_n^{(\alpha)} = 0 \quad .$$

Here $\lambda_n = n(n+2\alpha+1)$ is the eigenvalue associated with the N'th order ultraspherical polynomial, $P_n^{(\alpha)}(x)$.

Repeated integration by parts of $\hat{u}_n$ yields

$$\hat{u}_n = \frac{(-1)^p}{\gamma_n \lambda_n^m} \int_{-1}^{1} \left[\mathcal{Q}^m u(x)\right] P_n(x) w(x)\, dx \quad ,$$

due to the singularity of $\mathcal{Q}$. We have

$$|\hat{u}_n|^2 \leq C \frac{1}{\gamma_n \lambda_n^{2m}} \|\mathcal{Q}^m u\|^2_{L_w^2[-1,1]} \quad ,$$

using Cauchy-Schwarz. To bound this, we recall that

$$\mathcal{Q}u = (1-x^2)\frac{d^2u}{dx^2} - 2x(1+\alpha)\frac{du}{dx} \quad .$$

Since $|x| \leq 1$, we have

$$\|\mathcal{Q}u\|_{L^2_w[-1,1]} \leq C\|u\|_{H^2_w[-1,1]} \ ,$$

which, by induction, yields

$$\|\mathcal{Q}^m u\|_{L^2_w[-1,1]} \leq C\|u\|_{H^{2m}_w[-1,1]} \ .$$

Combining the above results gives

$$\|u-\mathcal{P}_N u\|^2_{L^2_w[-1,1]} \leq C\|u\|^2_{H^{2m}_w[-1,1]} \sum_{n=N+1}^{\infty} \lambda_n^{-2m} \leq CN^{-4m}\|u\|^2_{H^{2m}_w[-1,1]} \ .$$

Taking $p = 2m$ establishes the result. QED

To arrive at a more general result in higher norms, we shall employ a result on the error due to loss of commutation of truncation and differentiation.

Theorem 54. *For any $u(x) \in H^p_w[-1,1]$ and $|\alpha| \leq \frac{1}{2}$ there exists a constant C, independent of N, such that*

$$\left\| \mathcal{P}_N \frac{du}{dx} - \frac{d}{dx}\mathcal{P}_N u \right\|_{H^q_w[-1,1]} \leq CN^{2q-p+3/2} \|u\|_{H^p_w[-1,1]} \ ,$$

where $1 \leq q \leq p$.

Proof: We shall give the proof in detail for N being even only as that for N being odd follows from an equivalent line of arguments.

Let us first of all recall that if

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n P_n^{(\alpha)}(x) \ ,$$

then we have

$$\frac{du}{dx} = \sum_{n=0}^{\infty} \hat{u}'_n P_n^{(\alpha)}(x) \ , \quad \hat{u}'_n = (2n+2\alpha+1) \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{\infty} \hat{u}_p \ ,$$

and

$$\frac{du}{dx} = \sum_{n=0}^{\infty} \hat{u}_n \frac{dP_n^{(\alpha)}}{dx} \quad , \quad \frac{dP_n^{(\alpha)}}{dx} = \sum_{\substack{k=0 \\ k+n \text{ odd}}}^{n-1} (2k+2\alpha+1) P_k^{(\alpha)}(x) \quad .$$

This yields that

$$\mathcal{P}_N \frac{du}{dx} - \frac{d}{dx}\mathcal{P}_N u = \sum_{n=0}^{N-1} \left[\hat{u}'_n P_n^{(\alpha)} - \hat{u}_n \frac{dP_n^{(\alpha)}}{dx} \right] + \hat{u}'_N P_N^{(\alpha)} \quad .$$

Inserting the expressions for $\hat{u}'_n$ and $\left(P_n^{(\alpha)}\right)'$ gives

$$\begin{aligned} &\hat{u}'_n P_n^{(\alpha)} - \hat{u}_n \left(P_n^{(\alpha)}\right)' \\ &= \sum_{\substack{k=0 \\ k+n \text{ odd}}}^{n-1} \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{\infty} \left[(2n+2\alpha+1)\hat{u}_p P_n^{(\alpha)} - (2k+2\alpha+1)\hat{u}_n P_k^{(\alpha)} \right] \quad . \end{aligned}$$

Carefully rearranging the terms we have

$$\mathcal{P}_N \frac{du}{dx} - \frac{d}{dx}\mathcal{P}_N u = \frac{1}{2N+2\alpha+1}\hat{u}'_N \left(P_{N+1}^{(\alpha)}\right)' + \frac{1}{2N+2\alpha-1}\hat{u}'_{N-1}\left(P_N^{(\alpha)}\right)' \quad .$$

To bound this, first recall that $\hat{u}'_k$ are scalars and that $\left(P_k^{(\alpha)}\right)'$ are orthogonal polynomials. Hence, we have

$$\left\| \mathcal{P}_N \frac{du}{dx} - \frac{d}{dx}\mathcal{P}_N u \right\|^2_{L^2_w[-1,1]} = \frac{|\hat{u}'_N|^2}{(2N+2\alpha+1)^2}\Pi^2_{N+1} + \frac{|\hat{u}'_{N-1}|^2}{(2N+2\alpha-1)^2}\Pi^2_N \quad ,$$

where

$$\Pi^2_N = \left\| \left(P_N^{(\alpha)}\right)' \right\|^2_{L^2_w[-1,1]} \quad .$$

We have that

$$|\hat{u}'_N|^2 \leq \|u' - P_{N-1}u'\|^2_{L^2_w[-1,1]} \leq CN^{2(1-p)}\|u\|^2_{H^p_w[-1,1]} \quad ,$$

using Theorem 53 with $p \geq 1$ and Parsevals identity. Furthermore we have

$$\left\| \left(P_N^{(\alpha)} \right)' \right\|^2_{L^2_w[-1,1]} = \left\| \sum_{\substack{k=0 \\ k+N \text{ odd}}}^{N-1} (2k+2\alpha+1) P_k^{(\alpha)} \right\|^2_{L^2_w[-1,1]}$$
$$\leq \sum_{\substack{k=0 \\ k+N \text{ odd}}}^{N-1} (2k+2\alpha+1)^2 \gamma_k \leq CN^3$$

using orthogonality of $P_k^{(\alpha)}(x)$ and that γ_k, Eq.(6.60), is bounded in k provided $|\alpha| \leq \frac{1}{2}$.

Combining these estimates yields

$$\left\| \mathcal{P}_N \frac{du}{dx} - \frac{d}{dx} \mathcal{P}_N u \right\|^2_{L^2_w[-1,1]} \leq CN^{3-2p} \|u\|^2_{H^p_w[-1,1]} \quad .$$

Generalization to higher norms follows by first considering

$$\|u - \mathcal{P}_N u\|^2_{H^q_w[-1,1]} = \sum_{m=0}^{q} \left\| u^{(m)} - \frac{d^m}{dx^m} \mathcal{P}_N u \right\|^2_{L^2_w[-1,1]} \quad .$$

We have

$$\frac{d^m}{dx^m} \mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n \frac{d^m}{dx^m} P_n^{(\alpha)}(x) \quad ,$$

which, in combination with

$$\left| \frac{d^m}{dx^m} P_n^{(\alpha)}(x) \right| \leq CN^{2m} \left| P_n^{(\alpha)}(x) \right| \quad ,$$

from the Sturm-Liouville equations and the Poincaré inequality, yields

$$\|u - \mathcal{P}_N u\|^2_{H^q_w[-1,1]} \leq CN^{2q} \|u - \mathcal{P}_N u\|^2_{L^2_w[-1,1]} \quad ,$$

by using the Bessel inequality. Combining this with the previous result yields the theorem. QED

This result paves the way for a generalization of Theorem 53 as

Theorem 55. *For any $u(x) \in H^p_w[-1,1]$ there exists a constant C, independent of N, such that*

$$\|u - \mathcal{P}_N u\|_{H^q_w[-1,1]} \leq C N^{\sigma(q,p)} \|u\|_{H^p_w[-1,1]} \quad ,$$

where

$$\sigma(q,p) = \begin{cases} \frac{3}{2}q - p & 0 \leq q \leq 1 \\ 2q - p - \frac{1}{2} & q \geq 1 \end{cases} \quad ,$$

and $0 \leq q \leq p$.

Proof: It suffices to prove it for integer values of q and then apply interpolation between spaces [?]. Hence, for $q = 0$ we recover the result of Theorem 53. For higher norms we can use the triangle inequality to obtain

$$\left\| \frac{du}{dx} - \frac{d}{dx}\mathcal{P}_N u \right\| \leq \left\| \frac{du}{dx} - \mathcal{P}_N \frac{du}{dx} \right\| + \left\| \mathcal{P}_N \frac{du}{dx} - \frac{d}{dx}\mathcal{P}_N u \right\| \quad ,$$

where each term can be bounded by Theorems 53 and 54. QED

A couple of remarks are in place regarding these results. First of all we note that, as expected, all results confirm that the convergence rate of the continuous expansion depends solely on the regularity of the function being approximated. However, Theorem 55 also suggest that one can construct a function, $u(x) \in H^1_w[-1,1]$, for which $\mathcal{P}_N u$ converges with the prescribed rate but the derivative of the truncated approximation does not converge to $u'(x)$. This is in contrast to the Fourier case where $u(x) \in H^1_p[0, 2\pi]$ suffices to guarantee L^2-convergence of the derivative. To appreciate this difference, let us consider a simple example.

Example 27. Consider the function

$$u(x) = \frac{1}{2N+1} \left[P_{N+1} - P_{N-1} \right] \quad ,$$

where $P_n(x) = P_n^{(0)}(x)$ represents the Legendre polynomials as usual and N is assumed even. This particular choice means that

$$\frac{du}{dx} = P_N(x) \quad ,$$

using the recurrence for the Legendre polynomials, Eq.(6.25). From

Parseval's identity we immediately get

$$\|u\|^2_{L^2[-1,1]} = \frac{1}{(2N+1)^2}\left(\frac{2}{2N+3} + \frac{2}{2N-1}\right) \ ,$$

which is bounded for all values of N. Furthermore, we have that

$$\|u'\|^2_{L^2[-1,1]} = \frac{2}{2N+1} \ ,$$

i.e. $u(x) \in H^1[-1,1]$ (but not in $H^2[-1,1]$).

Let us now assume that we wish to approximate $u(x)$ by a truncated Legendre expansion as

$$\mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n P_n(x) \ .$$

The expansion coefficients follows directly from the definition of $u(x)$.

Consider the error induced by the loss of commutation as

$$\frac{d}{dx}\mathcal{P}_N u - \mathcal{P}_N u' = -\frac{1}{2N+1}\frac{dP_{N+1}}{dx} \ .$$

Using Parseval's identity we recover

$$\begin{aligned}
\left\| \frac{d}{dx}\mathcal{P}_N u - \mathcal{P}_N u' \right\|^2_{L^2[-1,1]} &= \frac{1}{(2N+1)^2} \sum_{\substack{n=0 \\ n \text{ even}}}^{N} (2n+1)^2 \gamma_n \\
&= \frac{2}{(2N+1)^2} \sum_{\substack{n=0 \\ n \text{ even}}}^{N} 2n+1 \\
&= \frac{(N+1)(N+2)}{(2N+1)^2} \ .
\end{aligned}$$

Hence, for large N we have

$$\left\| \frac{d}{dx}\mathcal{P}_N u - \mathcal{P}_N u' \right\|_{L^2[-1,1]} \simeq \frac{1}{2} \ .$$

However, since we have that

$$\frac{1}{2N+1} \leq \|u\|^2_{H^1[-1,1]} \leq \frac{2}{2N+1} \ ,$$

one recovers that

$$\left\| \frac{d}{dx}\mathcal{P}_N u - \mathcal{P}_N u' \right\|_{L^2[-1,1]} \leq C\sqrt{N}\|u\|^2_{H^1[-1,1]} \quad ,$$

which diverges as N increases. The divergence is caused by the inability of the derivative of the truncated approximation to approximate the derivative of u, as predicted in Theorem 54.

As the example confirms, the bounds in Theorem 54 and, hence, Theorem 55, are sharp and we can construct functions in $H^1[-1,1]$ for which the derivative of the truncated expansion diverges.

6.6.1.2 The Discrete Expansion.

The analysis of the properties of the discrete interpolation expansion,

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n P_n^{(\alpha)}(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad ,$$

is considerably more complex than for the continuous expansion discussed above and the theory remains incomplete As for the discrete Fourier expansion, discussed in detail in Sec. 4.3, the main reason for this added complexity is the aliasing error introduced as a consequence of the use of a discrete, grid based representation of the function being approximated. Under the assumption of sufficient smoothness, e.g., $u(x) \in H^1_w[-1,1]$, the aliasing error is reflected in

$$\tilde{u}_n = \hat{u}_n + \frac{1}{\tilde{\gamma}_n} \sum_{k>N}^{\infty} \left[P_n^{(\alpha)}, P_k^{(\alpha)} \right]_w \hat{u}_k \quad ,$$

where we recall $[\cdot,\cdot]_w$ as being the discrete inner product introduced in Sec. 6.3.3. It follows from orthogonality that

$$\|u - \mathcal{I}_N\|^2_{L^2_w[-1,1]} = \|u - \mathcal{P}_N\|^2_{L^2_w[-1,1]} + \|\mathcal{R}_N u\|^2_{L^2_w[-1,1]} \quad ,$$

where the aliasing error takes the form

$$\mathcal{R}_N u(x) = \sum_{n=0}^{N} \frac{1}{\tilde{\gamma}_n} \left(\sum_{k>N}^{\infty} \left[P_n^{(\alpha)}, P_k^{(\alpha)} \right]_w \hat{u}_k \right) P_n^{(\alpha)}(x) \quad .$$

Interchanging the two summations we recover the simple expression

$$\mathcal{R}_N u(x) = \sum_{k>N}^{\infty} \left(\mathcal{I}_N P_k^{(\alpha)} \right) \hat{u}_k \quad .$$

We can thus interpreted the aliasing error as the error induced by using the interpolation of the basis, $\mathcal{I}_N P_k^{(\alpha)}$, rather than the basis itself to represent the higher modes. As we can not distinguish between lower and higher modes at a finite grid, this introduces an error exactly as in the discrete Fourier case.

Let us attempt to arrive at a qualitative understanding of the aliasing error before we continue with more rigorous results. Among other things this will help to identify situations where the aliasing error may become significant.

Example 28. We restrict ourselves to expansions based on Legendre polynomials and the associated Gauss-quadratures. However, the main conclusions are valid for the ultraspherical polynomials and associated general quadratures also.

Let us first recall that

$$\|\mathcal{R}_N u\|^2_{L^2[-1,1]} \leq \sum_{k>N}^{\infty} \left(\mathcal{I}_N P_k\right)^2 |\hat{u}_k|^2 \leq C k^{-2p} \|u\|^2_{H^p[-1,1]} \quad ,$$

as an consequence of Theorem 53. Furthermore we have

$$\mathcal{I}_N P_k = \sum_{n=0}^{N} \tilde{p}_n P_n \quad , \quad \tilde{p}_n = \frac{1}{\gamma_n} \sum_{j=0}^{N} P_k(z_j) P_n(z_j) v_j \quad ,$$

where z_j and u_j are the Legendre-Gauss points and weights, respectively, as discussed in Sec. 6.3.2.4. Let us also recall that while the Gauss quadrature is exact for all $f \in \mathsf{B}_{2N+1}$, using the quadrature on general functions introduces an error as

$$E_N(f) = \int_{-1}^{1} f \, dx - \sum_{j=0}^{N} f(z_j) u_j = \frac{2^{2N+3}}{2N+3} \frac{\Gamma(N+2)^4}{\Gamma(2N+3)} \frac{d^{2N+2}}{dx^{2N+2}} f(\xi) \quad ,$$

provided $f \in C^{2N+2}[-1,1]$. Here $\xi \in [-1,1]$. The latter expression is a standard error term for Gauss quadratures and can be found in e.g. [**?**].

Utilizing the orthogonality of P_n and P_k – recall $n \leq N < k$ – we

recover

$$\tilde{p}_n = -\frac{1}{\gamma_n} E_N(P_n P_k) = -\frac{2N+1}{2} \frac{2^{2N+3}}{2N+3} \frac{\Gamma(N+2)^4}{\Gamma(2N+3)} \frac{d^{2N+2}}{dx^{2N+2}} P_n(\xi) P_k(\xi) \ ,$$

for $n+k > 2N+1$ and zero otherwise. Assuming N large and using Stirling's formula for $\Gamma(N)$, we recover the bound

$$|\tilde{p}_n| \leq C \exp(-N) N^N \left\| \frac{d^{2N+2}}{dx^{2N+2}} P_n(\xi) P_k(\xi) \right\|_{L^2[-1,1]} .$$

The latter term is be bounded as

$$\left\| \frac{d^{2N+2}}{dx^{2N+2}} P_n(\xi) P_k(\xi) \right\|_{L^2[-1,1]} \leq C(n+k)^{4N+4} \left\| P_n(\xi) P_k(\xi) \right\|_{L^2[-1,1]} .$$

With this, one derives a bound on the aliasing error as

$$\left\| \mathcal{R}_N u \right\|_{L^2[-1,1]} \leq C \exp(-N) N^{\alpha(N)} \left(\sum_{k>N}^{\infty} k^{\beta(N)-p} \right) \|u\|_{H^p[-1,1]},$$

where C is independent of N while $\alpha(N)$ and $\beta(N)$ represent linear algebraic expressions in N. Clearly, for u being sufficiently smooth, we can always find p such that the sum is finite, i.e., $p > \beta(N) + 1$, and, furthermore, we can find a larger p, i.e., $p > \alpha(N) + \beta(N) + 1$, sufficient to guarantee exponential decay of the aliasing error as N increases. In other words, for smooth well resolved functions the aliasing error is not expected to affect the convergence properties of the discrete expansion in any significant way.

Quantitative results, the proof of which are highly technical and omitted in the following, for the interpolation using ultraspherical Gauss- and Gauss-Lobatto nodes are discussed in detail in [?]. In particular we quote the following result

Theorem 56. *Assume that $u \in H^p_w[-1,1]$ with $p > \frac{1}{2}\max(1, 1+\alpha)$ where $\mathcal{I}_N u$ is constructed using ultraspherical polynomials, $P^\alpha_n(x)$, with $|\alpha| \leq 1$. Then there exists a constant, C, depending on α and p but not on N such that*

$$\|u - \mathcal{I}_N u\|_{L^2_w[-1,1]} \leq C N^{-p} \|u\|_{H^p_w[-1,1]}.$$

This holds for Gauss and Gauss-Lobatto based interpolations.

Hence, the result confirms that for well resolved smooth functions the qualitative behavior of the continuous and the discrete expansion is similar for all practical purposes.

To be more specific and reach results in higher norms we leave the general ultraspherical expansion and consider discrete expansions based on Legendre and Chebyshev polynomials and the associated Gauss-type quadrature points.

Results for the Discrete Legendre Expansion. Consider first the discrete Legendre expansion

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n P_n(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad .$$

The most general result is given as [11]

Theorem 57. *For any $u(x) \in H^p[-1,1]$ with $p > \frac{1}{2}$ and $0 \leq q \leq p$, there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{I}_N u\|_{H^q[-1,1]} \leq C N^{2q-p+1/2} \|u\|_{H^p[-1,1]} \quad .$$

The proof of this result, again somewhat technical, can be found in [11]. However, we note that for $q = 0$, the result is suboptimal for Gauss and Gauss-Lobatto based interpolation and Theorem 56 presents a sharper bound.

Again we observe that for functions with sufficient smoothness, i.e., for any $u(x) \in H^p[-1,1]$ with $p \geq 1$, we recover results similar to those for the continuous expansion stated in Theorem 55, although the exact convergence rate is lowered by 1. However, for smooth functions the aliasing error does not modify the convergence rate of the discrete expansion as compared to the continuous expansion in any substantial way and spectral convergence for analytic functions is maintained [**?**]

A straightforward combination of Theorems 56 and 57 yields an estimated of the commutation error

$$\left\| \mathcal{I}_N \frac{du}{dx} - \frac{d}{dx} \mathcal{I}_N u \right\|_{L^2[-1,1]} \leq N^{\frac{5}{2}-p} \|u\|_{H^p[-1,1]} \quad ,$$

which should be compared with the result of Theorem 54, confirming the

modification of the convergence rate but the persistence of the spectral convergence.

Results for the Discrete Chebyshev Expansion. The behavior of the discrete Chebyshev expansion

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n T_n(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad ,$$

is most easily understood by utilizing the close connection between the Chebyshev expansion and the Fourier/Cosine series. Indeed, if we introduce the transformation, $x = \cos(\theta)$, the aliasing error can be expressed on a simple form since

$$\tilde{p}_n = \frac{1}{\tilde{\gamma}_n} [T_k(x), T_n(x)]_N = \begin{cases} 1 \ k = 2Np \pm n \ p = 0, \pm 1, \pm 2 \ldots \\ 0 \ \text{ else} \end{cases} \quad ,$$

as a direct consequence of the discrete orthogonality of the exponential function and hence the cosine function. This immediately yields that

$$\mathcal{R}_N u = \sum_{k>N}^{\infty} (\mathcal{I}_N T_k) \hat{u}_k = \sum_{\substack{p=-\infty \\ p \neq 0}}^{p=\infty} (\hat{u}_{2Np+n} + \hat{u}_{2Np-n}) T_n(x) \quad .$$

Comparing with the Lemma 5 highlights the close connection between the Chebyshev and the Fourier expansion.

In light of this it comes as no surprise that

Theorem 58. *For any $u(x) \in H^p[-1,1]$ with $p > \frac{1}{2}$ and $0 \leq q \leq p$, there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{I}_N u\|_{H^q_w[-1,1]} \leq C N^{2q-p} \|u\|_{H^p_w[-1,1]} \quad .$$

Proof: For $q = 0$ the result follows by evenly extending a function $u(\cos(\theta))$, $\theta \in [0, \pi]$ around $\theta = \pi$ to cover the whole domain $[0, 2\pi]$. In this case the Fourier interpolation becomes a cosine series and we can thus take the result from Theorem 11 to obtain

$$\|u - \mathcal{I}_N u\|_{L^2_w[-1,1]} = \frac{1}{\sqrt{2}} \|u - \mathcal{I}_N u\|_{L^2[0,2\pi]} \leq C N^{-p} \|u\|_{H^p_w[-1,1]} \quad .$$

The extension to higher Sobolev norms follows with equal ease by realizing that

$$\|u - \mathcal{I}_N u\|_{H^1_w[-1,1]} \leq \|u - \mathcal{I}_N u\|^2_{L^2_w[-1,1]} + \|u' - (\mathcal{I}_N u)'\|_{L^2_w[-1,1]} \ ,$$

where the latter term can be bounded by an inverse inequality as

$$\|u' - (\mathcal{I}_N u)'\|_{L^2_w[-1,1]} \leq N^2 \|u - \mathcal{I}_N u\|^2_{L^2_w[-1,1]} \ .$$

The general result follows by induction. QED

Let us finally quote the L^∞-error for the discrete Chebyshev expansion as [11]

Theorem 59. *For any $u(x) \in H^p_w[-1,1]$ with $p > \frac{1}{2}$, there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{I}_N u\|_{L^\infty[-1,1]} \leq C N^{1/2-p} \|u\|_{H^p_w[-1,1]} \ .$$

6.6.2 Approximation by Laguerre Polynomial Expansions.

For reference and completeness of the discussion let us also briefly summarize some central approximation results for expansions of functions, $u(x) \in L^2_w[0,\infty]$, using continuous and discrete Laguerre expansions.

In the spirit of the analysis for ultraspherical expansions, let us first consider the behavior of the continuous expansion, Eqs. (6.134)-(6.135). From [90] we have the result

Theorem 60. *For any $u(x) \in H^p_w[0,\infty]$, $p > 0$ there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{P}_N u\|_{H^q_w[0,\infty]} \leq C N^{q-p/2} \|u\|_{H^p_w[0,\infty]} \ ,$$

where $0 < q < p$.

This is the most general result needed in the context of spectral methods as we have seen in Sec. 6.6.1. A significant difference is the lower convergence rate for p finite, i.e., $\sqrt{N}^p$ rather than N^p as we have seen previously. In particular, it follows immediately that

$$\left\| \frac{d}{dx}\mathcal{P}_N u - \mathcal{P}_N \frac{du}{dx} \right\|_{L^2_w[0,\infty]} \leq CN^{1-p/2}\|u\|_{H^p_w[0,\infty]} \ ,$$

for $p \geq 1$. As for the ultraspherical polynomials this indicates that one can construct functions, $u(x) \in H^1_w[0,\infty]$, for which the derivative of the truncated approximation is divergent as exemplified for the ultraspherical expansion in Ex. 27.

A general idea of when the Laguerre expansion will work well can be obtained by viewing the Laguerre expansion as a regular L^2 expansion in the function

$$\psi_n(x) = L_n(x)\exp(-x/2) \ .$$

Clearly, for functions that exhibit a decay dramatically different from $e^{-x/2}$ the expansion is less natural and can be expected to exhibit very slow convergence. On the other hand, for simple rapidly decaying functions the Laguerre expansion can work well [90, 92].

For the Gauss-Radau based discrete expansion, Eq.(6.137), discussed in Sec. 6.4.2 a result similar to the one above has been proven in [90] as

Theorem 61. *For any $u(x) \in H^p_w[0,\infty]$, $p > \frac{1}{2}$ there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{I}_N u\|_{H^q_w[0,\infty]} \leq CN^{q-p/2+1/2}\|u\|_{H^p_w[0,\infty]} \ ,$$

where $0 \leq q \leq \frac{1}{2} < p$.

This indicates that we loose at most $\sqrt{N}$ due to aliasing.

6.6.3 Approximation by Hermite Polynomial Expansions.

Let us finally summarize the key results for the expansion of functions, $u(x) \in L^2_w[-\infty,\infty]$, using continuous and discrete Hermite expansions.

As in the previous sections, let us begin by considering the behavior of the continuous expansion, Eqs. (6.141)-(6.142). From [26] we have

Theorem 62. *For any $u(x) \in H^p_w[-\infty,\infty]$, $p > 0$ there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{P}_N u\|_{L^2_w[-\infty,\infty]} \leq CN^{-p/2}\|u\|_{H^p_w[-\infty,\infty]} \quad .$$

As for the Laguerre expansion we can appreciate the limitations of the Hermite expansion by viewing it as an L^2 expansion in the Hermite function

$$\psi_n(x) = H_n(x)\exp(-x^2/2) \quad .$$

If, indeed, the function exhibit a decay dramatically different from $e^{-x^2/2}$ the expansion is less natural and can be expected to exhibit very slow convergence On the other hand, for simple rapidly decaying functions the Hermite expansion can work well as examples in [**?**] verifies.

For the Gauss based discrete expansion, Eq.(6.144), discussed in Sec. 6.5.2 a result is given in [26] on the form

Theorem 63. *For any $u(x) \in H^p_w[-\infty,\infty]$, $p > 1$ there exists a positive constant, C, independent of N, such that*

$$\|u - \mathcal{I}_N u\|_{L^2_w[-\infty,\infty]} \leq CN^{-p/2+1/2}\|u\|_{H^p_w[-\infty,\infty]} \quad .$$

As for the Laguerre expansion, this indicates that we loose at most $\sqrt{N}$ due to aliasing.

Exercises

1. Assume that $u(x)$ has a point of discontinuity at x_0 and construct $v(x)$ such that it equals $u(x)$ outside the interval of $[x_0 - \delta, x_0 + \delta]$. Inside this interval, we construct $v(x)$ by a line segment connecting the point of $(x_0 - \delta, u(x_0 - \delta))$ with $(x_0 + \delta, u(x_0 + \delta))$.
   Prove that

$$\lim_{\delta \to 0} \|u(x) - v(x)\|_{L^2[0,1]} = 0 \quad ,$$

   i.e. prove that any function, $u(x) \in L^2[-1, 1]$, can be approximation arbitrarily well in mean by a $C^0[-1, 1]$ function.

2. Assume that $u(x) \in L^2_w[-1, 1]$ and that it is expanded in an orthogonal and complete basis as

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n \phi_n(x) \quad .$$

   Prove that if

$$\mathcal{P}_N u = \sum_{n=0}^{N} \hat{u}_n \phi_n(x) \quad .$$

   is convergent in the mean then

$$\int_{-}^{1} u^2(x) w(x)\, dx = (u, u)_w = \sum_{n=0}^{\infty} \gamma_n \hat{u}_n^2 \quad .$$

3. Prove that if $\phi_n(x)$ is an nth order polynomial, then the only solution to

$$-(1 - x^2)\phi_n''(x) + ((\alpha + \beta + 2)x + \alpha + \beta)\phi_n'(x) = \lambda_n \phi_n(x) \quad ,$$

   is

$$\lambda_n = n(n + \alpha + \beta + 1) \quad .$$

4. Prove that the recurrence relation in Theorem 32 has the coefficient given in Eq.(6.19).

   HINT: Prove first that

$$\begin{aligned} \tilde{a}_{n-1,n}^{(\alpha,\beta)} &= \frac{n + \alpha + \beta + 1}{n + \alpha + \beta} a_{n-2,n-1}^{(\alpha+1,\beta+1)} \quad , \\ \tilde{a}_{n,n}^{(\alpha,\beta)} &= a_{n-1,n-1}^{(\alpha+1,\beta+1)} \quad , \\ \tilde{a}_{n+1,n}^{(\alpha,\beta)} &= \frac{n + \alpha + \beta + 1}{n + \alpha + \beta + 2} a_{n,n-1}^{(\alpha+1,\beta+1)} \quad . \end{aligned}$$

5. Combine Eq.(6.40) and Eq.(6.45) to prove Lemma 8.

6. Prove Theorem 51.

7. Derive Eq.(6.139) using the definition of the Lagrange polynomials and the properties of the Laguerre polynomials.

8. Plot the Legendre polynomials, $P_n(x)$, for $n = 0 - 5$.

9. Plot the Chebyshev polynomials, $T_n(x)$, for $n = 0 - 5$.

10. Compute the discrete Chebyshev expansion, based on the Chebyshev Gauss Lobatto points, of the following functions ($x \in [-1, 1]$)

    (a) $u(x) = x^7$
    (b) $u(x) = |\sin(\pi x)|$
    (c) $u(x) = \text{sign}(x)$

    and compute the L^∞ error when increasing N, the length of the expansion.

11. (Continued) Repeat the computations using the Chebyshev Gauss points. Do you see a difference ?

12. Compute the derivative of the functions ($x \in [-1, 1]$)

    (a) $u(x) = x^7$
    (b) $u(x) = (1.1 - x)^{-1}$
    (c) $u(x) = |\sin(\pi x)|$

    using the differentiation matrix based on the Chebyshev Gauss-Lobatto points. Compute the L^∞ and L^2_w-error for increasing length N of the expansion.

13. (Continued) Repeat the computation using the backward recurrence. Do you see a difference ?

14. Derive the entries for the 2nd order Chebyshev differentiation matrix based on the Gauss-Lobatto points.

15. (Continued) Compare these entries to those computed by matrix-multiplication for increasing values of N (take N larger than 100).

16. (Continued) It has been suggested that one should initialize the diagonal of the differentiation matrices as

$$\mathrm{D}_{ii} = -\sum_{\substack{j=0 \\ j \neq i}}^{N} \mathrm{D}_{ij} \quad .$$

    Explain why this makes sense and verify whether it has any impact on the accuracy of the entires of $\mathrm{D}^{(2)}$ when these are computed from D.

7

Polynomial Spectral Methods

Having identified the orthogonal polynomials as a suitable choice of basis on which to base the development of spectral methods, we can now return to the actual development of such methods for solving partial differential equations.

As in Chapter 5, where we discussed schemes based on Fourier expansions, we shall discuss the issue of how to satisfy the equation, i.e., we shall concern ourselves with the details of the Galerkin, the tau and the Collocation methods when using polynomial expansions. The key difference between the previous discussion is that we need to consider methods for enforcing nontrivial boundary conditions. For the sake of simplicity, we shall also restrict ourselves to problems involving smooth solutions and return to the special issues related to non-smooth problems in the following chapter.

We consider the construction of schemes for the problem

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= \mathcal{L}u(x,t) \ , \quad x \in [a,b] \ , \ t \geq 0 \ , \\ \mathcal{B}_- u(a,t) &= 0 \ , \qquad t \geq 0 \\ \mathcal{B}_+ u(b,t) &= 0 \ , \qquad t \geq 0 \\ u(x,0) &= f(x) \ , \qquad x \in [a,b] \ , \ t = 0 \ . \end{aligned} \tag{7.1}$$

where $\mathcal{B}_\pm$ represent the boundary operator at $x = a$ and $x = b$, respectively, with a and/or b possibly being unbounded.

For ease of exposure we restrict much of the discussion to methods based on Legendre or Chebyshev expansions on the finite interval $[-1, 1]$. However, all results extend in a straightforward manner to schemes based

on ultraspherical polynomials in general. We also briefly discuss special concerns related to the construction of methods using Laguerre or Hermite expansions for problems on unbounded intervals.

7.1 Polynomial Methods on the Bounded Interval

The formulation of spectral methods for solving initial boundary value problems involve choosing the polynomial space, B_N, in which the approximate solution is sought, and the specification of the projection operator, $\mathcal{P}_N$, detailing how the equation is satisfied. As discussed in Chapter 3 these requirements split the development of the methods into three distinct categories to which we shall attend in what follows.

7.1.1 Galerkin Methods.

In the polynomial Galerkin method we seek solutions, $u_N(x,t) \in \mathsf{B}_N$, to Eq.(7.1) of the form

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t)\phi_n(x) \ ,$$

where $\phi_n(x)$ represents a polynomial basis and the polynomial space, B_N, in which we seek solutions is given as

$$\mathsf{B}_N = \text{span}\left\{\phi_n(x) \in \text{span}\left\{x^k\right\}_{k=0}^{n} | \mathcal{B}_-\phi_n(-1) = 0 \ , \ \ \mathcal{B}_+\phi_n(1) = 0\right\}_{n=0}^{N} \ .$$

The $N+1$ equations for the unknown expansion coefficients, $\hat{u}_n(t)$, are obtained from Eq.(7.1) by requiring the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \mathcal{L}u_N(x,t) \ ,$$

is orthogonal to B_N in $L^2_w[-1,1]$. Using the testfunctions, $\psi_k(x) = \phi_k(x)/\gamma_k$, this yields

$$\forall k \in [0,N] : \ \sum_{n=0}^{N} \mathrm{M}_{kn}\frac{d\hat{u}_n}{dt} = \sum_{n=0}^{N} \mathrm{S}_{kn}\hat{u}_n(t) \ ,$$

Hence, the basis-functions, $\phi_n(x)$, must satisfy the boundary conditions individually as there is no other mechanism by which to impose the boundary conditions.

We have introduced the mass-matrix, M, with the entries

$$\forall k, n \in [0, N] : \ \mathrm{M}_{kn} = \frac{1}{\gamma_k} \int_{-1}^{1} \phi_k(x) \phi_n(x) w(x) \, dx \ ,$$

where $w(x) \in L^1[-1, 1]$ signifies a weight-function. Likewise, we have the entries of the stiffness-matrix, S, as

$$\forall k, n \in [0, N] : \mathrm{S}_{kn} = \frac{1}{\gamma_k} \int_{-1}^{1} \phi_k(x) \mathcal{L} \phi_n(x) w(x) \, dx \ .$$

The basis, $\phi_n(x)$, is usually constructed from a linear combination of $P_n^{(\alpha)}(x)$ to ensure that the boundary conditions are satisfied for all $\phi_n(x)$. In this case it is natural to choose the weight-function, $w(x)$, such that $(P_n^{(\alpha)}, P_k^{(\alpha)})_w = \gamma_n \delta_{nk}$, thereby completing the specification of the scheme.

We consider in the following a few examples of polynomial Galerkin methods to illustrate the steps needed to formulate such methods.

Example 29. Consider first the linear hyperbolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \ ,$$

where we assume that $u(x, t) \in L_w^2[-1, 1]$ and the solution is subject to the boundary condition

$$u(1, t) = 0 \ ,$$

with the initial condition $u(x, 0) = f(x)$.

We wish to solve the problem using a Chebyshev Galerkin method. However, as $T_n(1) = 1$ we must modify the basis to obtain a Galerkin formulation. There are indeed many ways of specifying polynomials that satisfy the boundary conditions and spans B_N and a viable choice could be

$$\phi_n(x) = T_n(x) - 1 \ .$$

Thus, we seek solutions, $u_N(x, t) \in \mathsf{B}_N$, of the form

$$u_N(x, t) = \sum_{n=1}^{N} \hat{u}_n(t) \phi_n(x) = \sum_{n=1}^{N} \hat{u}_n(t) \left(T_n(x) - 1 \right) \ .$$

Note that the sum is from $n = 1$ rather than $n = 0$ since $\psi_0(x) = 0$. We now require that the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} \quad ,$$

is orthogonal to $\phi_n(x)$ in $L^2_w[-1,1]$ as

$$\forall k \in [1,N] : \ \frac{2}{\pi} \int_{-1}^{1} R_N(x,t)\phi_k(x) \frac{1}{\sqrt{1-x^2}}\, dx = 0 \quad .$$

We have chosen $w(x)$ as the weight for which $T_n(x)$ is orthogonal to simplify the scheme. While this is natural it is not necessary.

This yields the Chebyshev Galerkin scheme

$$\forall k \in [1,N] : \ \sum_{n=1}^{N} \mathrm{M}_{kn} \frac{d\hat{u}_n}{dt} = \sum_{n=1}^{N} \mathrm{S}_{kn} \hat{u}_n(t) \quad ,$$

where the mass-matrix has the entries

$$\mathrm{M}_{kn} = \frac{2}{\pi} \int_{-1}^{1} \left(T_k(x) - 1\right) \left(T_n(x) - 1\right) \frac{1}{\sqrt{1-x^2}}\, dx = 2 + \delta_{kn} \quad .$$

Computing the entries of the stiffness-matrix requires a little more work. Indeed, the entries are given as

$$\forall k,n \in [1,N] : \ \mathrm{S}_{kn} = \frac{2}{\pi} \int_{-1}^{1} \left(T_k(x) - 1\right) \frac{dT_n(x)}{dx} \frac{1}{\sqrt{1-x^2}}\, dx \quad .$$

Using Eq.(6.31) we obtain

$$\frac{dT_n(x)}{dx} = 2n \sum_{\substack{p=0 \\ p+n \text{ odd}}}^{n-1} \frac{T_p(x)}{c_p} \quad ,$$

where $c_0 = 2$ and $c_p = 1$ otherwise as usual. Introducing this into the expression for the stiffness-matrix yields

$$\mathrm{S}_{kn} = \frac{2}{\pi} \int_{-1}^{1} \left(T_k(x) - 1\right) 2n \sum_{\substack{p=0 \\ p+n \text{ odd}}}^{n-1} \frac{T_p(x)}{c_p} \frac{1}{\sqrt{1-x^2}}\, dx$$

$$= 2n \sum_{\substack{p=0 \\ p+n \text{ odd}}}^{n-1} (\delta_{kp} - \delta_{0p}) \quad .$$

This completes the specification of the Chebyshev Galerkin scheme and we have N equations for the N unknowns, $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_N)^T$, on the form

$$\frac{d\hat{\boldsymbol{u}}}{dt} = \mathrm{M}^{-1}\mathrm{S}\hat{\boldsymbol{u}}(t) \quad ,$$

with the initial conditions given as

$$\forall n \in [1, N] : \ \hat{u}_n(0) = \frac{2}{\pi} \int_{-1}^{1} f(x) \, (T_n(x) - 1) \, \frac{1}{\sqrt{1 - x^2}} \, dx \quad .$$

The formulation of the Chebyshev Galerkin is a bit cumbersome and involves, in the general case, the inversion of a mass-matrix, which depends on the specification of the basis. The mass-matrix is, however, symmetric and positive definite. The latter property follows by considering a general non-zero N-vector, $\boldsymbol{u}$, and using the definition of M to obtain

$$\begin{aligned} \boldsymbol{u}^T \mathrm{M} \boldsymbol{u} &= \sum_{i,j=0}^{N} u_i u_j \mathrm{M}_{ij} \\ &= \sum_{i,j=0}^{N} u_i u_j \int_{-1}^{1} \phi_i(x) \phi_j(x) w(x) \, dx \\ &= \int_{-1}^{1} \left(\sum_{i=0}^{N} u_i \phi_i(x) \right)^2 w(x) \, dx \geq 0 \quad , \end{aligned}$$

since $\|\cdot\|_{L^2_w[-1,1]}$ is a norm. Thus, the inverse of M is guaranteed to exist.

Example 30. Consider the linear parabolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad ,$$

where we assume that $u(x,t) \in L^2[-1,1]$ and the solution obeys the

boundary conditions

$$u(-1,t) = u(1,t) = 0 \ ,$$

with the initial conditions $u(x,0) = f(x)$.

We wish to use a Legendre Galerkin methods for solving this problem. However, since $P_n(1) = 1$ we must again modify the basis to derive the Galerkin scheme. Any choice of polynomials that satisfy the boundary conditions and spans B_N is acceptable. However, a convenient choice could be

$$\phi_n(x) = P_{n+1}(x) - P_{n-1}(x) \ ,$$

which has the desired property, $\phi_n(\pm 1) = 0$. An alternative option is

$$\phi_n(x) = \begin{cases} P_n(x) - P_0(x) & n \text{ even} \\ P_n(x) - P_1(x) & n \text{ odd} \end{cases} \ .$$

Choosing the former, we seek solutions, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=1}^{N-1} \hat{u}_n(t)\phi_n(x) = \sum_{n=1}^{N-1} \hat{u}_n(t)\left(P_{n+1}(x) - P_{n-1}(x)\right) \ ,$$

and require the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \frac{\partial^2 u_N}{\partial x^2} \ ,$$

to be L^2-orthogonal to $\phi_k(x)$ as

$$\forall k \in [1, N-1] : \ \frac{2k+1}{2} \int_{-1}^{1} R_N(x,t)\phi_k(x)\, dx = 0 \ .$$

This yields the Legendre Galerkin scheme

$$\forall k \in [1, N-1] : \ \sum_{n=1}^{N-1} \mathrm{M}_{kn} \frac{d\hat{u}_n}{dt} = \sum_{n=1}^{N-1} \mathrm{S}_{kn}\hat{u}_n(t) \ ,$$

with the mass-matrix having the entries

$$\mathrm{M}_{kn} = \frac{2k+1}{2} \int_{-1}^{1} \phi_k(x)\phi_n(x)\, dx$$

$$= \frac{2(2k+1)^2}{(2k-1)(2k+3)}\delta_{kn} - \frac{2k+1}{2k+3}\delta_{k,n+2} - \frac{2k+1}{2k-1}\delta_{k,n-2}$$

using the orthogonality of the Legendre polynomials. Note that the mass-matrix is only tridiagonal, i.e., the computation of M^{-1} is straightforward.

The computation of the entries of the stiffness matrix yields

$$\forall k,n \in [1,N-1]: \ \mathrm{S}_{kn} = \frac{2k+1}{2}\int_{-1}^{1} \phi_k(x)\frac{d^2\phi_n(x)}{dx^2}\,dx \ .$$

These entries can either be derived by using the properties of the Legendre polynomials or use a Gauss quadrature of sufficiently high accuracy.

This yields a complete scheme for the $N-1$ equations with $N-1$ unknowns, $\hat{\boldsymbol{u}} = (\hat{u}_1,\ldots,\hat{u}_{N-1})^T$, as

$$\frac{d\hat{\boldsymbol{u}}}{dt} = \mathrm{M}^{-1}\mathrm{S}\hat{\boldsymbol{u}}(t) \ ,$$

with the initial conditions given as

$$\forall n \in [1,N-1]: \ \hat{u}_n(0) = \frac{2n+1}{2}\int_{-1}^{1} f(x)\left(P_{n+1}(x) - P_{n-1}(x)\right)\,dx \ .$$

Similar to the Chebyshev Galerkin scheme for the wave equation, we arrive at a method that requires the inversion of a matrix, albeit in this case it is only tridiagonal. The complexities associated with the polynomial Galerkin methods result from the complicated relationships between the orthogonal polynomials and their derivatives, and the requirement that $\phi_n(x)$ satisfy the boundary conditions individually. One could of course simplify matters a bit by using quadratures of sufficient accuracy to compute the entries of the operators.

Example 31. Consider Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \ ,$$

where $u(x,t) \in L^2_w[-1,1]$ with the homogeneous boundary conditions

$$u(-1,t) = u(1,t) = 0 \ ,$$

and the initial condition $u(x,0) = f(x)$.

We shall solve Burgers equation using a Chebyshev Galerkin method, and use the modified basis

$$\phi_n(x) = T_{n+1}(x) - T_{n-1}(x) \quad ,$$

which has the desired property, $\phi_n(\pm 1) = 0$.

We seek solutions, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=1}^{N-1} \hat{u}_n(t)\phi_n(x) = \sum_{n=1}^{N-1} \hat{u}_n(t)\left(T_{n+1}(x) - T_{n-1}(x)\right) \quad ,$$

and require the residual,

$$R_N(x,t) = \frac{\partial u_N}{\partial t} + u_N \frac{\partial u_N}{\partial x} - \nu \frac{\partial^2 u_N}{\partial x^2} \quad ,$$

to be orthogonal to $\phi_k(x)$ in $L^2_w[-1,1]$ as

$$\forall k \in [1, N-1] : \ \frac{2}{\pi} \int_{-1}^{1} R_N(x,t)\phi_k(x) \frac{1}{\sqrt{1-x^2}}\, dx = 0 \ .$$

This yields the Chebyshev Galerkin scheme

$$\forall k \in [1, N-1] : \ \sum_{n=1}^{N-1} \mathrm{M}_{kn} \frac{d\hat{u}_n}{dt} = \sum_{n=1}^{N-1} \mathrm{S}_{kn}(\hat{\boldsymbol{u}}(t))\hat{u}_n(t) \quad ,$$

where the tridiagonal mass-matrix has the entries

$$\mathrm{M}_{kn} = \frac{2}{\pi} \int_{-1}^{1} \phi_k(x)\phi_n(x) \frac{1}{\sqrt{1-x^2}}\, dx = 2\delta_{kn} - \delta_{k,n+2} - \delta_{k,n-2} \ .$$

The derivation of the stiffness-matrix is more complicated, with the entries of S consisting of contributions from the hyperbolic part, S^h, and the parabolic part, S^p, respectively, as

$$\begin{aligned}
\mathrm{S}_{kn}(\hat{\boldsymbol{u}}(t)) &= -\mathrm{S}^h_{kn}(\hat{\boldsymbol{u}}(t)) + \nu \mathrm{S}^p_{kn} \\
&= -\frac{2}{\pi} \int_{-1}^{1} \phi_k(x)\phi_n(x) \sum_{l=1}^{N-1} \hat{u}_l(t) \frac{d\phi_l(x)}{dx} \frac{1}{\sqrt{1-x^2}}\, dx \\
&\quad + \nu \frac{2}{\pi} \int_{-1}^{1} \phi_k(x) \frac{d^2\phi_n(x)}{dx^2} \frac{1}{\sqrt{1-x^2}}\, dx \ .
\end{aligned}$$

If we first consider the term associated with the parabolic term, we recognize this term as the one obtained for the parabolic equation in the Example 30. The entries of this part depends only on the choice of basis functions and may be computed once and for all using the definition of $\phi_n(x)$ and the orthogonality of the Chebyshev polynomials or, alternatively, use a Gaussian quadrature.

The situation with the nonlinear part is worse. Here we have the entries on the form

$$\mathrm{S}^h_{kn}(\hat{\boldsymbol{u}}(t)) = \frac{2}{\pi}\int_{-1}^{1} \phi_k(x)\phi_n(x) \sum_{l=1}^{N-1} \hat{u}_l(t) \frac{d\phi_l(x)}{dx} \frac{1}{\sqrt{1-x^2}}\, dx \quad .$$

By invoking the definition of $\phi_n(x)$ and the identity

$$2T_n(x)T_l(x) = T_{|n+l|}(x) + T_{|n-l|}(x) \quad ,$$

this matrix may be simplified. However, it depends on $\hat{u}_l(t)$ due to the nonlinearity and will need to be computed whenever $\hat{u}_l(t)$ changes. The computation of the integrals involved in the stiffness-matrix is expensive. One could use a Gaussian quadrature to simplify matters but that would introduce aliasing error due to the nonlinear term and would be costly. Hence, apart from the considerable complexity involved in deriving the Chebyshev Galerkin scheme for Burgers equations, it also appears to be computationally intensive.

Nevertheless, once the two components of the stiffness-matrix are obtained, we have the $N-1$ equations for the $N-1$ unknowns, $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_{N-1})^T$,

$$\frac{d\hat{\boldsymbol{u}}}{dt} = \mathrm{M}^{-1}\left(-\mathrm{S}^h(\hat{\boldsymbol{u}}) + \nu \mathrm{S}^p\right)\hat{\boldsymbol{u}}(t) \quad ,$$

and the initial conditions

$$\forall n \in [1, N-1]: \ \hat{u}_n(0) = \frac{2}{\pi}\int_{-1}^{1} f(x)\left(T_{n+1}(x) - T_{n-1}(x)\right)\frac{1}{\sqrt{1-x^2}}\, dx \quad .$$

As illustrated in the above, the application of polynomial Galerkin methods for the solution of partial differential equations takes one through considerable complexity, analytically as well as computationally. Although the mass-matrix depends only on the basis being used, the computation of the entries of the stiffness-matrix has to be completed for

each individual problem and is by no means simple, even for linear problems such as the constant coefficient hyperbolic and parabolic problems. Moreover, as we saw for Burgers equation, the schemes for dealing with nonlinear terms become very complex and computationally intensive. Finally, we have only discussed Galerkin methods for problems with homogeneous boundary conditions which constitutes a special class of problems. More complicated boundary conditions results in a more complicated Galerkin scheme. For problems with time dependent boundary conditions the development of a polynomial Galerkin scheme lead to significant problems and it may in many cases not be possible to derive such schemes. In light of this, it is hardly surprising that polynomial Galerkin methods are used only in those cases where the stiffness- and mass-matrix may be obtained on a closed and simple form, e.g., for linear problems with constant or simple variable coefficient, and the boundary conditions are sufficiently simple. For such problems, however, the Galerkin scheme is fast as well as accurate.

7.1.2 Tau Methods

The problems encountered when formulating polynomial Galerkin methods can be attributed to the requirement that the basis functions, $\phi_n(x)$, obey the boundary conditions individually. This requires us to form polynomials with this property at the loss of orthogonality. In the polynomial tau method this problem is overcome by modifying the definition of the projection operator.

We seek solutions, $u_N(x,t) \in \mathsf{B}_N$, to Eq.(7.1) of the form

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t)\phi_n(x) \quad ,$$

where the polynomial space, B_N, can be characterized as

$$\mathsf{B}_N = \text{span}\left\{\phi_n(x) \in \text{span}\left\{x^k\right\}_{k=0}^{n} | \mathcal{B}_- u_N(-1,t) = 0\,, \quad \mathcal{B}_+ u_N(1,t) = 0\right\}_{n=0}^{N} \quad .$$

If we now define

$$\mathsf{P}_N = \text{span}\left\{\phi_n(x)\right\}_{n=0}^{N} \quad ,$$

then the definition of the projection operator consists of two parts. The first part is the projection of the residual onto P_{N-l} rather than B_N

as for the Galerkin method. Here l signifies the number of boundary conditions. Hence, we require that the first $N-l+1$ components of the residual be orthogonal to P_{N-l} in $L^2_w[-1,1]$ as

$$\forall k \in [0, N-k] : \ \frac{1}{\gamma_k} \int_{-1}^{1} \left(\frac{\partial u_N}{\partial t} - \mathcal{L} u_N \right) P_k^{(\alpha)}(x) w(x)\, dx = 0 \ ,$$

The remaining l equations are recovered as

$$\sum_{n=0}^{N} \hat{u}_n(t) \mathcal{B}_- \phi_n(-1) = 0 \ , \ \sum_{n=0}^{N} \hat{u}_n(t) \mathcal{B}_+ \phi_n(1) = 0 \ ,$$

to ensure that $u_N(\pm 1, t)$ obeys the boundary conditions. Although the formulation of the general scheme is slightly more complicated than the straightforward Galerkin method, the separation of the equation and boundary conditions leads to significantly simpler schemes. Moreover, it allows for dealing with general time dependent boundary conditions as we shall see in the following examples.

Example 32. Consider the linear hyperbolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \ ,$$

where $u(x,t) \in L^2[-1,1]$. The boundary condition is

$$u(1,t) = h(t) \ ,$$

and the initial condition, $u(x,0) = f(x)$.

We consider a Legendre tau method and seek solutions, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t) P_n(x) \ ,$$

with the additional constraint that

$$\sum_{n=0}^{N} \hat{u}_n(t) P_n(1) = \sum_{n=0}^{N} \hat{u}_n(t) = h(t) \ ,$$

to ensure that the solution obeys the boundary condition. To compare with the Galerkin method, let us assume that $h(t) = 0$ and introduce

the boundary operator into the expansion for $u_N(x,t)$ as

$$u_N(x,t) = \sum_{n=0}^{N-1} \hat{u}_n(t) P_n(x) - P_N(x) \sum_{n=0}^{N-1} \hat{u}_n = \sum_{n=0}^{N-1} \hat{u}_n(t) \phi_n(x) \quad ,$$

where $\phi_n(x) = P_n(x) - P_N(x)$, which satisfy $\phi_n(1) = 0$. Thus, the method is somewhat akin to a Galerkin method of order N with the exception that $\phi_N(x) \equiv 0$ and an additional equation is needed to obtain $\hat{u}_N(t)$. Note, that we can not in general give the explicit form of $\phi_n(x)$ since it is a nontrivial combination of the polynomial basis and the boundary conditions.

The first N equations are found by requiring that the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} \quad ,$$

is L^2-orthogonal to P_{N-1} as

$$\forall k \in [0, N-1] : \ \frac{2k+1}{2} \int_{-1}^{1} R_N(x,t) P_k(x) \, dx = 0 \quad .$$

This yields the first N equations

$$\forall n \in [0, N-1] : \frac{\partial \hat{u}_n}{\partial t} = (2n+1) \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{N} \hat{u}_p(t) \quad .$$

Here we recall the identity, Eq.(6.65),

$$\hat{u}_n^{(1)}(t) = (2n+1) \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{N} \hat{u}_p(t) \quad .$$

The equation to obtain $\hat{u}_N(t)$ appears directly as a constraint

$$\hat{u}_N(t) = - \sum_{n=0}^{N-1} \hat{u}_n(t) - h(t) \quad .$$

This reflects the ease by which time dependent boundary conditions are introduced into the tau method.

The initial conditions are recovered directly as

$$\forall n \in [0, N-1] : \hat{u}_n(0) = \frac{2n+1}{2} \int_{-1}^{1} f(x) P_n(x)\, dx \quad ,$$

with the final coefficient, $\hat{u}_N(0)$, being

$$\hat{u}_N(0) = -\sum_{n=0}^{N-1} \hat{u}_n(0) - h(0) \quad .$$

Since we project onto P_{N-1} rather than B_N the mass matrix remains diagonal due to orthogonality and the stiffness matrix is obtained directly from the properties of the polynomial basis, typically resulting in schemes being much simpler than the Galerkin approximation.

Example 33. Let us again consider the linear parabolic problem as

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad ,$$

where $u(x,t) \in L^2_w[-1,1]$. We ask that the solution obeys the general boundary conditions

$$\begin{aligned} \mathcal{B}_- u(-1,t) &= \alpha_1 u(-1,t) - \beta_1 \frac{\partial u(-1,t)}{\partial x} = g(t) \\ \mathcal{B}_+ u(1,t) &= \alpha_2 u(1,t) + \beta_2 \frac{\partial u(1,t)}{\partial x} = h(t) \quad , \end{aligned}$$

which leads to a wellposed problem for the constants, $\alpha_1, \beta_1, \alpha_2, \beta_2$, all being positive and related as $2\alpha_1\alpha_2 \geq \alpha_1\beta_2 + \alpha_2\beta_1$ [**?**]. As initial condition we have $u(x,0) = f(x)$.

We consider a Chebyshev tau method and seek solutions, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t) T_n(x) \quad ,$$

with the additional constraints from the boundary conditions that

$$\sum_{n=0}^{N} \hat{u}_n(t) \left(\alpha_1 T_n(-1) - \beta_1 T_n'(-1)\right) = g(t)$$
$$\sum_{n=0}^{N} \hat{u}_n(t) \left(\alpha_2 T_n(1) + \beta_2 T_n'(1)\right) = h(t) \ .$$

Using the boundary values of $T_n(x)$ and its derivative as

$$T_n(\pm 1) = (\pm 1)^n \ , \ T_n'(\pm 1) = (\pm 1)^{n+1} n^2 \ ,$$

the constraints become

$$\sum_{n=0}^{N} \hat{u}_n(t) \left(\alpha_1 (-1)^n - \beta_1 (-1)^{n+1} n^2\right) = g(t) \ , \tag{7.2}$$
$$\sum_{n=0}^{N} \hat{u}_n(t) \left(\alpha_2 + \beta_2 n^2\right) = h(t) \ ,$$

to ensure that the solution obey the boundary conditions.

We now require that the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \frac{\partial^2 u_N}{\partial x^2} \ ,$$

is orthogonal to P_{N-2} in $L_w^2[-1,1]$ as

$$\forall k \in [0, N-2] : \ \frac{2}{\pi c_k} \int_{-1}^{1} R_N(x,t) T_k(x) \frac{1}{\sqrt{1-x^2}} \, dx = 0 \ .$$

This yields the first $N-1$ equations

$$\forall n \in [0, N-2] : \frac{\partial \hat{u}_n}{\partial t} = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ p+n \text{ even}}}^{N} p(p^2 - n^2) \hat{u}_p(t) \ ,$$

using the identity

$$\hat{u}_n^{(2)}(t) = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ p+n \text{ even}}}^{N} p(p^2 - n^2) \hat{u}_p(t) \ ,$$

which relates the expansion coefficients for the function with those of

the second derivative. The final 2 equations required to obtain $\hat{u}_{N-1}(t)$ and $\hat{u}_N(t)$ appear by solving the linear 2-by-2 system of Eq.(7.2).

The initial conditions are obtained directly as

$$\forall n \in [0, N-2] : \hat{u}_n(0) = \frac{2}{c_n \pi} \int_{-1}^{1} f(x) T_n(x) \frac{1}{\sqrt{1-x^2}} \, dx \quad ,$$

with the final two coefficient being obtained from Eq.(7.2) at $t = 0$.

Even though we considered very general time dependent boundary conditions, the resulting Chebyshev tau method remains simple as the effect of the boundary conditions are separated from the approximation of the operator, $\mathcal{L}$.

While the emphasis in this text is on schemes for the solution of time-dependent problems we would like to take a small detour to illustrate the efficacy of tau methods for the solution of elliptic problems which is where they enjoy particular interest.

Example 34. Consider the elliptic problem

$$\frac{\partial^2 u}{\partial x^2} = f(x) \quad ,$$

where $u(x) \in L^2_w[-1, 1]$ and the solution is subject to the general boundary conditions

$$\alpha_1 u(-1) + \beta_1 \frac{\partial u(-1)}{\partial x} = c_-$$
$$\alpha_2 u(1) + \beta_2 \frac{\partial u(1)}{\partial x} = c_+ \quad ,$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are constants and the boundary values are given through $c_\pm$.

We now seek a solution, $u_N(x) \in \mathsf{B}_N$, on the form

$$u_N(x, t) = \sum_{n=0}^{N} \hat{u}_n(t) T_n(x) \quad ,$$

with the additional constraints from the boundary conditions that

$$\sum_{n=0}^{N} \hat{u}_n \left(\alpha_1 (-1)^n + \beta_1 (-1)^{n+1} n^2\right) = c_- \tag{7.3}$$
$$\sum_{n=0}^{N} \hat{u}_n \left(\alpha_2 + \beta_2 n^2\right) = c_+ \quad ,$$

similar to the approach taken in Example 33.

The residual is given as

$$R_N(x) = \frac{\partial^2 u_N}{\partial x^2} - f_{N-2}(x) \quad ,$$

where $f_{N-2}(x) \in \mathsf{P}_{N-2}$ as

$$f_{N-2}(x) = \sum_{n=0}^{N-2} \hat{f}_n T_n(x) \quad ,$$

where

$$\hat{f}_n = \frac{2}{c_n \pi} \int_{-1}^{1} f(x) T_n(x) \frac{1}{\sqrt{1-x^2}}\, dx \quad .$$

The introduction of $f_{N-2}(x)$ is consistent with $u_N(x)$ being a polynomial of order N, i.e. the approximation to $f(x)$ must be a polynomial of order $N-2$.

Requiring that the residual to be L^2_w-orthogonal to P_{N-2} yields the first $N-1$ equations

$$\forall n \in [0, N-2] : \hat{u}_n^{(2)} = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ p+n \text{ even}}}^{N} p(p^2 - n^2) \hat{u}_p = \hat{f}_n \quad .$$

At this point the solution to the problem is undetermined as we have $N+1$ unknowns, $\hat{u}_n$, but only $N-1$ equations. However, introducing the two boundary conditions, Eq.(7.3), as the remaining two rows into the matrix completes the specification of the scheme. Apart from the two equations appearing from the boundary conditions, the matrix is fairly sparse and upper triangular.

We observe that the complex boundary conditions do not pose a problem. The use of tau methods for solving linear elliptic problems or eigen-

value problems may well be the most efficient way to solve such problems using spectral methods as the resulting matrices typically are sparse and very efficient preconditioners can be developed citeCoutsias95.

Let us finally consider the formulation of a tau method for the solution of Burgers equation.

Example 35. Consider Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \quad ,$$

where $u(x,t) \in L^2_w[-1,1]$ and the boundary condition are

$$u(-1,t) = u(1,t) = 0 \quad ,$$

with the initial conditions, $u(x,0) = f(x)$.

We wish to solve Burgers equation using a Chebyshev tau method, and seek solutions, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=0}^{N} \hat{u}_n(t) T_n(x) \quad ,$$

with the constraints

$$\sum_{n=0}^{N} \hat{u}_n(t)(-1)^n = \sum_{n=0}^{N} \hat{u}_n(t) = 0 \quad .$$

Considering the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} + u_N\frac{\partial u_N}{\partial x} - \nu\frac{\partial^2 u_N}{\partial x^2} \quad ,$$

we require that the first $N-1$ components are L^2_w-orthogonal to P_{N-2} as

$$\forall k \in [0, N-2] : \ \frac{2}{c_k \pi}\int_{-1}^{1} R_N(x,t) T_k(x) \frac{1}{\sqrt{1-x^2}}\, dx = 0 \quad .$$

to recover

$$\forall k \in [0, N-2] : \frac{\partial \hat{u}_k}{\partial t} + \frac{2}{c_k \pi}\int_{-1}^{1} u_N \frac{\partial u_N}{\partial x} T_k(x) \frac{1}{\sqrt{1-x^2}}\, dx = \nu \hat{u}_k^{(2)}(t) \quad .$$

Recall that

$$\hat{u}_k^{(2)}(t) = \frac{1}{c_k} \sum_{\substack{p=n+2 \\ p+k \text{ even}}}^{N} p(p^2 - k^2)\hat{u}_p(t) \quad .$$

The specification of the nonlinear term requires a little more work. Introducing the expansions for $u_N(x,t)$ and its derivative we obtain

$$\frac{2}{c_k \pi} \int_{-1}^{1} u_N \frac{\partial u_N}{\partial x} T_k(x) \frac{1}{\sqrt{1-x^2}}\, dx =$$
$$\frac{2}{c_k \pi} \int_{-1}^{1} \sum_{n,l=0}^{N} \hat{u}_n(t)\hat{u}_l^{(1)}(t) T_n(x) T_l(x) T_k(x) \frac{1}{\sqrt{1-x^2}}\, dx \quad .$$

The identity for Chebyshev polynomials

$$T_n(x)T_l(x) = \frac{1}{2}\left(T_{n+l}(x) + T_{|n-l|}(x)\right) \quad ,$$

yields

$$\frac{2}{c_k \pi} \int_{-1}^{1} u_N \frac{\partial u_N}{\partial x} T_k(x) \frac{1}{\sqrt{1-x^2}}\, dx =$$
$$\frac{1}{2}\frac{2}{c_k \pi} \int_{-1}^{1} \sum_{n,l=0}^{N} \hat{u}_n(t)\hat{u}_l^{(1)}(t) \left(T_{n+l}(x) + T_{|n-l|}(x)\right) T_k(x) \frac{1}{\sqrt{1-x^2}}\, dx =$$
$$\frac{1}{2}\left(\sum_{\substack{n,l=0 \\ n+l=k}}^{N} \hat{u}_n(t)\hat{u}_l^{(1)}(t) + \sum_{\substack{k,l=0 \\ |n-l|=k}}^{N} \hat{u}_n(t)\hat{u}_l^{(1)}(t) \right) \quad ,$$

where

$$\hat{u}_l^{(1)}(t) = \frac{2}{c_l} \sum_{\substack{p=l+1 \\ p+l \text{ odd}}}^{N} p\hat{u}_p(t) \quad .$$

This establishes the equations for the first $N-1$ expansion coefficients, $\hat{u}_k(t)$. The remaining two are found by enforcing the boundary conditions as

$$\sum_{n=0}^{N} \hat{u}_n(t)(-1)^n = \sum_{n=0}^{N} \hat{u}_n(t) = 0 \ ,$$

and the initial conditions are given as

$$\forall n \in [0, N-2] : \hat{u}_n(0) = \frac{2}{c_n \pi} \int_{-1}^{1} f(x) T_n(x) \frac{1}{\sqrt{1-x^2}} \, dx \ .$$

Although the tau method avoids some of the problems of the Galerkin methods and allows for the development of fairly simple and efficient methods for solving the partial differential equation it is primarily being used for the solution of linear elliptic equations. Th e main reason is that the tau method still suffers from the need to derive equations for the expansion coefficients for each individual problem. For variable coefficient or nonlinear problems this process may be very complicated and in many cases impossible. However for linear constant coefficient or special cases of variable coefficient/non-linear problems the resulting tau method is as good as any other polynomial based approach for solving partial differential equations.

7.1.3 Collocation Methods

The problems involved in deriving Galerkin and tau schemes for partial differential equations can be attributed to the need to obtain equations that describe the temporal development of the expansion coefficients as a function of the expansion coefficients themselves. This involves the projection of the residual onto a polynomial space, resulting in a need to evaluate one or several integrals which may well be very complicated. This is in particular true when one considers variable coefficient or non-linear problems.

The collocation method, on the other hand, deals with linear, variable coefficient and nonlinear problems with equal ease, however, at the expense of introducing a grid and, thus, the associated aliasing error.

The appropriate choice of the grid on which to satisfy the equation depends on the actual problem being considered. However, as we are considering initial boundary value problems it seems natural to require that the grid includes the boundary points to allow the enforcement of the boundary conditions. This requirement suggests that the Gauss-

Lobatto quadrature points is the most natural set of grid points to be used with polynomial collocation methods and this set of grid points is indeed the choice in the vast majority of spectral schemes developed to date.

We wish to reiterate that there are generally two sets of grid points to choose in connection with formulating a collocation scheme. One set is associated with the polynomial space, B_N, and are the grid points on which the polynomial solution, u_N, is based. For accuracy these are often taken to be some set of Gauss quadrature points. The other set, on which we require the equation to be satisfied, is generally independent of the former grid and can be chosen with other objectives that accuracy in mind, e.g., stability. However, they are often taken to be the same and we shall also restrict the attention to this case in the following. We stress that this in one among many choices only.

Let us assume that we choose to construct our collocation method based on the Gauss-Lobatto quadrature points for the polynomial family, $P_n^{(\alpha)}(x)$, given as

$$\forall j \in [0,N] : x_j = \left\{ x | (1-x^2)(P_N^{(\alpha)})'(x) = 0 \right\} \quad .$$

The nodes, x_j, are assumed ordered such that $-1 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$. Note that although the present discussion is centered around the use of Gauss-Lobatto points, the development of collocation methods based on other grid points, e.g., the Gauss or Gauss-Radau quadrature points, follows an equivalent route.

We shall seek solutions, $u_N(x,t) \in \mathsf{B}_N$, to Eq.(7.1) of the form

$$u_N(x,t) = \sum_{n=0}^{N} \tilde{u}_n(t)\phi_n(x) = \sum_{j=0}^{N} u_N(x_j,t) l_j(x) \quad ,$$

where the space, B_N, in which we seek solutions is given as

$$\mathsf{B}_N = \left\{ \text{span}\left\{ l_j(x) \right\}_{j=0}^{N} | \mathcal{B}_- u_N(-1,t) = 0 \ , \quad \mathcal{B}_+ u_N(1,t) = 0 \right\} \quad .$$

We assume that the discrete expansion coefficients, $\tilde{u}_n(t)$, are found using the Gauss-Lobatto quadrature rule

$$\tilde{u}_n(t) = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u_N(x_j,t) P_n^{(\alpha)}(x_j) w_j \quad ,$$

where w_j refers to the Gauss-Lobatto weights, Chapter 6.3.2. We recall that using the Gauss-Lobatto quadrature rule we may also express the polynomial, $u_N(x,t)$, using the interpolating Lagrange polynomial, $l_j(x)$, based on the Gauss-Lobatto quadrature points

Introducing the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \mathcal{L}u_N(x,t) \ ,$$

we proceed by requiring this to vanish exactly at the interior grid points, x_j, as

$$\forall j \in [1, N-1] : \ \mathcal{I}_N R_N(x_j,t) = 0 \ ,$$

leading to the $N-1$ equations

$$\forall j \in [1, N-1] : \ \left.\frac{\partial u_N}{\partial t}\right|_{x_j} = (\mathcal{L}u_N)|_{x_j} \ ,$$

with the additional requirements that

$$\mathcal{B}_- u(x_0,t) = 0 \ , \ \ \mathcal{B}_+ u(x_N,t) = 0 \ .$$

This results in $N+1$ equations for the $N+1$ unknowns which are the grid point values. The initial conditions are obtained as $u_N(x_j,0) = \mathcal{I}_N f(x_j)$.

It is instructive to reiterate the differences between the Galerkin/tau methods and the collocation method. In the former two cases we require the residual to be orthogonal to some specific polynomial space and the unknowns are the expansion coefficients. In the latter case we request that the residual vanishes at all interior collocation points and the unknowns are in this case the value of the solution at the grid points. As interpolation is much easier than projection, the collocation schemes are derived in a more straightforward manner.

Let us now consider a few examples similar to those discussed in detail for the Galerkin and the tau methods.

Example 36. Consider first the linear hyperbolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \ ,$$

where we assume that $u(x,t) \in L^2_w[-1,1]$ and the solution is subject to

the boundary condition

$$u(1,t) = h(t) \ ,$$

with the initial condition $u(x,0) = f(x)$.

We shall solve the problem using a Chebyshev collocation method and choose to use the Gauss-Lobatto quadrature points

$$\forall j \in [0,N] : \ x_j = -\cos\left(\frac{\pi}{N} j\right) \ ,$$

to define the polynomial space as well as those points on which to satisfy the equation.

We seek a solution, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=0}^{N} \tilde{u}_n(t) T_n(x) = \sum_{j=0}^{N} u_N(x_j,t) l_j(x) \ ,$$

where the discrete expansion coefficients are given as

$$\tilde{u}_n(t) = \frac{2}{\overline{c}_n N} \sum_{j=0}^{N} \frac{1}{\overline{c}_j} u_N(x_j) T_n(x_j) \ ,$$

where $\overline{c}_0 = \overline{c}_N = 2$ and $\overline{c}_n = 1$ otherwise. Alternatively, the polynomial solution can be expressed through the interpolating Lagrange polynomial as

$$l_j(x) = \frac{(-1)^{j+1}(1-x^2)T'_N(x)}{\overline{c}_j N^2 (x - x_j)} \ .$$

We now require that the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} \ ,$$

vanish at all interior grid points as

$$\forall j \in [0,N-1] : \ \mathcal{I}_N R_N(x_j,t) = \left.\frac{\partial u_N}{\partial t}\right|_{x_j} - \mathcal{I}_N \left.\frac{\partial u_N}{\partial x}\right|_{x_j} = 0 \ ,$$

yielding N equations

$$\forall j \in [0,N-1] : \ \left.\frac{du_N}{dt}\right|_{x_j} = \mathcal{I}_N \left.\frac{\partial u_N}{\partial x}\right|_{x_j} \ .$$

The computation of the derivative at the quadrature points can be accomplished in two different ways. One can use

$$\mathcal{I}_N \frac{\partial u_N}{\partial x}\bigg|_{x_j} = \sum_{j=0}^{N} \tilde{u}_n^{(1)}(t) T_n(x_j) \quad ,$$

where the discrete expansion coefficients are found through a backward recursion relation

$$c_{n-1}\tilde{u}_{n-1}^{(1)}(t) = \tilde{u}_{n+1}^{(1)}(t) + 2n\tilde{u}_{n-1}(t) \quad .$$

Alternatively, we may compute the derivative through the introduction of the differentiation matrix, D, as

$$\mathcal{I}_N \frac{\partial u_N}{\partial x}\bigg|_{x_j} = \sum_{i=0}^{N} \mathrm{D}_{ji} u_N(x_i, t) \quad .$$

The final equation is simply given through the boundary condition as

$$u_N(x_N, t) = h(t) \quad ,$$

with the initial condition becoming

$$u_N(x_j, 0) = \mathcal{I}_N f(x_j) = f(x_j) \quad .$$

The hyperbolic problem is one of the cases where the Gauss-Radau quadrature points are useful as the boundary condition need to be enforced at one boundary only. The development of such a scheme follows the approach for the Gauss-Lobatto method exactly with the only difference being in the computation of the discrete expansion coefficients and the entries of the differentiation matrix.

Let us consider another example of a polynomial collocation method for the solution of a partial differential equation.

Example 37. Consider the linear parabolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad ,$$

where $u(x,t) \in L^2[-1,1]$. We ask that the solution obeys the boundary

conditions

$$\mathcal{B}_- u(-1,t) = \alpha_1 u(-1,t) - \beta_1 \frac{\partial u(-1,t)}{\partial x} = g(t)$$
$$\mathcal{B}_+ u(1,t) = \alpha_2 u(1,t) + \beta_2 \frac{\partial u(1,t)}{\partial x} = h(t) \ ,$$

where the necessary bounds for well posedness on $\alpha_1, \beta_1, \alpha_2, \beta_2$ are discussed in Ex. 33. As initial condition we have $u(x,0) = f(x)$.

We consider a Legendre Gauss-Lobatto collocation method suitable for solving this problem and seek solutions, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{j=0}^{N} u_N(x_j,t) l_j(x) \ ,$$

where $l_j(x)$ signifies the interpolating Lagrange polynomial based on the Legendre Gauss-Lobatto quadrature points given as

$$\forall j \in [0,N] : \ x_j = \{x | (1-x^2) P_N'(x) = 0\} \ .$$

We choose the points to constrain the solution on and require that the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \frac{\partial^2 u_N}{\partial x^2} \ ,$$

vanishes at interior points, yielding the $N-1$ equations

$$\forall j \in [1,N-1] : \ \frac{du_N(x_j)}{dt} = \sum_{i=0}^{N} \mathrm{D}_{ji}^{(2)} u_N(x_i,t) \ .$$

Here $\mathrm{D}^{(2)}$ represents the second order differentiation matrix based on the Legendre Gauss-Lobatto quadrature points.

The final two equations needed to obtain the full solution has to be obtained from the boundary conditions. If we introduce the first order differentiation matrix, D, and requiring the boundary condition to be obeyed exactly we have

$$\alpha_1 u_N(x_0,t) - \beta_1 \sum_{j=0}^{N} \mathrm{D}_{0j} u_N(x_j,t) = g(t) \ ,$$

$$\alpha_2 u_N(x_N,t) + \beta_2 \sum_{j=0}^{N} \mathrm{D}_{Nj} u_N(x_j,t) = h(t) \ .$$

This yields a 2-by-2 system for the computation of the boundary values as

$$(\alpha_1 - \beta_1 \mathrm{D}_{00}) u_N(x_0,t) - \beta_1 \mathrm{D}_{0N} u_N(x_N,t) = g(t) + \beta_1 \sum_{j=1}^{N-1} \mathrm{D}_{0j} u_N(x_j,t)$$

$$\beta_2 \mathrm{D}_{N0} u_N(x_0,t) + (\alpha_2 + \beta_2 \mathrm{D}_{NN}) u_N(x_N,t) = h(t) - \beta_2 \sum_{j=1}^{N-1} \mathrm{D}_{Nj} u_N(x_j,t) \ .$$

In this way the boundary conditions are enforced exactly. Note that in the case of Dirichlet boundary conditions, i.e. $\beta_1 = \beta_2 = 0$, the scheme becomes equivalent to the approach discussed for the hyperbolic problem in the previous example.

Example 38. Consider Burgers equation

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \ ,$$

where $u(x,t) \in L^2_w[-1,1]$ and the boundary condition are

$$u(-1,t) = u(1,t) = 0 \ .$$

The initial conditions are $u(x,0) = f(x)$.

We solve the problem using a Chebyshev collocation method based on the Gauss-Lobatto quadrature points

$$\forall j \in [0,N] : \ x_j = -\cos\left(\frac{\pi}{N} j\right) \ ,$$

and seek a solution, $u_N(x,t) \in \mathsf{B}_N$, of the form

$$u_N(x,t) = \sum_{n=0}^{N} \tilde{u}_n(t) T_n(x) = \sum_{j=0}^{N} u_N(x_j,t) l_j(x) \ ,$$

where the discrete expansion coefficients, $\tilde{u}_n$, and the interpolating Lagrange polynomials, $l_j(x)$, are discussed in Ex. 36.

We require that the residual

$$R_N(x,t) = \frac{\partial u_N}{\partial t} + u_N \frac{\partial u_N}{\partial x} - \nu \frac{\partial^2 u_N}{\partial x^2} \ ,$$

vanish at all interior points of the Gauss-Lobatto grid as

$$\forall j \in [1, N-1] : \ \mathcal{I}_N R_N(x_j,t) = \left.\frac{\partial u_N}{\partial t}\right|_{x_j} + \mathcal{I}_N u_N \left.\frac{\partial u_N}{\partial x}\right|_{x_j} - \nu \left.\frac{\partial^2 u_N}{\partial x^2}\right|_{x_j} = 0 \ .$$

This yields the $N-1$ equations

$$\forall j \in [1, N-1] : \ \left.\frac{du_N}{dt}\right|_{x_j} = -u_N(x_j,t) \left.\frac{\partial u_N}{\partial x}\right|_{x_j} + \nu \left.\frac{\partial^2 u_N}{\partial x^2}\right|_{x_j} \ ,$$

where the spatial derivatives are obtained in one of two ways as

$$\left.\frac{\partial u_N}{\partial x}\right|_{x_j} = \sum_{n=0}^{N} \tilde{u}_n^{(1)}(t) T_n(x_j) = \sum_{i=0}^{N} \mathrm{D}_{ji} u_N(x_i,t) \ ,$$

and

$$\left.\frac{\partial^2 u_N}{\partial x^2}\right|_{x_j} = \sum_{n=0}^{N} \tilde{u}_n^{(2)}(t) T_n(x_j) = \sum_{i=0}^{N} \mathrm{D}_{ji}^{(2)} u_N(x_i,t) \ .$$

Here D and $\mathrm{D}^{(2)}$ signifies the differentiation matrices of first and second order, respectively, and the expansion coefficients for the first, $\tilde{u}_n^{(1)}(t)$, and second order, $\tilde{u}_n^{(2)}(t)$, derivative are obtained by the repeated application of the backward recursion

$$c_{n-1}\tilde{u}_{n-1}^{(1)}(t) = \tilde{u}_{n+1}^{(1)}(t) + 2n\tilde{u}_{n-1}(t) \ .$$

The final two equations are obtained from the boundary conditions as

$$u_N(x_0,t) = u_N(x_N,t) = 0 \ .$$

This last example illustrated the ease by which collocation methods are developed for solving nonlinear problems, contrary to the case for the Galerkin and tau methods which both required considerable work just to be derived. The price for choosing the collocation method is the requirement of a grid and the introduction of the associated aliasing error. However, for many problems there is no real alternative to using the collocation method as the equations for the expansion coefficients

required for the Galerkin and tau methods are impossible to obtain.

7.2 Polynomial Methods on the Unbounded Interval

7.2.1 Laguerre Based Methods on $[0, \infty]$.

7.2.2 Hermite Based Methods on $[-\infty, \infty]$.

7.3 Connecting the Methods

When seeking approximate solutions to partial differential equations an interesting question to raise is what equation is actually being solved, i.e., consider the equation of which the obtained solution is an exact solution in a pointwise sense. Such results sheds some light on the connection between the Galerkin, the Tau and the Collocation methods.

The differential equation we aim to solve is given on the form

$$\frac{\partial u}{\partial t} = \mathcal{L}u(x,t) \quad ,$$

while the numerical solution is obtained by solving the equation

$$\frac{\partial u_N}{\partial t} = \mathcal{L}_N u_N(x,t) \quad ,$$

where $\mathcal{L}_N = \mathcal{P}_N \mathcal{L} \mathcal{P}_N$ is the approximation of the operator, $\mathcal{L}$. The residual, or the error equation, is defined as

$$R_N(x,t) = \frac{\partial u_N}{\partial t} - \mathcal{L}u_N(x,t) = (\mathcal{L}_N - \mathcal{L})\, u_N(x,t) \quad .$$

In the literature of finite differences this equation is frequently referred to as the modified equation. Our interest is here to derive explicit forms of $R_N(x,t)$.

In the following we shall, as an illustration, derive this residual on explicit form for Chebyshev approximation of linear hyperbolic and parabolic problems. The process for other schemes and choice of basis is equivalent.

7.3.1 The Hyperbolic Problem.

Consider the linear hyperbolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \quad ,$$

subject to the boundary condition, $u(1,t) = 0$, and with the initial condition, $u(x,0) = f(x)$.

We begin by considering solution of the problem using a Chebyshev Galerkin method as discussed in Ex. 29, i.e., we require the residual to vanish as

$$\forall k \in [0,N] : \int_{-1}^{1} R_N(x,t)(T_k(x)-1)\frac{1}{\sqrt{1-x^2}}\,dx = 0 \quad .$$

It is easily verified that

$$\frac{1}{2} + \sum_{n=1}^{N} T_n(x) \quad ,$$

satisfies this equation. Since $R_N(x,t) \in \mathsf{B}_N$ is must be proportional to this sum. To determine the constant we observe that as $u_N(1,t)$ is constant we obtain

$$R_N(1,t) = -\left.\frac{\partial u_N}{\partial x}\right|_{x=1} \quad ,$$

and the error equation for the *Chebyshev Galerkin method* becomes

$$\frac{\partial u_N}{\partial t} = \frac{\partial u_N}{\partial x} - \frac{2}{2N+1}\left.\frac{\partial u_N}{\partial x}\right|_{x=1}\left[\frac{1}{2} + \sum_{n=1}^{N} T_n(x)\right] \quad . \tag{7.4}$$

This is the equation being solved exactly. This emphasizes that $\mathcal{P}_N u \neq u_N$, i.e., the computed solution is generally not the projection of the exact solution, even for a problem as simple as the wave equation.

Let us also consider the Chebyshev tau method, for which the error equation becomes

$$\frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} = \tau(t) T_N(x) \quad ,$$

since $u_N(x,t) \in \mathsf{B}_N$. Again using that

$$R_N(1,t) = -\left.\frac{\partial u_N}{\partial x}\right|_{x=1} \quad ,$$

we recover the error equation for the *Chebyshev tau method* as

$$\frac{\partial u_N}{\partial t} = \frac{\partial u_N}{\partial x} - \left.\frac{\partial u_N}{\partial x}\right|_{x=1} T_N(x) \quad .$$

Recall that the Chebyshev Gauss quadrature points of order $N-1$ are the roots of $T_N(x)$, i.e., the Chebyshev tau method is equivalent to some Chebyshev collocation method of order $N-1$ based on the Gauss points.

Let us finally consider the error equation for the Chebyshev collocation method. If we first consider the Chebyshev Gauss-Lobatto method we obtain the error equation as

$$\frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} = \tau(t)(1+x)T_N'(x) \quad ,$$

since the remainder must vanish at all interior nodal points. Using

$$R_N(1,t) = -\left.\frac{\partial u_N}{\partial x}\right|_{x=1} \quad ,$$

we immediately recover the error equation for the *Chebyshev Gauss-Lobatto method* as

$$\frac{\partial u_N}{\partial t} = \frac{\partial u_N}{\partial x} - \frac{1}{2N^2}\left.\frac{\partial u_N}{\partial x}\right|_{x=1}(1+x)T_N'(x) \quad .$$

Let us instead consider the Chebyshev Gauss-Radau method, where the quadrature points are given as

$$\forall j \in [0,N]: \; y_j = -\cos\left(\frac{(2j+1)\pi}{2N+1}\right) \quad ,$$

i.e., $y_N = 1$ is included in the nodal set. If we now consider the polynomial

$$p_N(x) = \frac{1}{2} + \sum_{n=1}^{N} T_n(x) = \frac{\sin\left[\left(N+\frac{1}{2}\right)\cos^{-1}x\right]}{2\sin\frac{1}{2}\cos^{-1}x} \quad ,$$

then it may be verified that $p_N(y_j) = 0$ at all the interior points, while $p_N(1) = N + \frac{1}{2}$. Hence, it follows directly that the error equation for the *Chebyshev Gauss-Radau method* is given as

$$\frac{\partial u_N}{\partial t} = \frac{\partial u_N}{\partial x} - \frac{2}{2N+1}\left.\frac{\partial u_N}{\partial x}\right|_{x=1}\left[\frac{1}{2} + \sum_{n=1}^{N} T_n(x)\right] \quad .$$

It is thus equivalent to the error equation for the Chebyshev Galerkin method, Eq.(7.4). In other words, the Chebyshev Galerkin method and the Chebyshev Collocation method based on the Gauss-Radau quadrature points are identical as they are solving the same equation exactly.

7.3.2 The Parabolic Problem.

Let us also consider the error equation when solving parabolic problems of the type

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \ ,$$

subject to the boundary condition

$$u(-1,t) = u(1,t) = 0 \ ,$$

and the initial conditions, $u(x,0) = f(x)$.

We begin by considering the error equation for the Chebyshev Galerkin method with a basis as $\phi_n(x) = T_{n+1}(x) - T_{n-1}(x)$. Proceeding exactly as for the hyperbolic problem we realize that the remainder can be written as (N assumed even, but similar for odd)

$$R_N(x,t) = A(t) \sum_{\substack{n=0 \\ n \text{ even}}}^{N} \frac{1}{c_n} T_n(x) + B(t) \sum_{\substack{n=1 \\ n \text{ odd}}}^{N-1} T_n(x) \ ,$$

which is orthogonal to $\phi_n(x)$ for $n \in [1, N-1]$. The two constants are recovered by enforcing the boundary conditions as

$$R_N(\pm 1,t) = - \left. \frac{\partial^2 u_N}{\partial x^2} \right|_{\pm 1} \ .$$

This yields the error equation for the *Chebyshev Galerkin method* as

$$\begin{aligned} \frac{\partial u_N}{\partial t} = \frac{\partial^2 u_N}{\partial x^2} &- \frac{1}{N+1} \left(\left. \frac{\partial^2 u_N}{\partial x^2} \right|_1 + \left. \frac{\partial^2 u_N}{\partial x^2} \right|_{-1} \right) \sum_{\substack{n=0 \\ n \text{ even}}}^{N} \frac{1}{c_n} T_n(x) \\ &- \frac{1}{N} \left(\left. \frac{\partial^2 u_N}{\partial x^2} \right|_1 - \left. \frac{\partial^2 u_N}{\partial x^2} \right|_{-1} \right) \sum_{\substack{n=1 \\ n \text{ odd}}}^{N-1} T_n(x) \ . \end{aligned}$$

The error equation for the Chebyshev collocation method based on the Gauss-Lobatto quadrature points is recovered by exploiting that the remainder must vanish at all the interior collocation points. This allows us to recover

$$R_N(x,t) = A(t)(1+x)T'_N(x) + B(t)(1-x)T'_N(x) \ .$$

Applying the boundary conditions directly yields the error equation for the *Chebyshev Gauss-Lobatto method* as

$$\frac{\partial u_N}{\partial t} = \frac{\partial^2 u_N}{\partial x^2} - \frac{1}{2N^2}\left((1+x)T_N'(x)\left.\frac{\partial^2 u_N}{\partial x^2}\right|_1 + (-1)^{N+1}(1-x)T_N'(x)\left.\frac{\partial^2 u_N}{\partial x^2}\right|_{-1}\right) \quad .$$

Unlike the case for the hyperbolic problem, there is no direct relation between the different methods.

7.4 Stability of Polynomial Methods

7.4.1 Polynomial Methods on a Finite Interval

7.4.1.1 Stability of Galerkin and Tau Methods

7.4.1.2 Stability of Collocation Methods

7.4.2 Polynomial Methods on an Unbounded Interval

7.4.2.1 Stability of Galerkin and Tau Methods

7.4.2.2 Stability of Collocation Methods

Exercises

1. Consider

$$\frac{\partial u}{\partial t} + (2+x)\frac{\partial u}{\partial x} = 0 \quad , \quad x \in [-1,1] \quad ,$$

   with

$$u(1,t) = 0 \quad , \quad u(x,0) = f(x) \quad .$$

   Construct a Legendre-Galerkin method for the problem.

2. (Continued) Construct a Legendre-Tau method for this problem.

3. (Continued) Construct a Legendre-Collocation method based on this using Legendre-Gauss-Lobatto quadrature points for both approximating the solution and satisfying the equation.

4. (Continued) Construct a Legendre-Collocation method using the Legendre-Gauss quadrature points to represent the solutions and the Legendre-Gauss-Radau points to satisfy the equation.

5. Consider

$$\frac{\partial u}{\partial t} = \frac{\partial^3 u}{\partial x^3} \quad , \quad x \in [-1,1] \quad .$$

Assume the general boundary conditions take the form

$$\alpha_{\pm} u(\pm 1, t) + \beta_{\pm} \left. \frac{\partial u}{\partial t} \right|_{x=\pm 1} + \gamma_{\pm} \left. \frac{\partial^2 u}{\partial x^2} \right|_{x=\pm 1} = f_{\pm}(t) \quad ,$$

where the constants $\alpha_{\pm}$, $\beta_{\pm}$ and $\gamma_{\pm}$ are chosen to ensure wellposedness.

Formulate a Chebyshev-Tau scheme to solve this problem.

6. (Continued) Formulate Chebyshev-Collocation methods to solve this problem.

7. Consider

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \quad , x \in [-1, 1] \quad ,$$

and

$$u(1, t) = 0 \quad , \quad u(x, 0) = f(x) \quad .$$

Assume that we seek solutions $u_N \in \mathsf{B}_N$ of the form

$$u_N(x, t) = (1 - x) \sum_{j=0}^{N-1} u_N(x_j, t) l_j(x) \quad ,$$

where x_j are the Chebyshev-Gauss points of order $N - 1$.

If we require that the residual vanishes at the Chebyshev-Gauss points, show that the resulting scheme is equivalent to a Chebyshev-Tau method of order N.

8

Spectral Methods for Non-Smooth Problems

So far we have focused almost exclusively on problems with a minimal degree of smoothness, i.e., at least continuous in a classical sense. Indeed, as spectral methods are centered around the use of expansions of the unknown solutions in terms of smooth polynomials, one could think that such high-order methods are less interesting when one attempts to solve problems containing genuinely discontinuous solutions.

While it is true that the straightforward use of spectral methods works best when considering smooth problems there are a number of important reasons for considering the use of spectral methods for truly discontinuous problems. To illustrate this, let us consider an example.

Example 39. Consider the scalar hyperbolic equation

$$\begin{aligned} \frac{\partial u}{\partial t} &= -2\pi \frac{\partial u}{\partial x} \ , \\ u(0,t) &= u(2\pi,t) \ , \end{aligned} \tag{8.1}$$

and assume that we wish to advance the initial conditions using an odd Fourier collocation method with $N = 64$ grid points in space and a 4'th order Runge-Kutta method in time with a time-step well below the stability limits.

Let us also take the initial condition as

$$u(x,0) = \begin{cases} x & 0 \leq x \leq \pi \\ x - 2\pi & \pi < x \leq 2\pi \end{cases} \ ,$$

and assume that it is periodically extended. Clearly, $u(x)$ has a sharp

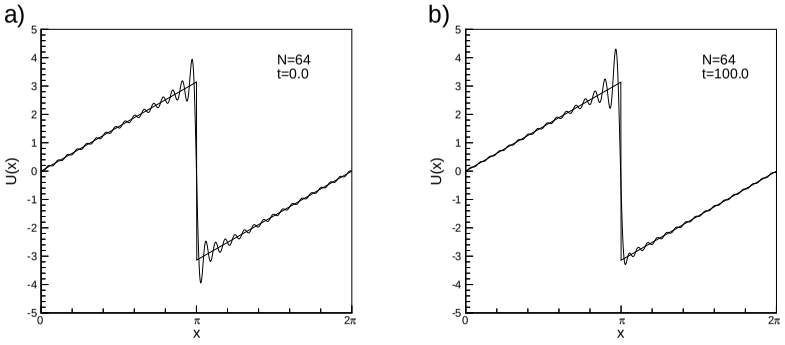

figure 8.1. a) The initial saw-tooth function and the associated discrete Fourier series approximation. b) The saw-tooth function and the convected Fourier approximation at $t = 100$ obtained by solving Eq.(8.1).

discontinuity at $x = \pi$.

In Fig. 8.1 we plot the computed approximation to the convected saw-tooth function at $t = 0$ and after 100 periodic revolutions.

From this simple experiment we may observe a number of central issues related to the spectral solution of discontinuous problems. Already at $t = 0$, i.e., in the initial approximation, do we observe some very strong oscillations in the approximations and it is clear that the approximation is not very good in the neighborhood of the discontinuity although the steep gradient characterizing the discontinuity is fairly well represented.

On the other hand, by inspecting the computed solution after very long time, in this case at $t = 100$, we realize that the discontinuity remains to be well approximated even after very long time and that only little smearing of the initial front can be observed.

While the example clearly illustrates that the initial approximation of non-smooth functions introduces very significant oscillations in the approximation it is remarkable that the superior phase properties of the spectral methods remains intact, hence allowing for the very accurate advection of the initial waveform observed in Fig. 8.1.

Inspired by the above example, we shall devote this chapter to an

analysis of the above example. In particular we shall discuss in detail the appearance of the oscillations, known as the Gibbs phenomenon, and the non-uniform convergence that appears as a result of this. This discussion shall consider not only the Gibbs phenomenon in trigonometric and polynomial expansions but also the implications of the oscillations on the stability of general spectral methods.

The second part of this chapter is devoted to techniques that will allow us to overcome or at least decrease the effect of the Gibbs phenomenon when solving general partial differential equations. A very powerful tool in this respect is the use of filters and we shall discuss these techniques in detail before we turn to the development of a general theory that shall allow us to recover exponentially convergent series for piecewise continuous problems such as to completely overcome the Gibbs phenomenon.

8.1 Trigonometric Approximation of Non-Smooth Problems

Let us begin by considering the behavior of trigonometric expansions of piecewise continuous functions. Prior to obtaining estimates for the convergence of the Fourier series for the approximation of such problems, let us return to Ex. 39 and consider solely the initial approximation problem.

Example 40. Consider again the function

$$u(x) = \begin{cases} x & 0 \leq x \leq \pi \\ x - 2\pi & \pi < x \leq 2\pi \end{cases},$$

and assume that it is periodically extended. The continuous Fourier series expansion coefficients are found as

$$\hat{u}_n = \begin{cases} i(-1)^{|n|}/n & n \neq 0 \\ 0 & n = 0 \end{cases}.$$

In Fig. 8.2 we show the Fourier series approximation to $u(x)$ and observe strong oscillations around the discontinuity while the approximation converges, albeit slowly, away from the discontinuity. Note that the oscillations around the point of discontinuity persists when increasing the resolution. The convergence of the approximation away from the

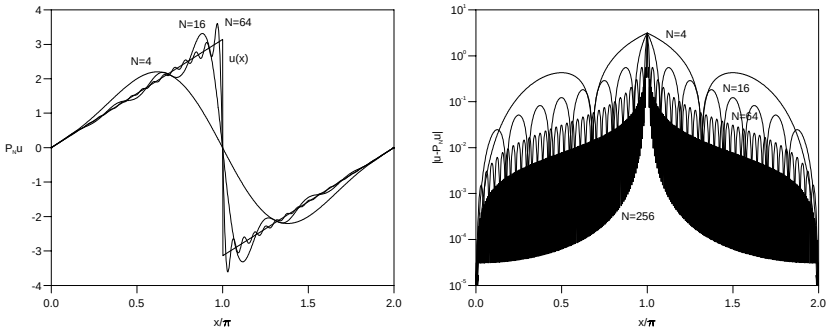

figure 8.2. On the left we show the continuous Fourier series approximation of a discontinuous function illustrating the appearance of the Gibbs phenomenon. The figure on the right displays the pointwise error of the continuous Fourier series approximation for increasing resolution. Note that the value of the pointwise error at the discontinuity corresponds exactly to $(u(\pi+)+u(\pi-))/2$.

discontinuity is also confirmed in Fig. 8.2 where we plot the pointwise error. We find linear convergence at most in correspondence with the decay of the expansion coefficients.

As we clearly observe from Fig. 8.2 the error at the discontinuity does not disappear as we increase the resolution. Thus, the approximation is no longer pointwise convergent although it remains convergent in the mean as a result of completeness as discussed in Chapter 4. This may also be seen by directly estimating the L^2-error using Parseval's identity, Eq.(4.1.5), as

$$\|u - \mathcal{P}_N u\|_{L^2[0,2\pi]} = \left(\sum_{|n|>N} \frac{1}{n^2} \right)^{1/2} \simeq \frac{1}{\sqrt{N}} \ ,$$

i.e., the approximating is convergent but at an extremely slow rate. Although the function is smooth and periodic away from the discontinuity, the global rate of convergence is dominated by the presence of the discontinuity. It is this characteristic oscillatory behavior in the neighborhood of a discontinuity of a truncated Fourier series of a non-smooth function, $u(x)$, that is known as the Gibbs phenomenon.

8.1.1 The Gibbs Phenomenon.

Let us in the following consider the phenomenon in a little more detail. We introduce the truncated symmetric Fourier series sum as

$$\mathcal{P}_N u(x) = \sum_{n=-N/2}^{N/2} \hat{u}_n \exp(inx) \quad ,$$

where the continuous expansion coefficients are recovered as

$$\hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(x) \exp(-inx)\, dx \quad .$$

The behavior of the truncated approximation for a piecewise continuous function is given as follows

Theorem 64. *Every piecewise continuous function, $u(x) \in L^2[0, 2\pi]$, has a Fourier series which is pointwise convergent as*

$$\mathcal{P}_N u(x) \to \frac{u(x^+) + u(x^-)}{2} \quad as \quad N \to \infty \quad .$$

Proof: We begin by writing the truncated series as

$$\begin{aligned}
\mathcal{P}_N u(x) &= \frac{1}{2\pi} \sum_{n=-N/2}^{N/2} \left(\int_0^{2\pi} u(y) \exp(-iny)\, dy \right) \exp(inx) \\
&= \frac{1}{2\pi} \int_0^{2\pi} u(y) \left(\sum_{n=-N/2}^{N/2} \exp(in(x-y)) \right) dy \\
&= \frac{1}{2\pi} \int_{x-2\pi}^{x} D_N(t) u(x-t)\, dt \quad ,
\end{aligned}$$

where we have introduced the Dirichlet kernel

$$D_N(t) = \sum_{n=-N/2}^{N/2} \exp(int) = \frac{\sin((N+1)t/2)}{\sin(t/2)} \quad . \tag{8.2}$$

The Dirichlet kernel can be viewed as the projection of a delta function onto the space spanned by the Fourier basis and is an even function in

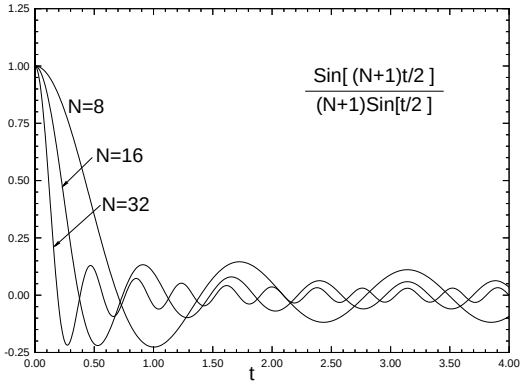

figure 8.3. The normalized Dirichlet kernel for various values of N.

t which oscillates while changing signs at $t_j = 2\pi j/(N+1)$. The kernel is shown for illustration in Fig. 8.3 for increasing values of N.

Due to the strong oscillatory behavior away from the origin, we may also assume that the main contribution of the integral originates from a narrow region around zero, i.e.,

$$\mathcal{P}_N u(x) \simeq \frac{1}{2\pi} \int_{-\varepsilon}^{\varepsilon} D_N(t) u(x-t)\, dt \quad ,$$

where $\varepsilon \ll 1$. Since $u(x)$ is at least piecewise continuous we may also assume that $u(x-t) \simeq u(x^-)$ for $0 \le t \le \varepsilon$ and $u(x-t) \simeq u(x^+)$ for $0 \ge t \ge -\varepsilon$ with at most a jump at $t = 0$. If we in addition assume $\sin(t/2) \simeq t/2$ we recover the result

$$\mathcal{P}_N u(x) \simeq \frac{1}{\pi}(u(x^+) + u(x^-)) \int_0^{\varepsilon} \frac{\sin[(N+1)t/2]}{t}\, dt \quad .$$

However, since

$$\begin{aligned} \frac{1}{\pi} \int_0^{\varepsilon} \frac{\sin[(N+1)t/2]}{t}\, dt &= \frac{1}{\pi} \int_0^{(N+1)\varepsilon/2} \frac{\sin s}{s}\, ds \\ &\simeq \frac{1}{\pi} \int_0^{\infty} \frac{\sin s}{s}\, ds = \frac{1}{2} \quad \text{as} \quad N \to \infty \quad , \end{aligned}$$

this implies the asymptotic result

$$\mathcal{P}_N u(x) \to \frac{u(x^+) + u(x^-)}{2} \quad ,$$

thereby concluding the proof. QED

Thus, pointwise convergence to the average is ensured. However, the important point is that the convergence rate is non-uniform close to a discontinuity. To see that, let us consider the behavior of the approximation in the neighborhood of a discontinuity at x_0 and express the limit of $\mathcal{P}_N u(x)$ as

$$\mathcal{P}_N u\left(x_0 + \frac{2z}{N+1}\right) \simeq \frac{1}{2\pi}\int_{-\varepsilon}^{\varepsilon} D_N(t) u\left(x_0 + \frac{2z}{N+1} - t\right) dt \quad .$$

Again utilizing the local nature of the Dirichlet kernel as in the proof of Theorem 64 we arrive at the asymptotic result for $N \to \infty$

$$\begin{aligned}\mathcal{P}_N u\left(x_0 + \frac{2z}{N+1}\right) &\simeq \frac{u(x_0^+)}{\pi}\int_{-\infty}^{z} \frac{\sin s}{s}\, ds + \frac{u(x_0^-)}{\pi}\int_{z}^{\infty} \frac{\sin s}{s}\, ds \\ &\simeq \frac{1}{2}(u(x_0^+) + u(x_0^-)) + \frac{1}{\pi}(u(x_0^+) - u(x_0^-))\mathrm{Si}(z) \quad .\end{aligned}$$

Here we have introduced the Sine integral function, $\mathrm{Si}(z)$, which we recall is defined

$$\mathrm{Si}(z) = \int_0^z \frac{\sin s}{s}\, ds \quad , \quad \lim_{z\to\infty} \mathrm{Si}(z) = \frac{\pi}{2} \quad , \quad \mathrm{Si}(-z) = -\mathrm{Si}(z) \quad ,$$

and is plotted in Fig. 8.4 for the purpose of illustration. Clearly, for a smooth function we recover the pointwise convergence result. However, for a piecewise smooth function with a jump at x_0 we have $\mathcal{P}_N u(x) - \frac{1}{2}(u(x_0^+) + u(x_0^-)) = \mathcal{O}(1)$ provided $x - x_0 = \mathcal{O}(N^{-1})$, i.e., z is constant such that $\mathrm{Si}(z) \simeq \pi/2$. Consequently, in the neighborhood of the point of discontinuity we must expect non-uniform convergence, while the convergence is linear away from x_0.

The maximum size of the overshoot occurs where the Sine integral has its maximum as happens for $z = \pi$ where

$$\frac{1}{\pi}\mathrm{Si}(\pi) = 0.58949 \quad .$$

Thus, the maximum overshoot/undershoot at a point discontinuity asymptotically approaches

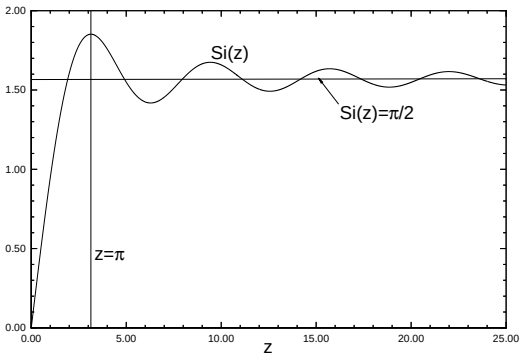

figure 8.4. The Sine integral function, $Si(z)$.

$$u_N(x) - u(x_0^-) \simeq 0.08949(u(x_0^+) - u(x_0^-)) \ .$$

Hence, it is approximately 9% of the magnitude of the jump, and happens approximately for $x = x_0 \pm 2\pi/(N+1)$. Let us conclude this section with a few examples.

Example 41. Consider again the function from Ex. 40

$$u(x) = \begin{cases} x & 0 \le x \le \pi \\ x - 2\pi & \pi < x \le 2\pi \end{cases} \ ,$$

and assume that it is periodically extended. Clearly, $u(x)$ has a sharp discontinuity at $x = \pi$ and using the theory just developed we recover that the maximum overshoot around $x_0 = \pi$ should be

$$|\mathcal{P}_N u(x) - u(\pi^{\pm})| \simeq 0.08949|u(\pi^+) - u(\pi^-)| = 0.08949\, 2\pi \simeq 0.5623 \ ,$$

which corresponds well with the results displayed in Fig. 8.2.

Example 42.

Consider the unit step function, $u(x)$, defined as

$$u(x) = \begin{cases} -\frac{1}{2} & -\pi < x < 0 \\ \frac{1}{2} & 0 < x < \pi \end{cases} ,$$

and assume that it is periodically extended. The continuous Fourier series expansion is found as

$$u(x) = \frac{2}{\pi} \sum_{k=0}^{\infty} \frac{\sin(2k+1)x}{2k+1} .$$

The simpleminded approach would be to look for the extremum of the truncated expansion,

$$\mathcal{P}_N u(x) = \frac{2}{\pi} \sum_{k=0}^{N} \frac{\sin(2k+1)x}{2k+1} ,$$

found through the derivative

$$\mathcal{P}_N u'(x) = \frac{2}{\pi} \sum_{k=0}^{N} \cos(2k+1)x = \frac{\sin 2(N+1)x}{\pi \sin x} ,$$

which is zero for $z = \pi/(2(N+1))$. If we insert this into the truncated expansion one obtains, through the use of the Riemann sum, the asymptotic result

$$\mathcal{P}_N u(z) = \frac{1}{\pi(N+1)} \sum_{k=0}^{N} \frac{\sin \frac{(2k+1)\pi}{2(N+1)}}{\frac{2k+1}{2(N+1)}} \simeq \frac{1}{\pi} \int_0^1 \frac{\sin \pi t}{t} \, dt = \frac{1}{\pi} \mathrm{Si}(\pi) = 0.58949 ,$$

which is exactly what we would expect from the more general analysis, since x approaches $x_0 = 0$ linearly and the size of the jump is one.

So far we have only considered the Gibbs phenomenon and its behavior in the continuous expansions. However, as we have discussed extensively in Chapter 4, the convergence behavior of the discrete expansion is very similar, if not in quantitative terms, then certainly in qualitative terms. Indeed, the Gibbs phenomenon appears in the discrete expansions with characteristics similar to those discussed above. This is most easily understood by recalling that the sole difference between the discrete and the continuous expansions comes in through the aliasing error.

While we are unable to directly quantify this for non-smooth problems it is clear that one should expect the two types of expansions to yield similar results in the asymptotic limit, which is where we arrived at the results discussed above.

8.2 Filters

The slow and nonuniform convergence of $\mathcal{P}_N u(x)$ for a piecewise continuous function can be traced to two facts

- The linear decay of the expansion coefficients.
- The global nature of the polynomial approximation where the expansion coefficients are obtained by integration/summation over the full domain, including the point(s) of discontinuity.

These two factors are clearly not independently contributing to the Gibbs phenomenon but rather two different manifestations of the same phenomenon. However, in trying to resolve the Gibbs phenomenon, two different approaches can be taken each associated with one of the interpretations of the source of the oscillations.

Let us first consider the truncated continuous polynomial approximations as

$$\mathcal{P}_N u(x) = u_N^\sigma(x) = \sum_{n=0}^{N} \sigma(\eta)\hat{u}_n\phi_n(x) \quad , \tag{8.3}$$

where $\hat{u}_n$ are the usual continuous expansion coefficients and the parameter $\sigma(\eta)$ with $\eta = n/N$ represents the filter. Utilizing the definition of the continuous expansion coefficients, we obtain the physical space approximation as

$$u_N^\sigma(x) = \int_{\mathsf{D}} u(y)w(y)S(x,y)\,dy \quad ,$$

where we have introduced the filter function

$$S(x,y) = \sum_{n=0}^{N} \sigma(\eta)\frac{1}{\gamma_n}\phi_n(x)\phi_n(y) \quad . \tag{8.4}$$

If we assume that $\sigma(\eta) = 1$, the filter function can be summed exactly using the Christoffel-Darboux identity for polynomials in which case we realize that $S(x,y)$ is nothing more than the polynomial Dirichlet

kernel. This also emphasizes that the Dirichlet kernel is nothing more than the continuous equivalent of the cardinal functions, or Lagrange polynomials, discussed thoroughly in Chapter 4 for Fourier series and in Chapter 6 for the general case of orthogonal polynomials. Hence, any action taken on the behavior of the Dirichlet kernel can be expected to have a similar impact on the cardinal functions.

In the case of trigonometric expansions this simplifies considerably since

$$S(x) = \sum_{n=-N/2}^{N/2} \sigma(\eta) \exp(inx) \quad , \tag{8.5}$$

which we recall as the very oscillatory Dirichlet kernel shown in Fig. 8.3.

For reasons that will become apparent in Sec. 8.2.2, the filter, $\sigma(\eta)$, is defined as follows

Definition 8. *A real and even function $\sigma(\eta) \in C^{q-1}[-\infty, \infty]$ is called a filter of order q if it has the following properties*

a)

$$\sigma(\eta) = 0 \quad for \quad |\eta| > 1 \quad .$$

b)

$$\sigma(0) = 1 \quad and \quad \sigma(1) = 0 \quad .$$

c)

$$\forall m \in [1, \ldots, q-1] : \quad \sigma^{(m)}(0) = \sigma^{(m)}(1) = 0 \quad .$$

Equations (8.3)-(8.4) suggest, as mentioned briefly in the above, that we can think of at least two different ways of designing and implementing filtering. Equation (8.3) suggests that by enhancing the decay of the expansion coefficients, e.g., by choosing $\sigma(\eta)$ decaying for increasing η, would result in a faster convergent series. This line of argumentation, however, is not without traps as modifying the expansion coefficients has global consequences. Hence, the aim of the analysis is devise filter functions, $\sigma(\eta)$, that has only a localized effect on the approximation which otherwise must remain unchanged in smooth parts of the function.

One way to attempt to overcome this apparent problem with localized modifications of a global expansion is to turn the attention to direct

modifications of the physical space Dirichlet kernel, Eq.(8.4). Since the highly oscillatory behavior of the Dirichlet kernel is a representation of the global nature of the approximation, one could attempt to specify $\sigma(\eta)$ with the aim of localizing the influence of $S(x,y)$. This, however, requires one to specify $\sigma(\eta)$ in a very delicate fashion such as to minimize the oscillations of $S(x,y)$ away from $x=y$ in some prescribed manner. Moreover, localizing the kernel too dramatically essentially reduces the order af the scheme thereby counteracting the effort of devising high-order methods. As it turns out, proper localization of the Dirichlet kernel is at least as complex as that of increasing the decay rate and it is indeed the latter procedure that has received most of the attention in the past.

In the following we shall therefore mainly utilize the first interpretation although deriving the modified associated Dirichlet kernel, as is possible in a few cases, will be seen to yield additional information about the observed behavior of the filter functions.

8.2.1 A First Look at Filters and Their Use.

As we have realized, the Gibbs phenomenon results in very slow decay of the expansion coefficients. Thus, it seems natural to apply filters in an attempt to attenuate the high order coefficients with the aim of increasing the rate of convergence away from the discontinuity. However, care has to be taken when doing so. The high order expansion coefficients carry important information concerning the behavior of the function close to the discontinuity, and this information should not be wasted. Too strong a smoothing procedure results in a strongly smeared function, thus rendering the approximation less useful.

In this section we shall take a first practical look at the use of filters and the associated implications. For the sake of simplicity we shall restrict the examples to continuous Fourier series. While the results certainly are quantitatively different from those obtained when considering filtering of polynomial expansions, the results are nevertheless qualitatively similar and provides a good background for understanding the central implications of using filters in general cases.

Example 43 (Cesáro Filter). This filter represents an arithmetic mean of the truncated series as

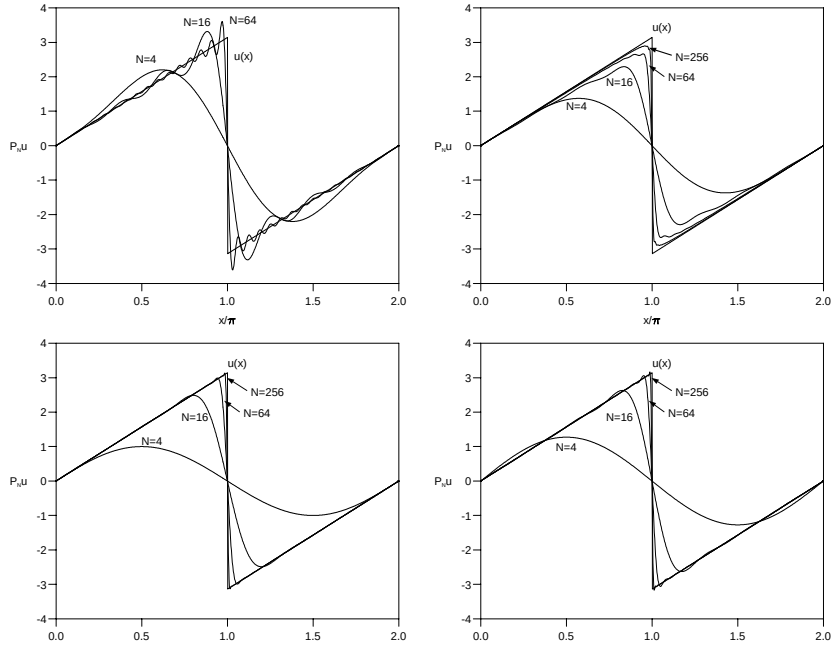

figure 8.5. Effects on the smoothness of the approximation when applying various filters to a Fourier approximation of a discontinuous function. Upper left) Un-filtered approximation. Upper right) Cesáro filtered approximation. Lower left) Raised cosine filtered approximation. Lower right) Lanczos filtered approximation.

$$\sigma(\eta) = 1 - \eta \ ,$$

and is only a linear filter. In Fig. 8.5 we have plotted the Cesáro filtered Fourier series approximation of the discontinuous function introduced in Ex. 40 and compare it with the un-smoothed approximation. We observe that the Cesáro filter inhibit the Gibbs phenomenon close to the discontinuity, however, it also produces a heavily smeared approximation to the original function. In Fig. 8.6 we plot the point-wise error of the filtered approximation, and note that only little is gained in accuracy away from the discontinuity as compared to the un-filtered approximation.

A further understanding of the effect of the Cesáro filter can be obtained by considering the modified Dirichlet kernel, Eq.(8.5), as

$$S(x) = \frac{2}{N+2} \begin{cases} \frac{\sin((N/2+1)x/2)}{\sin(x/2)} & x \neq 2\pi p \\ 1 & x = 2\pi p \end{cases} \quad p = 0, \pm 1, \pm 2, \ldots \quad .$$

We observe, among other things, that $S(x) \geq 0$, i.e., the Cesáro filtered

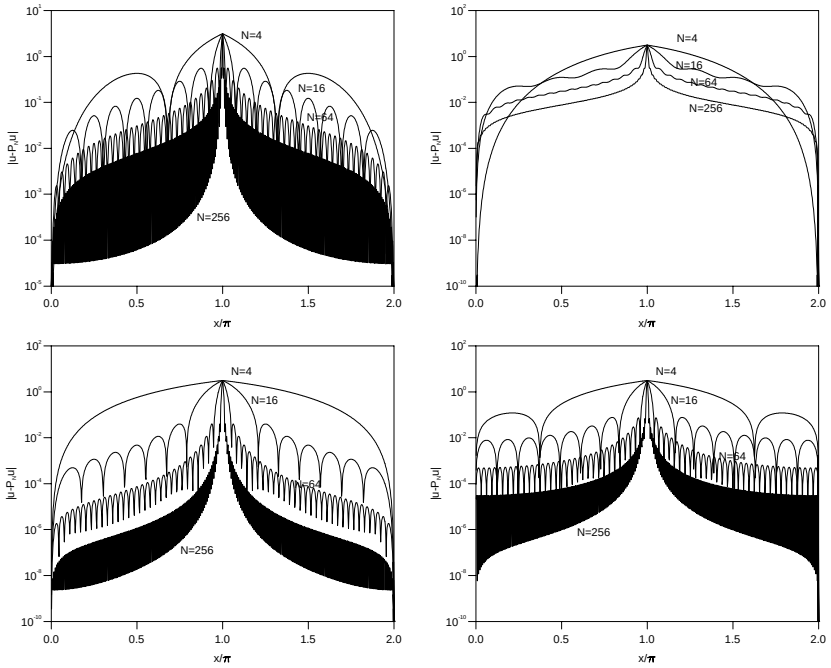

figure 8.6. Pointwise errors of the approximation when applying various filters to a Fourier approximation of a discontinuous function. Upper left) Un-filtered approximation (Note different scale than the rest). Upper right) Cesáro filtered approximation. Lower left) Raised cosine filtered approximation. Lower right) Lanczos filtered approximation.

Fourier series approximation can be expected to be non-oscillatory in physical space as observed in Fig. 8.5. However, if we consider the first zero of the modified kernel it appears at $x = 4\pi/(N+2)$ while the original Dirichlet kernel has it first zero at $x = 2\pi/(N+1)$. A consequence of this is that we should expect a significant smearing of the discontinuity as we observe in Fig. 8.5. Indeed, the smearing is so severe that we loose the ability to accurately identify the location of the discontinuity, in reality reducing the Cesáro filter to a tool of analysis only.

Example 44 (Raised Cosine Filter). This filter is given as

$$\sigma(\eta) = \frac{1}{2}\left(1 + \cos\left(\pi\eta\right)\right) \quad ,$$

and is formally of second order. In fact the application of this filter is equivalent to taking the spatial average as

$$u_j \simeq \frac{u_{j+1} + 2u_j + u_{j-1}}{4} \quad ,$$

as is realized by deriving the modified kernel, Eq.(8.5), on the form

$$S(x) = \frac{1}{4}\left[D_N\left(x - \frac{2\pi}{N}\right) + 2D_N(x) + D_N\left(x + \frac{2\pi}{N}\right)\right] \quad ,$$

where the Fourier kernel, $D_N(x)$, is given in Eq.(8.2).

From Fig. 8.5 we observe that the Cosine filter does not inhibit the Gibbs phenomenon, although compared to the un-filtered approximation the oscillations are severely reduced. Also, from Fig. 8.6 we find that away from the discontinuity the approximation error is clearly reduced by applying the filter.

Example 45 (Lanczos Filter). The Lanczos filter has a filter function as

$$\sigma(\eta) = \frac{\sin \pi\eta}{\pi\eta} \quad ,$$

and is formally only of first order, although we observe that for $\eta = 0$ the filter does satisfy the conditions for being a second order filter. As is evident from Fig. 8.5, applying a Lanczos filter does not inhibit the Gibbs phenomenon, although it clearly results in a strong reduction. From Fig. 8.6 we note that the pointwise error indeed decreases as compared to the un-filtered approximation. However, the error is slightly larger than that obtained by using the Raised Cosine filter.

The classical filters being considered so far all lead to a significant reduction of the Gibbs phenomenon. However, the convergence of the pointwise error remains algebraic in N away from the discontinuity.

We could choose to value recovery of exponential convergence away

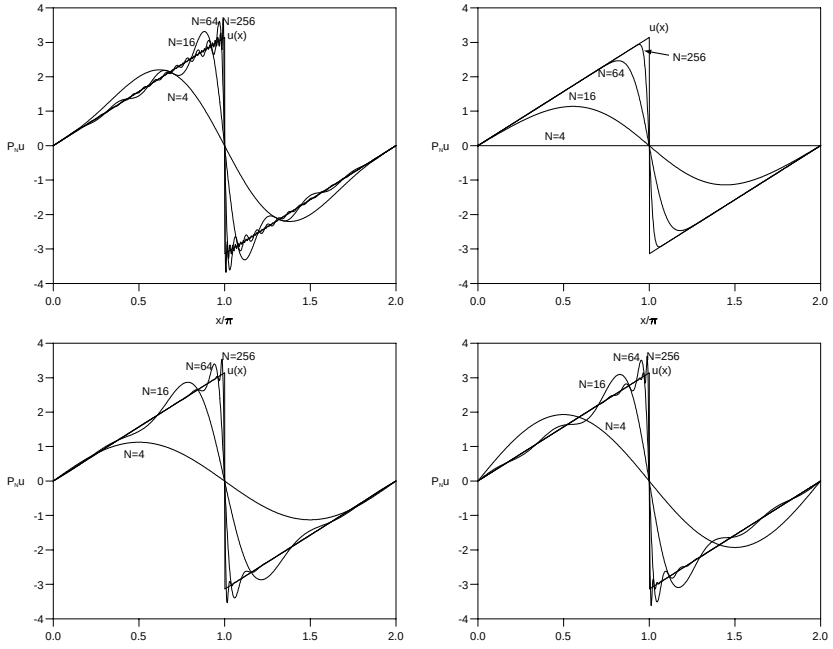

figure 8.7. Effects on the smoothness of the approximation when applying exponential filters of increasing order to a Fourier approximation of a discontinuous function. In all cases we used $\alpha = -\log \varepsilon_M \sim 35$. Upper left) Un-filtered approximation. Upper right) Exponential filter of order 2. Lower left) Exponential filter of order 6. Lower right) Exponential filter of order 10.

from discontinuity higher than actual removal of the Gibbs phenomenon. This can be accomplished by applying exponential filters.

Example 46 (Exponential Filters). This family of filters are defined as

$$\sigma(\eta) = \begin{cases} 1 & |\eta| \leq \eta_c \\ \exp\left(-\alpha\left(\frac{\eta-\eta_c}{1-\eta_c}\right)^p\right) & \eta > \eta_c \end{cases},$$

where α is a measure of how strong the modes should be filtered and p signifies the order of the filter. Note, that the exponential filter does not conform with the definition of a filter as put forward in Def. 8 as $\sigma(1) = \exp(-\alpha)$. However, in practice will we chose α such that $\sigma(1) \simeq \mathcal{O}(\varepsilon_M)$ where ε_M represents the machine accuracy of the actual machine.

In Fig. 8.7 and Fig. 8.8 we illustrate the effect of applying an exponential filter of various order to the discontinuous function introduced in Ex. 40. We have chosen $\alpha = -\log \varepsilon_M$, where ε_M is the machine accuracy, such that mode $|n| = N/2$, i.e., $\eta = 1$, is completely removed. This is by no means a unique choice and results in a fairly strong filtering. We observe that eventhough the Gibbs phenomenon remains, we recover exponential convergence in N away from the discontinuity. The recovery

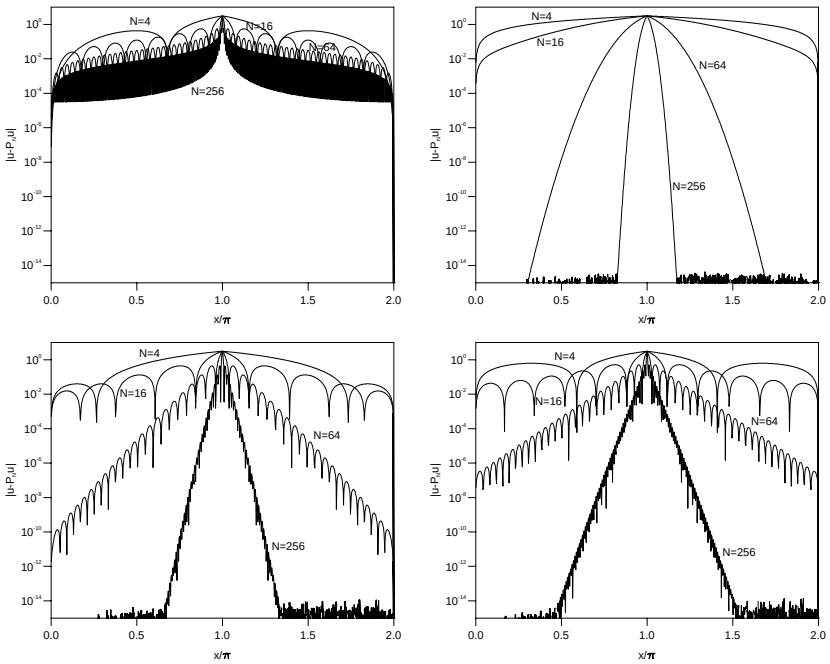

figure 8.8. Pointwise error of the approximation when applying exponential filters of increasing order to a Fourier approximation of a discontinuous function. In all cases we used $\alpha = -\log \varepsilon_M \sim 35$. Upper left) Un-filtered approximation. Upper right) Exponential filter of order 2. Lower left) Exponential filter of order 6. Lower right) Exponential filter of order 10.

of the exponential convergence has made the exponential filter a popular choice when performing simulations of non-linear partial differential equations, where the Gibbs phenomenon may drive an otherwise stable scheme unstable due to amplification of oscillations appearing from the Gibbs phenomenon.

8.2.1.1 The Use of Filters in Fourier Methods.

We may gain a little intuition on the impact of filtering through a simple analogy. The filtered function, $u_N^\sigma(x)$, is found as

$$u_N^\sigma(x) = \sum_{n=-N/2}^{N/2} \sigma(\eta)\hat{u}_n \exp(-inx) \quad .$$

Let us now assume that the filter function, $\sigma(\eta)$, can be approximated as

$$\sigma(\eta) \simeq 1 + c_1\eta^q \quad ,$$

which is a reasonable approximation for all the filters considered in the above provided only that η is not too large. Here c_1 is some constant, specific to the filter.

Let us introduce a new function, $v(x)$, defined as

$$v(x) = u(x) + c_2\frac{\partial^q u}{\partial x^q} \quad ,$$

and continue by expanding $v(x)$ in a Fourier series as

$$\mathcal{P}_N v(x) \simeq \sum_{n=-N/2}^{N/2} \hat{u}_n(1 + c_2(-in)^p)\exp(-inx) \quad ,$$

such that by defining

$$c_2 = \frac{c_1}{(-iN/2)^p} \ll 1 \quad ,$$

we have that $\hat{u}_n = \hat{v}_n$. Hence, the effect of applying a filter in the Fourier expanded function is to modify the function by a small term proportional to a high order derivative of the function. Consequently, in point space it will only have a very localized effect around the point of discontinuity. This corresponds well with what we may observe from the many numerical experiments shown in the last section.

Let us finally consider the details of how filters can be applied and implemented in connection with Fourier spectral methods. As we recall, the use of continuous and discrete expansion coefficients leads to different methods and different implementations when computing derivatives. A similar situation appears for the filtering of Fourier expanded functions.

Filtering of the Continuous Expansion. The practical use of filter in connection with the continuous expansions is straightforward as it only involves summing the filtered series as

$$u_N^\sigma = \sum_{n=-N/2}^{N/2} \sigma\left(\frac{|n|}{N/2}\right) \hat{u}_n \exp(inx) \ ,$$

and likewise for the filtered differentiation operators as

$$\frac{d^q}{dx^q} u_N^\sigma = \sum_{n=-N/2}^{N/2} \sigma\left(\frac{|n|}{N/2}\right) \hat{u}_n^{(q)} \exp(inx) \ .$$

Filtering of the Discrete Expansion. As we have seen previously, there are two mathematically equivalent but computationally different method of expressing the discrete expansions, leading to two different ways of filtering the expansions.

The first method is equivalent to the approach used for the continuous expansion with the only difference being the use of the discrete expansion coefficients rather than the continuous ones.

For the second method, involving matrices rather than summation of series, the scenario is slightly different. Let us first consider the case of an even number of points as

$$x_j = \frac{2\pi}{N} j \ , \quad j \in [0, \ldots, N-1] \ ,$$

with the corresponding approximation

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N-1} u(x_j) g_j(x) \ ,$$

where $g_j(x)$ signifies the interpolating Lagrange polynomial given in Theorem 4.2.2. To recover the filtered approximation

$$u_N^\sigma(x) = \sum_{j=0}^{N-1} u(x_j) g_j^\sigma(x) \ ,$$

we need to obtain the filtered version of interpolating Lagrange polynomial. For a general choice of filter and since $\sigma(\eta)$ is even we may express these as

$$g_j^\sigma(x) = \frac{2}{N}\sum_{n=0}^{N/2}\frac{1}{c_n^\sigma}\sigma\left(\frac{n}{N/2}\right)\cos\left[n(x-x_j)\right] \quad ,$$

where we have introduced the constants $c_0^\sigma = c_{N/2}^\sigma = 2$ and $c_n^\sigma = 1$ otherwise.

This allows for expressing the filtering as a matrix operation

$$u_N^\sigma(x_l) = \sum_{j=0}^{N-1} u_N(x_j) g_j^\sigma(x_l) \quad ,$$

where we note that the filter matrix is a symmetric, Toeplitz matrix.

Likewise, we may obtain matrix forms for the combination of filtering and differentiation where the matrices have the entries

$$\mathrm{D}_{lj}^{(q),\sigma} = \frac{2}{N}\sum_{n=0}^{N/2}\frac{1}{c_n^\sigma}\sigma\left(\frac{n}{N/2}\right)\begin{cases}(in)^q\cos\left[n(x_l-x_j)\right] & q \text{ even}\\ i(in)^q\sin\left[n(x_l-x_j)\right] & q \text{ odd}\end{cases} \quad ,$$

such that the filtered and differentiated approximation is recovered directly as

$$\frac{d^q}{dx^q}u_N^\sigma(x_l) = \sum_{j=0}^{N-1} u_N(x_j)\mathrm{D}_{lj}^{(q),\sigma} \quad .$$

We note that $\mathrm{D}^{(q),\sigma}$ has the same properties as $\mathrm{D}^{(q)}$, i.e., it is a circulant Toeplitz matrix that is symmetric for q being even and skew-symmetric for q being odd.

For completeness we note that the filtered versions of the interpolation Lagrange polynomials and the differentiation matrices for an odd number of collocation points as

$$y_j = \frac{2\pi}{N+1}j \quad , \quad j \in [0,\ldots,N] \quad ,$$

are obtained for the above results by setting $c_n^\sigma = 1$ for all values of n.

8.2.1.2 The Use of Filters in Polynomial Methods.

As for the trigonometric expansions, let us attempt to gain a little understanding of the impact of filters in polynomial expansions. As we shall see there is a small but very important difference between the two

cases.

Let us begin by recalling that the expansion coefficients is given as

$$\hat{u}_n = \frac{1}{\gamma_n} \left(u, \phi_n\right)_w \quad ,$$

and that ϕ_n is the solution to a singular Sturm-Liouville eigenvalue problem as

$$\mathcal{L}\phi_n = -\frac{d}{dx}p(x)\frac{d}{dx}\phi_n = \lambda_n w(x)\phi_n \quad ,$$

where $p(x)$ is singular at $x = \pm 1$ and $w(x)$ signifies the weight.

We consider the filtered approximation

$$u_N^\sigma(x) = \sum_{n=0}^{N} \sigma(\eta)\hat{u}_n\phi_n(x) \quad ,$$

and assume that the filter function, $\sigma(\eta)$, may be approximated as

$$\sigma(\eta) \simeq 1 + c_1\eta^q \quad .$$

We also recall from Section 6.2 that the expansion coefficients in general decay as

$$\hat{u}_n = \frac{1}{\gamma_n\lambda_n^q} \left(u_{(q)}, \phi_n\right)_w \quad ,$$

where

$$u_{(q)} = \frac{1}{w}\mathcal{L}u_{(q-1)} \quad ,$$

and $u_{(0)} = u(x)$. Using that $\lambda_n \sim n^2$ for the ultraspherical polynomials we have the leading term approximation as

$$n^{2q}\hat{u}_n \simeq \frac{1}{\gamma_n} \left(u_{(q)}, \phi_n\right)_w \quad ,$$

or, in other words, an approximate relation as

$$\sum_{n=0}^{N} n^{2q}\hat{u}_n\phi_n(x) \simeq \frac{1}{\gamma_n w^q}\frac{d^q}{dx^q}p(x)\frac{du}{dx} \quad .$$

This implies, as in the case of filtering on trigonometric expansions, that $u(x)$ and its filtered version is related through a high-order derivative as

$$u_N^\sigma(x) \simeq u_N(x) + \frac{1}{\gamma_n w^q} \frac{d^q}{dx^q} p(x) \frac{du}{dx} \quad ,$$

where $2q$ is the order of the filter. At first, this seems fine. However, by recalling that $p(x)$ is singular we realize that filtering of the expansion coefficients in general has no effect on the approximation at the boundaries of the domain as the smoothing high derivative vanishes exactly at $x = \pm 1$. Hence, filtering of the polynomial expansions only has an effect in the interior of the computational domain.

Let us finally consider the practical aspects of employing filters in polynomial expansions, be they based on the continuous or the discrete expansion coefficients.

Filtering of the Continuous Expansion. As for the Fourier expansions, the filtering of the continuous expansions is straightforward as it involves the summation of the filtered series as

$$u_N^\sigma = \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \hat{u}_n \phi_n(x) \quad ,$$

and likewise for the filtered differentiation operators as

$$\frac{d^q}{dx^q} u_N^\sigma = \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \hat{u}_n^{(q)} \phi_n(x) \quad ,$$

where the expansion coefficients of $\hat{u}_n^{(q)}$ can be recovered by using the backward recurrence relation discussed in Sec. 6.3.1.

Filtering of the Discrete Expansion. As we have two equivalent formulations of the discrete polynomial approximation we may also formulate two equivalent ways in which to apply a filter.

Utilizing the discrete expansion coefficients rather than the continuous ones yields a method of filtering equivalent to the one discussed in the above.

However, if we choose to employ the interpolating Lagrange polynomials as the basis of our scheme, the scenario is slightly different. In this case we have

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) \quad ,$$

where x_j signifies a chosen set of grid points and $l_j(x)$ represents the associated Lagrange polynomials. Let us for simplicity assume that the Gauss-Lobatto quadrature points associated with the ultraspherical polynomials, $P_n^{(\alpha)}(x)$, in which case the Lagrange polynomials take the general form

$$l_j(x) = w_j \sum_{n=0}^{N} \frac{P_n^{(\alpha)}(x) P_n^{(\alpha)}(x_j)}{\tilde{\gamma}_n} \quad ,$$

where w_j represents the Gauss-Lobatto quadrature weights given in Eq. (6.3.17). As we recall from Chapter 6, these polynomials may be expressed on closed form by introducing the Christoffel-Darboux formula.

To obtain the filtered approximation

$$u_N^{\sigma}(x) = \sum_{j=0}^{N} u(x_j) l_j^{\sigma}(x) \quad ,$$

we must form the filtered Lagrange polynomials as

$$l_j^{\sigma}(x) = w_j \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \frac{P_n^{(\alpha)}(x) P_n^{(\alpha)}(x_j)}{\tilde{\gamma}_n} \quad ,$$

which we can not in general express on a simple form. However, in this formulation, the action of the filter at the grid points is expressed through a filter matrix as

$$u_N^{\sigma}(x_i) = \sum_{j=0}^{N} \mathrm{F}_{ij} u(x_j) \quad ,$$

where $\mathrm{F}_{ij} = l_j^{\sigma}(x_i)$. It should be noted that F_{ij} is centro-symmetric as a consequence of the symmetry of the quadrature points and that a similar approach can be taken in case Gauss quadrature points is chosen.

Likewise, we may obtain matrix forms for the combination of filtering and differentiation where the matrices have the entries

$$\mathrm{D}_{ij}^{(q),\sigma} = w_j \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \frac{P_n^{(\alpha)}(x_j)}{\tilde{\gamma}_n} \left.\frac{d^q P_n^{(\alpha)}}{dx^q}\right|_{x_i} \quad ,$$

such that the filtered and differentiated approximation is recovered directly as

$$\frac{d^q}{dx^q} u_N^\sigma(x_i) = \sum_{j=0}^{N} \mathrm{D}_{ij}^{(q),\sigma} u(x_j) \quad .$$

The filtered matrices, $\mathrm{D}_{ij}^{(q),\sigma}$, share the properties of the unfiltered versions, i.e., they are centro-antisymmetric for q odd and centro-symmetric when q is even.

8.2.2 Approximation Theory for Filters.

In the previous section we saw that by applying a filter of a given order p we may improve the convergence of the expansions away from the discontinuity. In this section we shall prove this result in a more rigorous sense and, during the development of the proof, attempt to obtain an understanding of what factors, associated with the filter, are important in order to obtain an a priori specified convergence rate away from the point of discontinuity.

For the sake of simplicity we shall restrict the attention to an analysis of the trigonometric expansions and we shall assume that we have the first $2N-1$ expansion coefficients of a piecewise analytic function, u, given and that the function is known to have a point of discontinuity at $x = \xi$. The aim is to recover the value of the function, u, at any point in the interval $[0, 2\pi]$. For this purpose we introduce a filter $\sigma(\eta)$ with the hope that the modified approximation

$$u_N^\sigma(x) = \sum_{n=-N}^{N} \sigma(\eta) \hat{u}_n \exp(inx) \quad ,$$

where $\eta = n/N$, converges faster than the original series

$$u_N(x) = \sum_{n=-N}^{N} \hat{u}_n \exp(inx) \quad .$$

Note that we have changed the notation for reasons of simplicity. Following the definition, Def. 8, of the filter, $\sigma(\eta)$, we have that $\sigma(\eta) = 0$ for $\eta > N$, such that we equally well may express the modified approximation as

$$u_N^\sigma(x) = \sum_{n=-\infty}^{\infty} \sigma(\eta)\hat{u}_n \exp(inx) \ ,$$

to obtain

$$u_N^\sigma(x) = \frac{1}{2\pi}\int_0^{2\pi} S(x-y)u(y)\,dy \ , \tag{8.6}$$

where the filter function is given as

$$S(z) = \sum_{n=-\infty}^{\infty} \sigma(\eta)\exp(inz) \ , \tag{8.7}$$

which is obtained directly by inserting the definition of the continuous expansion coefficients and rearranging the terms. We remind that $z \in [0, 2\pi]$. Note that this expression is equivalent to Eq. (8.5) provided the filter is properly defined. Note also that we in Eq. (8.6) have replaced the truncated series with the actual function, u, thus eliminating the effect of the truncation.

Prior to continuing, let us introduce the concept of a filter function family

Definition 9. *We define a family of periodic filter functions, $S_l(z)$, as*

$$\begin{aligned} S_0(z) &= S(z) \\ S_l'(z) &= S_{l-1}(z) \\ \int_0^{2\pi} S_l(z)\,dz &= 0 \ , \ l \geq 1 \ . \end{aligned}$$

The filter function defined in Eq.(8.7) leads to several equivalent representations of the corresponding filter family as stated in the following lemma.

Lemma 9. Assume that $\sigma(\eta)$ represents a filter of order p, as defined in Def. 8, with the associated filter function

$$S(z) = S_0(z) = \sum_{n=-\infty}^{\infty} \sigma(\eta) \exp(inz) \ ,$$

where $\eta = n/N$. The corresponding filter family, $S_l(z)$, as defined in Def. 9, has the following equivalent representations ($1 \leq l \leq p$);

a)

$$S_l(z) = \frac{1}{N^l} \sum_{n=-\infty}^{\infty} G_l(\eta) i^l \exp(inz) \ ,$$

where

$$G_l(\eta) = \frac{\sigma(\eta) - 1}{\eta^l} \ .$$

b)

$$S_l(z) = \frac{1}{N^{l-1}} \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(iN(z + 2\pi m)\eta\right) G_l(\eta) i^l \, dz \ .$$

c)

$$l = 1 \ : \ S_1(z) = z - \pi + \sum_{\substack{n=-\infty \\ n \neq 0}}^{n=\infty} \sigma(\eta)(in)^{-1} \exp(inx)$$

$$l \geq 2 \ : \ S_l(z) = B_l(z) + \sum_{\substack{n=-\infty \\ n \neq 0}}^{n=\infty} \sigma(\eta)(in)^{-l} \exp(inx) \ ,$$

where $B_l(z)$ represents the Bernoulli polynomial of order l.

The importance of the properties of the filter family and its integrals becomes evident in the following theorem

Theorem 65. *Let $u(x)$ be a piecewise $C^p[0, 2\pi]$ function with a single point of discontinuity at $x = \xi$. Then we have*

$$u_N^\sigma(x) - u(x) = \frac{1}{2\pi} \sum_{l=0}^{p-1} S_{l+1}(c) \left(u^{(l)}(\xi^+) - u^{(l)}(\xi^-) \right) + \frac{1}{2\pi} \int_0^{2\pi} S_p(x-y) u^{(p)}(y) \, dy$$

where $c = x - \xi$ *for* $x > \xi$ *and* $c = 2\pi + x - \xi$ *for* $x < \xi$.

Proof: To establish the proof, let us first realize that an immediate consequence of the actual definition of the filter family as given in Def. 9 is

$$\begin{aligned} S_1(2\pi) - S_1(0) &= 2\pi \\ S_l(2\pi) - S_l(0) &= 0 \quad , \quad l \geq 2 \quad , \end{aligned} \tag{8.8}$$

as can be seen by using expression c) for $S_l(z)$ given in Lemma 9 for $l = 1$ and the integral condition on $S_l(z)$ for $l > 1$.

We continue by considering the case of $x > \xi$. Integrating Eq. (8.6) by parts p times, being careful not to integrate over the point of discontinuity, yields

$$\begin{aligned} 2\pi u_N^\sigma(x) &= \int_0^{\xi^-} S(x-y)u(y)\,dy + \int_{\xi^+}^{2\pi} S(x-y)u(y)\,dy \\ &= \sum_{l=0}^{p-1} \left(S_{l+1}(2\pi) - S_{l+1}(0)\right) u^{(l)}(x) \\ &\quad + \sum_{l=0}^{p-1} S_{l+1}(x-\xi)\left(u^{(l)}(\xi^+) - u^{(l)}(\xi^-)\right) \\ &\quad + \int_0^{2\pi} S_p(x-y)u^{(p)}(y)\,dy \\ &= 2\pi u(x) + \sum_{l=0}^{p-1} S_{l+1}(x-\xi)\left(u^{(l)}(\xi^+) - u^{(l)}(\xi^-)\right) \\ &\quad + \int_0^{2\pi} S_p(x-y)u^{(p)}(y)\,dy \quad , \end{aligned}$$

where the last reduction follows from Eq.(8.8) and we have used that u as well as S_l are assumed to be periodically extended. Also note that we use the notation that

$$\int_0^{2\pi} \quad dy = \int_0^{\xi^-} \quad dy + \int_{\xi^+}^{2\pi} \quad dy \quad .$$

Likewise, we obtain for $x < \xi$ the result

$$\begin{aligned}
2\pi u_N^\sigma(x) &= \int_0^{\xi^-} S(x-y)u(y)\,dy + \int_{\xi^+}^{2\pi} S(x-y)u(y)\,dy \\
&= \sum_{l=0}^{p-1} \left(S_{l+1}(2\pi) - S_{l+1}(0)\right) u^{(l)}(x) \\
&\quad + \sum_{l=0}^{p-1} S_{l+1}(2\pi + x - \xi)\left(u^{(l)}(\xi^+) - u^{(l)}(\xi^-)\right) \\
&\quad + \int_0^{2\pi} S_p(x-y)u^{(p)}(y)\,dy \\
&= 2\pi u(x) + \sum_{l=0}^{p-1} S_{l+1}(2\pi + x - \xi)\left(u^{(l)}(\xi^+) - u^{(l)}(\xi^-)\right) \\
&\quad + \int_0^{2\pi} S_p(x-y)u^{(p)}(y)\,dy \quad ,
\end{aligned}$$

which proves the theorem. QED

The result stated in Theorem 65 provides a precise estimate of the error between the unknown point value of $u(x)$ and its filtered and truncated approximating series, $u_N^\sigma(x)$. The goal of the following is to estimate the two terms on the right hand side of the expression in Theorem 65.

Let us first consider the last term in Theorem 65. The estimation of this is classic. Indeed, if $u(x) \in C^{p-1}[0, 2\pi]$ the first term of Theorem 65 vanishes and we are left with the last term only. Thus, this last term is really the error term for smooth functions as stated in the following lemma.

Lemma 10. Let $S_l(x)$ be defined as in Eq. (8.5) and Def. 9 then

$$\frac{1}{2\pi}\left|\int_0^{2\pi} S_p(x-y)u^{(p)}(y)\,dy\right| \leq C\frac{\sqrt{N}}{N^p}\left(\int_0^{2\pi}\left|u^{(p)}\right|^2 dx\right)^{1/2} \quad ,$$

where C is independent of C and u.

Proof: Applying the Cauchy-Schwartz inequality yields

$$\frac{1}{2\pi}\left|\int_0^{2\pi} S_p(x-y)u^{(p)}(y)\,dy\right| \leq \frac{1}{2\pi}\left(\int_0^{2\pi} S_p^2(x)\,dx\right)^{1/2}\left(\int_0^{2\pi}\left|u^{(p)}(x)\right|^2 dx\right)^{1/2}$$

using that S_p is periodic. To estimate the first term we express S_p using a) of Lemma 9 to obtain

$$\begin{aligned}
\frac{1}{2\pi}\int_0^{2\pi} S_p^2(x)\,dx &= \frac{1}{2\pi}\int_0^{2\pi}\sum_{n=-\infty}^{\infty}\left(\frac{1}{N^p}\frac{\sigma(\eta)-1}{\eta^p}i^p\exp(inx)\right)^2 \\
&= \sum_{n=-\infty}^{\infty}\frac{1}{N^{2p}}\left(\frac{\sigma(\eta)-1}{\eta^p}\right)^2 \\
&= \frac{1}{N^{2p-1}}\sum_{n=-N}^{N}\frac{2}{2N}\left(\frac{\sigma(\eta)-1}{\eta^p}\right)^2 + \sum_{|n|>N}\frac{1}{n^{2p}} \\
&\leq \frac{1}{N^{2p-1}}\int_{-1}^{1}\left(\frac{\sigma(\eta)-1}{\eta^p}\right)^2 d\eta + \frac{1}{N^{2p-1}} \\
&\leq C\frac{N}{N^{2p}}\ ,
\end{aligned}$$

where we have used the orthogonality of exponential function and bounded the Riemann sum by its integral which is convergent under condition c) of Definition 8 as can be proved by partial integration. QED

Let us now turn to the estimate for the first term of the expression given in Theorem 65. We will show that the conditions given in Lemma 9 implies that this is bounded.

Lemma 11. Let $S_l(x)$ be defined as in Eq.(8.5) and Def. 9 for $l \geq 1$, then

$$|S_l(x)| \leq C\frac{1}{N^{p-1}}\left(\frac{1}{|x|^{p-l}} + \frac{1}{|2\pi - x|^{p-l}}\right)\int_{-\infty}^{\infty}|G_l^{(p-l)}(\eta)|\,d\eta\ ,$$

where $G_l(\eta)$ is defined in Lemma 9.

Proof: Consider representation b) in Lemma 9 of $S_l(x)$. Since $G_l(x)$ is $p-1$ times differentiable and the filter is defined such that $\sigma^{(l)}(0) = \sigma^{(l)}(1) = 0$, we obtain by $p-l$ times partial integration that

$$|S_l(x)| = \frac{1}{N^{l-1}}\left|\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\exp[iN(x+2\pi m)\eta]}{N^{p-l}(x+2\pi m)^{p-l}}G_l^{(p-l)}(\eta)\,d\eta\right|\ .$$

Since $x \in [0, 2\pi]$, the dominating terms are found for $m = 0$ and $m = -1$, such that, using the triangle inequality and taking these two contribu-

tions outside the integral we obtain the result. QED

We are now in a position to state the main theorem as

Theorem 66. *Let $u(x)$ be a piecewise $C^p[0,2\pi]$ function with only one point of discontinuity at $x=\xi$. Assume also that the filter $\sigma(\eta)$ satisfies Def 8. Let us now denote the distance between a point $x \in [0,2\pi]$ and the discontinuity as $d(x)=|x-\xi|$ and define the filtered truncated series approximation to $u(x)$ as*

$$u_N^\sigma(x) = \sum_{n=-N}^{N} \sigma(\eta)\hat{u}_N \exp(inx) \quad .$$

The pointwise difference between $u(x)$ and $u_N^\sigma(x)$ at all x with the exception of ξ is then bounded as

$$|u(x)-u_N^\sigma(x)| \leq C_1 \frac{1}{N^{p-1}} \frac{1}{d(x)^{p-1}} K(u) + C_2 \frac{\sqrt{N}}{N^p} \left(\int_0^{2\pi} \left| u^{(p)} \right|^2 \, dx \right)^{1/2} \quad ,$$

where

$$K(u) = \sum_{l=0}^{p-1} d(x)^l \left| u^{(l)}(\xi^+) - u^{(l)}(\xi^-) \right| \int_{-\infty}^{\infty} \left| G_l^{(p-l)}(\eta) \right| \, d\eta \quad .$$

Proof: The second part of the estimate follows directly from Lemma 10. From Lemma 11 we know that the first part of the expression in Theorem 65 is bounded.

Since $c = x-\xi$ for $x > \xi$ and $c = 2\pi + x - \xi$ for $x < \xi$ in Theorem 65 we have that

$$\frac{1}{|c|^{p-l}} + \frac{1}{|2\pi - c|^{p-l}} \leq \frac{2}{d(x)^{p-l}} \quad .$$

Combining this with the estimate of Lemma 11 yields the proof. QED

Note that for $u(x) \in C^p[0,2\pi]$ we recover the result of Theorem 66 as expected.

This theorem proves that the filtering process works away from the discontinuity as all terms can be bounded by $\mathcal{O}(N^{1-p})$, such that provided $d > 0$ convergence depends only on the regularity of the piecewise continuous function away from the discontinuity and the order of the filter.

In Examples 43-45 we considered classical types of filters that all fall under the appropriate definition of a filter as put forward in Def. 8, such that the expected convergence rate away from the point of discontinuity may be predicted using Theorem 66. Indeed, remembering that the Cesáro and the Lanczos are both first order filter, we should expect no more than first order convergence as is also seen in Fig. 8.6. Likewise, for the case of the Raised Cosine filter, being a second order filter, we find in Fig. 8.6 the expected second order convergence away from the point of discontinuity.

The case for the exponential filter is somewhat different in the sense that this filter does not conform with Definition 8, thereby introducing an additional error term into the above analysis. However, in most computations we will set the parameter α such that $\sigma(1) \simeq \varepsilon_M$ making it plausible that the conclusions from the above analysis carries over which is also confirmed from Example 46 where we see that convergence is achieved away from the discontinuity and with a convergence rate being close to spectral.

8.3 The Resolution of the Gibbs Phenomenon

8.3.1 General Theory.

8.3.2 Reconstruction for Trigonometric Expansions.

8.3.3 Reconstruction for Polynomial Expansions.

9

Computational Aspects

The purpose of the previous chapters has been the development of spectral methods from a theoretical point of view, a second and equally important issue when solving partial differential equations is the question of efficient implementations of the methods for problems of a more general character than we have discussed hitherto.

This chapter is devoted to a discussion of tools that allows for efficient implementations of spectral methods based on trigonometric as well as polynomial expansions. We shall also discuss the problems, and quite significant they are, of round-off errors in spectral methods and address the issues of aliasing in connection with quadratic nonlinearities and how to remove the aliasing errors in such cases.

Finally, we address problem requiring the use of mappings before we turn to the important discussion of how to extend the general one-dimensional framework to problems in multiple dimensions.

9.1 Fast Computation of Interpolation and Differentiation

Only at very rare instances does a numerical algorithm appear that in a few year revolutionizes entire fields within science and engineering. It is, however, fair to say that the appearance of the Fast Fourier Transform in 1965 did exactly that. By establishing a fast way of evaluating discrete Fourier series and their inverses, this single algorithm opened up for the use of methods that were hitherto considered less interesting due to excessive computational requirements.

Among methods that benefited enormously from the development

of the Fast Fourier Transforms (FFT) are the spectral methods and a significant part of the fundamental theory of spectral methods were developed in the years immediately following the introduction of the FFT.

Unfortunately, the idea behind the FFT is valid only when dealing with trigonometric polynomials in some form. Hence, these methods are not in general applicable to the expansions based on orthogonal polynomials.

In the following we shall briefly discuss the idea behind the FFT and its relatives. However, since the polynomials play such an important role in spectral methods we shall continue by discussing alternative fast methods for the computation of interpolation and differentiation for such methods. We conclude by including a brief section on how to compute the general Gaussian quadrature points and weights necessary to compute the discrete polynomial expansion coefficients.

9.1.1 Fast Fourier Transforms.

Let us first recall the discrete Fourier series expansion of a function, $u \in L^2[0, 2\pi]$, using an even number of points

$$x_j = \frac{2\pi}{N} j \quad , \quad j \in [0, N-1] \ ,$$

given as

$$\mathcal{I}_N u(x) = \sum_{n=-N/2}^{N/2} \tilde{u}_n \exp(-inx) \quad , \quad \tilde{u}_n = \frac{1}{c_n N} \sum_{j=0}^{N-1} u(x_j) \exp(inx_j) \quad .$$

If we now restrict the attention to $x = x_j$, i.e. interpolation at the collocation points and recall that $\tilde{u}_{-N/2} = \tilde{u}_{N/2}$ is assumed for uniqueness, we arrive at the discrete Fourier expansion on the form

$$\mathcal{I}_N u(x_j) = \sum_{n=-N/2}^{N/2-1} \tilde{u}_n \exp\left(-i\frac{2\pi}{N} jn\right) \quad , \quad \tilde{u}_n = \frac{1}{N} \sum_{j=0}^{N-1} u(x_j) \exp\left(i\frac{2\pi}{N} jn\right) \quad , \tag{9.1}$$

i.e. we have N terms in both summations. Direct evaluation of the expansion coefficients or interpolation to the grid, x_j, requires $\mathcal{O}(8N^2)$ real operations since $\tilde{u}_n$ in general is a complex number.

Let us now assume that N is a power of two, i.e. $N = 2^M$, and introduce the new index, j_1, as

$$j = \begin{cases} 2j_1 & j \text{ even} \\ 2j_1 + 1 & j \text{ odd} \end{cases} , \quad j_1 \in [0, N/2 - 1] \ .$$

Using this new index, we express the discrete expansion coefficients

$$\tilde{u}_n = \frac{1}{N} \left(\sum_{j_1=0}^{N/2-1} u(x_{2j_1}) \exp\left[i \frac{2\pi}{N} (2j_1) n \right] + \sum_{j_1=0}^{N/2-1} u(x_{2j_1+1}) \exp\left[i \frac{2\pi}{N} (2j_1 + 1) n \right] \right)$$

Introducing $N_1 = N/2$, we obtain

$$\tilde{u}_n = \frac{1}{N} \left(\sum_{j_1=0}^{N_1-1} u(x_{2j_1}) \exp\left[i \frac{2\pi}{N_1} j_1 n \right] + \exp\left[i \frac{2\pi}{N} n \right] \sum_{j_1=0}^{N_1-1} u(x_{2j_1+1}) \exp\left[i \frac{2\pi}{N_1} j_1 n \right] \right) \ .$$

At this point we realize that the two sums, each of half the length of the original sum, have exactly the same form as the original sum, Eq.(9.1). Thus, we may repeat the process and break the computation down to 4 sums each of length $N/4$. Repeating this process results in an algorithm that computes the discrete expansion coefficients in $\mathcal{O}(5N \log_2 N)$ real operations provided the twiddle factors, $\exp(2\pi n/N)$, $\exp(2\pi n/N_1)$ and so on, are precomputed. This decomposition is precisely the Fast Fourier Transform and the reduction, as we saw, in computations are significant.

The application of the FFT for computing the interpolating follows the same line of thought. Indeed, if we consider

$$u(x_j) = \sum_{n=-N/2}^{N/2-1} \tilde{u}_n \exp\left[-i \frac{2\pi}{N} j n \right] \ ,$$

and introduce the new index, n_1, as

$$n = \begin{cases} 2n_1 & n \text{ even} \\ 2n_1 + 1 & n \text{ odd} \end{cases} , \quad n_1 \in [-N/4, N/4 - 1] \ ,$$

we recognize the exact same structure that made the fast computation of the expansion coefficients possible. Hence, the FFT is applicable for computing the discrete expansion coefficients as well as the interpolation

at the collocation points.

At this point in time numerous highly efficient and accurate implementations of the FFT algorithm exists and it is in general safe to say that when considering Fourier methods one should always use the FFT if possible. The brief account for the idea behind the FFT presented above introduces the assumption that N is a power of 2. However, it is easy to see that also if N is a power a 3 can one design a fast transform, by splitting the original sum into three sums each of length $1/3N$ - such an algorithm being known as a radix-3 transform. Indeed, the algorithm can be formulated for any prime-decomposition of a given number, allowing for the use of a very wide range of values of N.

It should also be clear that choosing the alternative and very appealing grid set

$$y_j = \frac{2\pi}{N+1} \ , \ j \in [0, N] \ ,$$

inhibits the use of the FFT as the decomposition into smaller sums is impossible with this choice of grid points.

In many applications of spectral methods the direct use of the complex FFT is unnecessarily expensive, e.g., $u(x)$ is in many case a real function. If we consider a real function of length N we may construct a complex function, v, of half the length as

$$v_j = u(x_{2j}) + iu(x_{2j+1}) \ \ , j \in [0, N/2 - 1] \ .$$

Applying the FFT yields the complex expansion coefficients of v as

$$\tilde{v}_n = \tilde{v}_n^e + i\tilde{v}_n^o = \frac{1}{N} \sum_{j=0}^{N/2-1} u(x_{2j}) \exp\left(\frac{2\pi}{N/2} nj\right) + i\frac{1}{N} \sum_{j=0}^{N/2-1} u(x_{2j+1}) \exp\left(\frac{2\pi}{N/2} nj\right) \ ,$$

for $n \in [0, N/2 - 1]$. Recalling the development of the FFT algorithm, the first level of decomposition establishes the relationship

$$\forall n \in [-N/2, N/2 - 1] : \ \tilde{u}_n = \tilde{v}_n^e + \exp\left[i\frac{2\pi}{N} n\right] \tilde{v}_n^o \ .$$

Moreover, since $u(x)$ is real we have that $\tilde{u}_n = \overline{\tilde{u}}_{-n}$, i.e. we need only compute the elements $n \in [0, N/2]$ complex numbers. Using $\tilde{v}_0 = \tilde{v}_{N/2}$, implies that these two components are purely real, and we obtain that only $N/2$ complex numbers need to be stored, i.e. the transformation can be done in place. Using the decomposition, we recover the discrete

expansion coefficients as the sum of the even and odd parts of $\tilde{v}$ as

$$\tilde{u}_n = \frac{1}{2}\left(\tilde{v}_n + \overline{\tilde{v}}_{N/2-n}\right) - \frac{i}{2}\exp\left[i\frac{2\pi}{N}n\right]\left(\tilde{v}_n - \overline{\tilde{v}}_{N/2-n}\right) \quad .$$

To obtain the interpolation at the grid points we assemble the two components of $\tilde{v}_n$ as

$$\tilde{v}_n = \frac{1}{2}\left(\tilde{u}_n + \tilde{u}_{N/2-n}\right) + \frac{i}{2}\exp\left[-i\frac{2\pi}{N}n\right]\left(\tilde{u}_n - \overline{\tilde{u}}_{N/2-n}\right) \quad .$$

which recovers the exact same structure as for the forward transform. The fast transformation using the complex FFT yields an algorithm as $\mathcal{O}(\frac{5}{2}N\log_2 N)$ for the radix-2 scheme, i.e. twice as fast as just performing the complex FFT with all imaginary components being zero.

The FFT may also be applied for the fast computation of Sine transformations although it requires a little additional work. If we first consider the discrete Sine expansion coefficients given as

$$\tilde{u}_n = \frac{1}{\gamma_n}\sum_{j=1}^{N-1} u(x_j)\sin\left(\frac{\pi}{N}nj\right) \quad ,$$

it is clear that there is a factor of 2 difference in the argument of the Sine transformation compared to Eq.(9.1). However, let us extend the function, $u(x)$, around $j = N$ and introduce a new odd function as

$$\forall k \in [0, 2N-1]: \; v(x_k) = \begin{cases} u(x_k) & k < N \\ 0 & k = N \\ -u(x_{2N-k}) & k > N \end{cases} \quad .$$

Doing an FFT on $v(x)$ yields

$$\tilde{v}_n = \frac{1}{2N}\sum_{k=0}^{2N-1} v(x_k)\exp\left[i\frac{2\pi}{2N}kn\right] \quad ,$$

which we may rewrite as

$$\tilde{v}_n = \sum_{k=0}^{N-1} u(x_k)\exp\left[i\frac{2\pi}{2N}kn\right] - \sum_{k=N}^{2N-1} u(x_{2N-k})\exp\left[i\frac{2\pi}{2N}kn\right]$$

$$= \sum_{j=1}^{N-1} u(x_j) \exp\left[i\frac{2\pi}{2N}kn\right] - \sum_{j=1}^{N-1} u(x_j) \exp\left[i\frac{2\pi}{2N}(2N-j)n\right]$$
$$= 2i \sum_{j=1}^{N-1} u(x_j) \sin\left(\frac{\pi}{N}nj\right) \quad ,$$

i.e. the FFT of the function v exactly results in the discrete Sine expansion coefficients, although, at first it seems at the expense of computing an expansion of twice the length. However, since $u(x)$ is real the expansion can be performed using of FFT of half length for real functions.

The exact same trick can be applied for computing the discrete Cosine expansion coefficients given as

$$\tilde{u}_n = \sum_{j=0}^{N} u(x_j) \cos\left(\frac{\pi}{N}nj\right) \quad ,$$

although in this case we must form a new even function as

$$\forall k \in [0, 2N-1] : \; v(x_k) = \begin{cases} u(x_k) & k \leq N \\ u(x_{2N-k}) & k > N \end{cases} \quad ,$$

such that the Cosine expansion coefficients are obtained as

$$\tilde{v}_n = \sum_{k=0}^{2N-1} v(x_k) \exp\left[i\frac{2\pi}{2N}kn\right] = 2\sum_{j=0}^{N} u(x_j) \cos\left(\frac{\pi}{N}nj\right) \quad .$$

Again we can use the complex transform of a real function to minimize the computational workload. Thus, we have fast transforms for the Sine series, and, much more importantly, for the Cosine series which is very close to the Chebyshev expansion. Indeed, if we recall the Chebyshev Gauss-Lobatto quadrature rule and expansion

$$\mathcal{I}_N u(x_j) = \sum_{n=0}^{N} \tilde{u}_n \cos\left(\frac{\pi}{N}nj\right) \quad , \quad \tilde{u}_n = \frac{1}{\overline{c}_n N} \sum_{j=0}^{N} \frac{1}{\overline{c}_j} u(x_j) \cos\left(\frac{\pi}{N}nj\right) \quad ,$$

it is immediate that the FFT can be used to compute both sums in $\mathcal{O}(\frac{5}{2}N\log_2 N + 4N)$, where the latter contribution appears from the packing and unpacking required to utilize the FFT. The option of using the FFT for computing the Chebyshev Gauss-Lobatto expansion is yet another reason for the wide use of Chebyshev polynomials for the construction of efficient spectral methods.

It should be noted that implementations of the Cosine transforms are of very varying quality and it is in general not possible to estimate when a fast Cosine transform should be used at a specific machine rather than applying a matrix multiply directly. However, a good rule of thump is that if $N > 64$ it is most certainly worth to consider the use of a fast Cosine transform.

9.1.2 The Even-Odd Decomposition.

Unfortunately, the approach that leads to the Fast Fourier Transform does not extend to orthogonal polynomials beyond the Chebyshev polynomials. Hence, if one insists on using expansion coefficients, $\tilde{u}_n$, to compute the derivative at the collocation points, there is in general no way around summing the series directly.

However, using the interpolating Lagrange polynomials and the associated differentiation matrices, there is still room for improvement. Let us consider the vector, $\boldsymbol{u} = (u(x_0), \ldots, u(x_N))$, and the differentiation matrix, D, associated with the chosen set of collocation points, x_j. We recall that the derivative of $\boldsymbol{u}$ at the grid points, $\boldsymbol{u}'$, is obtained as

$$\boldsymbol{u}' = \mathrm{D}\boldsymbol{u} \quad .$$

Using ultraspherical polynomials as the basis for the approximation we have previous established that

$$\mathrm{D}_{ij} = -\mathrm{D}_{N-i,N-j} \quad ,$$

i.e. that D is centro-antisymmetric. This property shall allow us to develop a fast algorithm for the computation of $\boldsymbol{u}'$.

We shall decompose $\boldsymbol{u}$ into its even, $\boldsymbol{e}$, and odd parts, $\boldsymbol{o}$, as $\boldsymbol{u} = \boldsymbol{e} + \boldsymbol{o}$ where the two new vectors have the entries

$$e_j = \frac{1}{2}\left(u_j + u_{N-j}\right) \quad , \quad o_j = \frac{1}{2}\left(u_j - u_{N-j}\right) \quad ,$$

where u_j signifies entry j in $\boldsymbol{u}$ and similarly for e_j and o_j. We observe that

$$e_j = e_{N-j} \quad , \quad o_j = -o_{N-j} \quad .$$

By linearity of the differentiation operation we have

$$\boldsymbol{u}' = \mathrm{D}\boldsymbol{u} = \mathrm{D}\boldsymbol{e} + \mathrm{D}\boldsymbol{o} = \boldsymbol{e}' + \boldsymbol{o}' \ .$$

Let us now first consider the case where N is odd such that we in total have an even number of collocation points.

If we compute the differentiation of $\boldsymbol{e}$ we obtain

$$\begin{aligned} e'_j &= \sum_{i=0}^{N} \mathrm{D}_{ji} e_i = \sum_{i=0}^{(N-1)/2} \mathrm{D}_{ji} e_i + \mathrm{D}_{j,N-i} e_{N-i} \\ &= \sum_{i=0}^{(N-1)/2} \left(\mathrm{D}_{ji} + \mathrm{D}_{j,N-i}\right) e_i \ . \end{aligned}$$

Moreover, $\boldsymbol{e}'$ is odd as

$$\begin{aligned} e'_{N-j} &= \sum_{i=0}^{N} \mathrm{D}_{N-j,i} e_i = \sum_{i=0}^{(N-1)/2} \left(\mathrm{D}_{N-j,i} + \mathrm{D}_{N-j,N-i}\right) e_i \\ &= \sum_{i=0}^{(N-1)/2} -\left(\mathrm{D}_{j,N-i} + \mathrm{D}_{ji}\right) e_i = -e'_j \ . \end{aligned}$$

Hence, it is only necessary to compute the first half of $\boldsymbol{e}'$. If we introduce the matrix, D^e, with the entries

$$\forall i, j \in [0, (N-1)/2] : \ \mathrm{D}^e_{ij} = \mathrm{D}_{ij} + \mathrm{D}_{i,N-j} \ ,$$

the computation is simply a matrix multiply as

$$\tilde{\boldsymbol{e}}' = \mathrm{D}^e \tilde{\boldsymbol{e}} \ ,$$

where $\tilde{\boldsymbol{e}} = (e_0, \ldots, e_{(N-1)/2})^T$ and similarly for $\tilde{\boldsymbol{e}}'$.

The computation of $\boldsymbol{o}'$ yields

$$\begin{aligned} o'_j &= \sum_{i=0}^{N} \mathrm{D}_{ji} o_i = \sum_{i=0}^{(N-1)/2} \mathrm{D}_{ji} o_i + \mathrm{D}_{j,N-i} o_{N-i} \\ &= \sum_{i=0}^{(N-1)/2} \left(\mathrm{D}_{ji} - \mathrm{D}_{j,N-i}\right) o_i \ , \end{aligned}$$

and $\boldsymbol{o}'$ is even as

$$o'_{N-j} = \sum_{i=0}^{N} \mathrm{D}_{N-j,i} o_i = \sum_{i=0}^{(N-1)/2} \left(\mathrm{D}_{N-j,i} - \mathrm{D}_{N-j,N-i}\right) o_i$$
$$= \sum_{i=0}^{(N-1)/2} \left(-\mathrm{D}_{j,N-i} + \mathrm{D}_{ji}\right) o_i = o'_j \ ,$$

i.e. also in this case is it sufficient to compute one half of the elements in the vector. Introducing the matrix, D^o, with the entries

$$\forall i,j \in [0,(N-1)/2] : \ \mathrm{D}^o_{ij} = \mathrm{D}_{ij} - \mathrm{D}_{i,N-j} \ ,$$

we recover the matrix multiplication as

$$\tilde{\boldsymbol{o}}' = \mathrm{D}^o \tilde{\boldsymbol{o}} \ ,$$

where $\tilde{\boldsymbol{o}} = (o_0, \ldots, o_{(N-1)/2})^T$ and similarly for $\tilde{\boldsymbol{o}}'$.

Finally, we can reconstruct $\boldsymbol{u}'$ as

$$\boldsymbol{u}' = \boldsymbol{e}' + \boldsymbol{o}' \ ,$$

using

$$u'_j = \tilde{e}'_j + \tilde{o}'_j \ , \ \ u'_{N-j} = e'_{N-j} + o'_{N-j} = -\tilde{e}'_j + \tilde{o}'_j \ ,$$

for $j \in [0,(N-1)/2]$.

Consequently, to compute the derivative of $\boldsymbol{u}$ at the collocation points we need to construct $\tilde{\boldsymbol{e}}$ and $\tilde{\boldsymbol{o}}$, perform two matrix-vector product of length $N/2$ and reconstruct $\boldsymbol{u}'$. The total operation count for this process becomes

$$2\frac{N}{2} + 2\left(2\left(\frac{N}{2}\right)^2 - \frac{N}{2}\right) + 2\frac{N}{2} = N^2 + N \ ,$$

provided the differentiation matrices are precomputed. This should be contrasted to the direct computation of $\boldsymbol{u}'$ which requires $2N^2 - N$ operations. Hence, utilizing the centro-antisymmetry of the differentiation matrix allows for decreasing the computational work with close to a factor of 2 which is very important for N being large.

Let us finally consider the case where N is even. If we first consider

differentiation of the even part of $\boldsymbol{u}$ and follow the outlined approach we obtain

$$e'_j = \sum_{i=0}^{N} \mathrm{D}_{ji} e_i = \sum_{i=0}^{N/2} \mathrm{D}_{ji} e_i + \mathrm{D}_{j,N-i} e_{N-i} \ ,$$

i.e. the term from $i = N/2$ is computed twice. This, however, is easily fixed by defining D^e slightly different as

$$\forall i \in [0, N/2-1], j \in [0, N/2] : \ \mathrm{D}^e_{ij} = \begin{cases} \mathrm{D}_{ij} + \mathrm{D}_{i,N-j} & j \neq N/2 \\ \mathrm{D}_{i,N/2} & j = N/2 \end{cases} \ .$$

Note that D^e is rectangular rather than quadratic. This is a consequence of $\boldsymbol{e}'$ being odd, i.e. $e'_{N/2} = 0$ and need not be computed.

In the same way we define a modified D^o as

$$\forall i \in [0, N/2], j \in [0, N/2-1] : \ \mathrm{D}^o_{ij} = \mathrm{D}_{ij} - \mathrm{D}_{i,N-j} \ ,$$

since $\boldsymbol{o}$ is odd such that the problem of counting the last entry twice does not appear in this case. This also implies that D^o is rectangular, since $\boldsymbol{o}'$ is even and therefore $o'_{N/2}$ needs to be computed.

In the situation where D is centro-symmetric as

$$\mathrm{D}_{ij} = \mathrm{D}_{N-i,N-j} \ ,$$

the exact same even-odd splitting can be applied with the only difference being that $\boldsymbol{e}'$ is even and $\boldsymbol{o}'$ is odd such that the final reconstruction becomes

$$u'_j = \tilde{e}'_j + \tilde{o}'_j \ , \ \ u'_{N-j} = e'_{N-j} + o'_{N-j} = \tilde{e}'_j - \tilde{o}'_j \ .$$

Indeed, all even order differentiation matrices appearing from the ultraspherical polynomials share the property of centro-symmetry, thus allowing for applying the splitting technique directly.

9.2 Computation of Gaussian Quadrature Points and Weights

When using polynomial methods the first requirement is to establish the position of the collocation points, these being the quadrature points of some chosen Gauss quadrature rule. However, with the exception of a few special cases, like the Chebyshev polynomials, no closed form

expression for the quadrature nodes are known. Nevertheless, as we shall discover in the following, there is a simple and elegant way of computing these nodes as well as the corresponding weights, although these are known on explicit form and can be computed using these.

We shall, as usual, restrict the attention to the case of ultraspherical polynomials, $P_n^{(\alpha)}(x)$, although we note that everything generalizes to the case of Jacobi polynomials as well as Laguerre and Hermite polynomials.

Let us begin by recalling the three-term recurrence relation for ultraspherical polynomials as

$$xP_n^{(\alpha)}(x) = a_{n-1,n}P_{n-1}^{(\alpha)}(x) + a_{n+1,n}P_{n+1}^{(\alpha)}(x) \quad , \tag{9.2}$$

where the recurrence coefficients are given as

$$a_{n-1,n} = \frac{n+2\alpha}{2n+2\alpha+1} \quad , \quad a_{n+1,n} = \frac{n+1}{2n+2\alpha+1} \quad .$$

Let us now normalize the polynomials slightly different and introduce the modified polynomials

$$\tilde{P}_n^{(\alpha)}(x) = \frac{1}{\sqrt{\gamma_n}}P_n^{(\alpha)}(x) \quad ,$$

such that $(\tilde{P}_n^{(\alpha)}, \tilde{P}_k^{(\alpha)})_w = \delta_{nk}$. With this normalization, the recurrence coefficients of the three-term recurrence relation, Eq.(9.2), becomes

$$\tilde{a}_{n-1,n} = \sqrt{\frac{\gamma_{n-1}}{\gamma_n}}a_{n-1,n} = \sqrt{\frac{n(n+2\alpha)}{(2n+2\alpha+1)(2n+2\alpha-1)}} \quad ,$$

and

$$\tilde{a}_{n+1,n} = \sqrt{\frac{\gamma_{n+1}}{\gamma_n}}a_{n+1,n} = \sqrt{\frac{(n+1)(n+2\alpha+1)}{(2n+2\alpha+3)(2n+2\alpha+1)}} \quad .$$

The key observation to make is that

$$\beta_n = \tilde{a}_{n+1,n} = \tilde{a}_{n,n+1} \quad ,$$

such that Eq.(9.2) reduces to

$$x\tilde{P}_n^{(\alpha)}(x) = \beta_{n-1}\tilde{P}_{n-1}^{(\alpha)}(x) + \beta_n\tilde{P}_{n+1}^{(\alpha)}(x) \quad . \tag{9.3}$$

If we now introduce the vector $\tilde{\boldsymbol{P}}^{(\alpha)}(x) = (\tilde{P}_0^{(\alpha)}(x), \ldots, \tilde{P}_N^{(\alpha)}(x))^T$, and the symmetric bi-diagonal matrix

$$\mathrm{J}_N = \begin{bmatrix} 0 & \beta_1 & 0 & 0 & 0 & \cdots & 0 \\ \beta_1 & 0 & \beta_2 & 0 & 0 & \cdots & 0 \\ 0 & \beta_2 & 0 & \beta_3 & 0 & \cdots & 0 \\ 0 & 0 & \beta_3 & 0 & \beta_4 & \cdots & 0 \\ 0 & 0 & 0 & \beta_4 & 0 & \ddots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \beta_{N-1} \\ 0 & 0 & 0 & 0 & 0 & \beta_{N-1} & 0 \end{bmatrix} ,$$

we may express Eq.(9.3) as

$$x\tilde{\boldsymbol{P}}^{(\alpha)}(x) = \mathrm{J}_N \tilde{\boldsymbol{P}}^{(\alpha)}(x) + \beta_N \tilde{P}_{N+1}^{(\alpha)}(x) \ .$$

However, since the Gauss quadrature points, z_j, are defined as the roots of $P_{N+1}^{(\alpha)}(x)$, and therefore also of $\tilde{P}_{N+1}^{(\alpha)}$ we realize that the real grid points, z_j, appear as the eigenvalues of the symmetric bi-diagonal matrix, J_N. The eigenvalue problem may be solved using the QR-algorithm and the corresponding weights may be computed using the exact formulas. However, we may in fact recover the weights from the eigenvectors of J_N. To realize that we recall the formula for the interpolating Lagrange polynomial associated with the Gauss quadrature points given as

$$\tilde{l}_j(z) = u_j \sum_{n=0}^{N} \frac{P_n^\alpha(z) P_n^\alpha(z_j)}{\gamma_n} = u_j \left(\tilde{\boldsymbol{P}}^{(\alpha)}(z)\right)^T \tilde{\boldsymbol{P}}^{(\alpha)}(z_j) \ .$$

Using the Christoffel-Darboux identity we established that

$$\tilde{l}_j(z_j) = u_j \left(\tilde{\boldsymbol{P}}^{(\alpha)}(z_j)\right)^T \tilde{\boldsymbol{P}}^{(\alpha)}(z_j) = 1 \ .$$

In other words, the normalized eigenvector, $\boldsymbol{Q}(z_j)$, corresponding to the eigenvalue, z_j, of J_N is given as

$$\boldsymbol{Q}(z_j) = \sqrt{u_j}\tilde{\boldsymbol{P}}^{(\alpha)}(z_j) \ ,$$

from which, by equating the first components of the two vectors, we obtain the expression of the weight as

$$u_j = \left(\frac{Q_0(z_j)}{\tilde{P}_0^{(\alpha)}(z_j)} \right)^2 = \gamma_0 \left(Q_0(z_j) \right)^2 = \sqrt{\pi} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+3/2)} \left(Q_0(z_j) \right)^2 \quad ,$$

since $\tilde{P}_0^{(\alpha)}(z_j) = 1/\sqrt{\gamma_0}$. Here $Q_0(z_j)$ signifies the first component of the eigenvector, $\boldsymbol{Q}(z_j)$. Hence, the quadrature points as well as the weights may be obtained directly by computing the $N+1$ eigenvalues and the first component of the corresponding eigenvectors.

The algorithm for computing the Gauss-Radau quadrature points and weights is very similar, the main difference being due to the definition of the Gauss-Radau quadrature points, which are found as the roots of the polynomial

$$q(y) = P_{N+1}^{(\alpha)}(y) + \alpha_N P_N^{(\alpha)}(y) = 0 \quad ,$$

where α_N is chosen such that $q(y)$ vanish at one of the two boundaries as

$$\alpha_N = -\frac{P_{N+1}^{(\alpha)}(\pm 1)}{P_N^{(\alpha)}(\pm 1)} = (\mp 1) \frac{N+1+2\alpha}{N+1} \quad ,$$

where the upper sign corresponds to $g(1) = 0$ while the lower sign yields α_N for $g(-1) = 0$.

The three-term recurrence relation, Eq.(9.3), yields

$$y \tilde{\boldsymbol{P}}^{(\alpha)}(y) = \mathrm{J}_N \tilde{\boldsymbol{P}}^{(\alpha)}(y) + \beta_N \tilde{P}_{N+1}^{(\alpha)}(y) \quad .$$

Utilizing the definition of the Gauss-Radau quadrature points, we express

$$\tilde{P}_{N+1}^{(\alpha)}(\pm 1) = \tilde{\alpha}_N \tilde{P}_N^{(\alpha)}(\pm 1) \quad ,$$

where

$$\tilde{\alpha}_N = -\sqrt{\frac{\gamma_N}{\gamma_{N+1}}} \alpha_N = (\pm 1) \sqrt{\frac{(N+1+2\alpha)(2N+2\alpha+3)}{(N+1)(2N+2\alpha+1)}} \quad .$$

Thus, the last row of J_N may be modified as

$$(\pm 1) \tilde{P}_N^{(\alpha)}(\pm 1) = \beta_{N-1} \tilde{P}_{N-1}^{(\alpha)}(\pm 1) + \beta_N \tilde{P}_{N+1}^{(\alpha)}(\pm 1)$$

$$= \beta_{N-1}\tilde{P}^{(\alpha)}_{N-1}(\pm 1) + \beta_N \tilde{\alpha}_N \tilde{P}^{(\alpha)}_N(\pm 1) \ ,$$

i.e. only the element $(i,j) = (N,N)$ of J_N needs to be modified while the remaining part of the algorithm follows. Hence, the Gauss-Radau quadrature points are found by solving the modified eigenvalue problem while the corresponding weights, v_j, are found from the first components of the corresponding eigenvectors as discussed in connection with the Gauss quadratures.

Let us finally consider the modifications necessary to compute the Gauss-Lobatto quadrature points and weights. In this case, the quadrature points are given as the roots of the polynomial

$$q(x) = P^{(\alpha)}_{N+1}(x) + \alpha_N P^{(\alpha)}_N(x) + \alpha_{N-1} P^{(\alpha)}_{N-1}(x) \ ,$$

where the coefficients, α_{N-1} and α_N, are found such that $q(\pm 1) = 0$, i.e. by solving the system

$$\begin{aligned} \alpha_N P^{(\alpha)}_N(-1) + \alpha_{N-1} P^{(\alpha)}_{N-1}(-1) &= -P^{(\alpha)}_{N+1}(-1) \\ \alpha_N P^{(\alpha)}_N(1) + \alpha_{N-1} P^{(\alpha)}_{N-1}(1) &= -P^{(\alpha)}_{N+1}(1) \ . \end{aligned}$$

If we then normalize the polynomials as usual, these constants are modified as

$$\tilde{\alpha}_N = \sqrt{\frac{\gamma_N}{\gamma_{N+1}}}\alpha_N \ , \quad \tilde{\alpha}_{N-1} = \sqrt{\frac{\gamma_{N-1}}{\gamma_{N+1}}}\alpha_{N-1} \ ,$$

and we recover the equation for the quadrature points as

$$\tilde{P}^{(\alpha)}_{N+1}(x) + \tilde{\alpha}_N \tilde{P}^{(\alpha)}_N(x) + \tilde{\alpha_{N-1}} P^{(\alpha)}_{N-1}(x) = 0 \ .$$

This may be enforced on the eigenvalue problem by changing two elements of J_N as

$$(\mathrm{J}_N)_{N,N-1} = \beta_{N-1} - \beta_N \tilde{\alpha}_{N-1} \ , \quad (\mathrm{J}_N)_{N,N} = -\beta_N \tilde{\alpha}_N \ ,$$

while the remaining part of the algorithm remains unchanged. Unfortunately, using this approach for computing the Gauss-Lobatto quadrature points, J_N looses its symmetry, thereby making the solution of the resulting eigensystem slightly harder.

A different approach can be taken by recalling that

$$\frac{dP_N^{(\alpha)}(x)}{dx} = (2\alpha+1)P_{N-1}^{(\alpha+1)}(x) \quad ,$$

i.e., the interior Gauss-Lobatto quadrature point for $P_N^{(\alpha)}(x)$ can be recovered as the roots of the polynomial $P_{N-1}^{(\alpha+1)}(x)$ Gauss quadrature points of the polynomial $P_{N-2}^{(\alpha+1)}(x)$ which be may recover directly using the symmetric approach discussed above.

9.3 Finite Precision Effects.

In the ideal world of approximation and stability theory, the accuracy of spectral methods depends, as we have seen, only on the regularity of the function being approximated and the operators being involved. However, in the non-ideal world of computation, the finite precision of the computer has a very significant effect on any algorithm being implemented.

For spectral methods the effect of round-off errors is most pronounced when derivatives are computed, with a very significant difference between the behavior of derivatives computed using continuous expansion coefficients and discrete expansion coefficients/interpolating Lagrange polynomials.

In the following we shall address this problem in detail and discover that for polynomial spectral methods in particular, the effects of the finite precision can have very significant consequences for the overall accuracy of the scheme, indeed, for certain problems the results are essentially useless due to overwhelming an amplification of round-off errors.

9.3.1 Finite Precision Effects in Fourier Methods.

Let us begin by considering any given function, $u(x) \in L^2[0, 2\pi]$, being approximated using a continuous Fourier series

$$\mathcal{P}_N u(x) = \sum_{n=-N/2}^{N/2} \hat{u}_n \exp(inx) \quad , \quad \hat{u}_n = \frac{1}{2\pi}\int_0^{2\pi} u(x)\exp(-inx)\,dx \quad .$$

As we have discussed in detail previously, the approximation to the m-derivative of $u(x)$ is then given exactly as

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.438E+00	0.575E+01	0.119E+02	0.255E+03
16	0.477E-01	0.105E+01	0.522E+01	0.106E+03
32	0.360E-03	0.139E-01	0.114E+00	0.434E+01
64	0.104E-07	0.778E-06	0.119E-04	0.873E-03
128	0.222E-14	0.160E-13	0.524E-13	0.335E-11
256	0.311E-14	0.160E-13	0.782E-13	0.398E-12
512	0.355E-14	0.160E-13	0.782E-13	0.568E-12
1024	0.355E-14	0.195E-13	0.853E-13	0.568E-12
Accuracy	$\mathcal{O}(\varepsilon_M(N_0/2))$	$\mathcal{O}(\varepsilon_M(N_0/2)^2)$	$\mathcal{O}(\varepsilon_M(N_0/2)^3)$	$\mathcal{O}(\varepsilon_M(N_0/2)^4)$

$$\mathcal{P}_N u^{(m)}(x) = \sum_{n=-N/2}^{N/2} (in)^m \hat{u}_n \exp(inx) \quad ,$$

which naturally means that the highest modes are being amplified. Let us study the effect of this phenomenon through an example.

Example 47. Let us consider the $C^\infty[0,2\pi]$ and periodic function

$$u(x) = \frac{3}{5-4\cos(x)} \quad ,$$

with the continuous expansion coefficients being

$$\hat{u}_n = 2^{-|n|} \quad ,$$

from which we directly obtain the approximation to the m-derivative as

$$\mathcal{P}_N u^{(m)}(x) = \sum_{n=-N/2}^{N/2} (in)^m 2^{-|n|} \exp(inx) \quad .$$

In Table 9.1 we show the maximum pointwise error of this approximation to $u^{(m)}(x)$ with increasing number of modes, N, being used.

We observe that once the function is well resolved the error approaches machine zero, ε_M, faster than any algebraic order of N. However, a close inspection reveals that the accuracy of the approximation decays with increasing order of the derivative as

$$\max_{x\in[0,2\pi]} \left| u^{(m)}(x) - \mathcal{P}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M (N_0/2)^m) \quad \text{as} \quad N \to \infty \ .$$

Here N_0 corresponds to the number of modes required to approximate the function, $u(x)$, to $\mathcal{O}(\varepsilon_M)$. For $N \gg N_0$ we have $\hat{u}_n \ll \varepsilon_M$ and, since $\hat{u}_n$ decays exponentially in this limit, $\hat{u}_k^{(m)} \ll \varepsilon_M$, i.e. the last term that contributes to the accuracy is $\hat{u}_{N_0/2} \sim \mathcal{O}(\varepsilon_M)$, being the limiting factor on accuracy.

The main observation to make is that the effect of the finite precision is independent of N once the function is well resolved and only a slight dependency of the order of the derivative on the accuracy is observed. Unfortunately, such behavior does not carry over to the discrete case.

Let us now consider the case where the function, $u(x) \in L^2[0,2\pi]$, is approximated using the discrete expansion coefficients as

$$\mathcal{I}_N u(x) = \sum_{n=-N/2}^{N/2} \tilde{u}_n \exp(inx) \ , \quad \tilde{u}_n = \frac{1}{N c_n} \sum_{j=0}^{N-1} u(x_j) \exp(-inx_j) \ ,$$

where we use the even grid

$$x_j = \frac{2\pi}{N} j \ , \quad j \in [0, N-1] \ .$$

The actual computation of the expansion coefficients may be performed using the FFT or by simply summing the series. Once the expansion coefficients are obtained, computation of the approximation to the m-derivative of $u(x)$ is obtained as for the continuous expansion like

$$\mathcal{I}_N u^{(m)}(x) = \sum_{n=-N/2}^{N/2} (in)^m \tilde{u}_n \exp(inx) \ .$$

Let us consider the accuracy of this approach in the following example.

Example 48. Consider again the $C^\infty[0,2\pi]$ and periodic function

$$u(x) = \frac{3}{5 - 4\cos(x)} \ .$$

The expansion coefficients are now found using an FFT, from which we immediately obtain the expansion coefficients for the m'th derivative as

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.654E+00	0.402E+01	0.134E+02	0.233E+03
16	0.814E-01	0.500E+00	0.618E+01	0.770E+02
32	0.648E-03	0.391E-02	0.173E+00	0.210E+01
64	0.198E-07	0.119E-06	0.205E-04	0.247E-03
128	0.380E-13	0.216E-11	0.116E-09	0.715E-08
256	0.657E-13	0.705E-11	0.680E-09	0.954E-07
512	0.272E-12	0.605E-10	0.132E-07	0.110E-05
1024	0.447E-12	0.253E-09	0.844E-07	0.157E-04
Accuracy	$\mathcal{O}(\varepsilon_M(N/2))$	$\mathcal{O}(\varepsilon_M(N/2)^2)$	$\mathcal{O}(\varepsilon_M(N/2)^3)$	$\mathcal{O}(\varepsilon_M(N/2)^4)$

$$\tilde{u}_n^{(m)} = (in)^m \tilde{u}_n \quad .$$

In Table 9.2 we show the maximum pointwise error of this approximation to $u^{(m)}(x)$ with increasing number of modes, N, being used.

We note a pronounced difference in the accuracy of the derivatives as compared to the results quoted in Table 9.1, where the error is constant for a given m once the function is well resolved. Using the FFT we find that the accuracy deteriorates with increasing order of the derivative, as in Table 9.1, but also for increasing N. Indeed, we observe a scaling of the error like

$$\max_{x\in[0,2\pi]} \left| u^{(m)}(x) - \mathcal{I}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M (N/2)^m) \quad \text{as} \quad N \to \infty \ ,$$

i.e. a significant reduction in the accuracy, in particular for high order derivatives. This is a consequence of the uniform error of $\mathcal{O}(\varepsilon_M)$ introduced by the numerical computation of the discrete expansion coefficients. As a result we have that $\tilde{u}_n \sim \mathcal{O}(\varepsilon_M)$ even for $N \gg N_0$, where we expect the expansion coefficients to decay exponentially. However, due to the finite accuracy, it is impossible to obtain these very small numbers. Thus, the maximum error is obtained from the maximum mode number, $N/2$, and the manifestation of this term, introduced by the uniform noise-level, is clearly seen in Table 9.2.

One way to reduce the effect of the round-off error in the calculation of the expansion coefficients, as there is no way of avoiding it, is to always compute the FFT with the highest possible accuracy and in general attempt to formulate the differential equations using as low order

derivatives as possible.

Although we have only illustrated the problem of round-off errors using the FFT and the discrete expansion coefficients, the general picture remains the same in the case where differentiation matrices are used. Indeed, it is usually found that using the FFT and the expansion coefficients results in the numerically most stable algorithm, i.e. the algorithm which suffers least from effects of the finite precision. However, if the entries of the differentiation matrices are carefully computed, i.e. the exact entries are computed for high order derivatives rather than obtained by multiplying several first order differentiation matrices, the two different algorithms yield a comparable accuracy.

9.3.2 Finite Precision in Polynomial Methods

The situation for polynomial methods is even worse than what we saw for the case of the discrete Fourier expansion and, as we shall see, great care has to be exercised when approximating high order derivatives using polynomial methods.

For reasons of simplicity, we shall restrict the discussion to the case of Chebyshev expansions and derivatives. However, most of the results carry directly over to the case of general ultraspherical polynomials unless otherwise stated.

As for the Fourier expansion we begin by considering the continuous Chebyshev expansion of $u(x) \in L^2_w[-1,1]$ as

$$\mathcal{P}_N u(x) = \sum_{n=0}^{N} \hat{u}_n T_n(x) \quad , \quad \hat{u}_n = \frac{2}{c_n \pi} \int_{-1}^{1} u(x) T_n(x) \frac{1}{\sqrt{1-x^2}}\, dx \quad .$$

The approximation to the m-derivative of $u(x)$ is directly obtained as

$$\mathcal{P}_N u^{(m)}(x) = \sum_{n=0}^{N} \hat{u}_n^{(m)} T_n(x) \quad ,$$

where the continuous expansion coefficients for $u^{(m)}(x)$ are found by repeated use of the backward recursion formula

$$\forall n \in [1,N] : \; c_{n-1} \hat{u}_{n-1}^{(m)} = \hat{u}_{n+1}^{(m)} + 2n \hat{u}_n^{(m-1)} \quad ,$$

with the assumption that $\hat{u}_N^{(m)} = \hat{u}_{N+1}^{(m)} = 0$. Let us examine the accuracy of this transform-recurrence-transform method through an exam-

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.273E+02	0.136E+04	0.552E+05	0.237E+07
16	0.232E+01	0.295E+03	0.247E+05	0.168E+07
32	0.651E-02	0.268E+01	0.678E+03	0.126E+06
64	0.164E-07	0.245E-04	0.221E-01	0.143E+02
128	0.568E-13	0.125E-11	0.582E-10	0.931E-09
256	0.568E-13	0.296E-11	0.873E-10	0.349E-08
512	0.568E-13	0.341E-11	0.873E-10	0.442E-08
1024	0.853E-13	0.341E-11	0.873E-10	0.442E-08
Accuracy	$\mathcal{O}(\varepsilon_M N_0)$	$\mathcal{O}(\varepsilon_M N_0^2)$	$\mathcal{O}(\varepsilon_M N_0^3)$	$\mathcal{O}(\varepsilon_M N_0^4)$

ple.

Example 49. Consider the function, $u(x) \in C^\infty[-1,1]$, as

$$u(x) = \frac{1}{x+a} \quad , \quad a > 1 \ ,$$

for which the continuous expansion coefficients are given as

$$\hat{u}_n = \frac{2}{c_n} \frac{1}{\sqrt{a^2-1}} (\sqrt{a^2-1} - a)^n \ .$$

As the function is smooth we find, as expected, that the expansion coefficients decay exponentially fast in n. Note that when a approaches 1 the function develops a strong gradient at $x = -1$ and becomes singular in the limit. In this example we used $a = 1.1$.

Using the backward recursion formula for Chebyshev polynomials we have calculated the expansion coefficients for higher derivatives and in Table 9.3 we list the maximum pointwise error of the expansion as a function of the order, N, of the approximating polynomial.

As for Fourier series we find that once the function, $u(x)$, is well approximated the error is close to machine zero, ε_M. However, a closer look shows that the maximum pointwise error approximately as

$$\max_{x\in[-1,1]} \left| u^{(m)}(x) - \mathcal{P}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M N_0^m) \quad \text{as} \quad N \to \infty \ ,$$

where N_0 corresponds to the maximum mode number required to approximate $u(x)$ to $\mathcal{O}(\varepsilon_M)$. Due to the rapid decay of the expansion

coefficients, we know that for $N \gg N_0$ $\hat{u}_n \ll \varepsilon_M$, i.e. the last term that contributes to the accuracy is $2N_0\hat{u}_{N_0}$ which is $\mathcal{O}(\varepsilon_M)$. Contrary to Fourier series, the expansion coefficient, $\hat{u}_n^{(m)}$, depends on all coefficients with higher n. However, due to the rapid decay of the coefficients the backward recursion is extremely stable, i.e. the last term of order $\mathcal{O}(\varepsilon_M)$ is carried backwards in the recursion without being amplified, thus leading to the observed scaling.

The situation for the discrete expansion is rather different. Let us consider the Chebyshev Gauss-Lobatto expansion as

$$\mathcal{I}_N u(x) = \sum_{n=0}^{N} \tilde{u}_n T_n(x) \quad , \quad \tilde{u}_n = \frac{2}{\overline{c}_n N} \sum_{j=0}^{N} \frac{1}{\overline{c}_j} u(x_j) T_n(x_j) \quad ,$$

where the Gauss-Lobatto quadrature points are given as

$$x_j = -\cos\left(\frac{\pi}{N} j\right) \quad , \quad j \in [0, N] \quad .$$

The discrete approximation to the m-derivative of $u(x)$ is obtained as for the continuous expansion using the backward recursion repeatedly. In the following example we consider the accuracy of this approach.

Example 50. Let us again consider the function, $u(x) \in C^\infty[-1, 1]$, being defined as

$$u(x) = \frac{1}{x + a} \quad , \quad a > 1 \quad ,$$

where we now compute the discrete Chebyshev expansion coefficients using a standard Fast Cosine Transform algorithm and use the analytic backward recursion formulas to calculate the expansion coefficients for the higher derivatives. In Table 9.4 we list the maximum pointwise error as obtained for increasing resolution and order of derivative for $a = 1.1$. The results should be compared with those in Table 9.3.

The effect of using the discrete Chebyshev expansion coefficients as compared to the continuous coefficients is very pronounced and quite discouraging. We note that the error increases rapidly with the number of modes as well as with the order of the derivative and find that the error scales approximately as

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.571E+01	0.403E+02	0.276E+04	0.119E+06
16	0.485E+00	0.936E+01	0.436E+04	0.443E+06
32	0.912E-03	0.771E-01	0.129E+03	0.404E+05
64	0.130E-08	0.514E-06	0.289E-02	0.333E+01
128	0.154E-09	0.474E-06	0.104E-02	0.179E+01
256	0.527E-08	0.636E-04	0.560E+00	0.390E+04
512	0.237E-08	0.374E-03	0.203E+02	0.723E+06
1024	0.227E-07	0.362E-01	0.458E+04	0.457E+09
Accuracy	$\mathcal{O}(\varepsilon_M N^3)$	$\mathcal{O}(\varepsilon_M N^5)$	$\mathcal{O}(\varepsilon_M N^7)$	$\mathcal{O}(\varepsilon_M N^9)$

$$\max_{x\in[-1,1]} \left| u^{(m)}(x) - \mathcal{I}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M N^{2m+1}) \quad \text{as} \quad N \to \infty \ .$$

As we observe in Table 9.4, this strong dependence on N implies that for high values of N and/or m it becomes impossible to approximate the derivative of the function to any reasonable error. The problem lies in the combination of the cosine transform and the backward recursion used to determine the expansion coefficients for the higher derivatives. The backward recursion leads to an $\mathcal{O}(N^2)$ amplification of the initial round-off error of $\mathcal{O}(\varepsilon_M N)$ resulting from the transform. This last term could be avoided by using a direct summation, which, however, may be prohibitively expensive for large N. The ill-conditioning of the backward recursion has the sad consequence that the approximation of high order derivatives remains a non-trivial task when using polynomial methods and great care has to be taken when attempting to do so. Comparing the results illustrated in Ex. 49 and Ex. 50 it becomes clear that the most important issue here is the accuracy by we compute the discrete expansion coefficients, i.e. this should always be done with the highest possible accuracy.

Although using the expansion coefficients and backward recursion is mathematically equivalent to using the differentiation matrices, the two methods are numerically very different as we shall observe shortly. Let us again consider the discrete Chebyshev approximation using the interpolating Lagrange polynomials as

$$\mathcal{I}_N u(x) = \sum_{j=0}^{N} u(x_j) l_j(x) = \sum_{j=0}^{N} u(x_j) \frac{(-1)^{j+1}(1-x^2)T'_N(x)}{\overline{c}_j N^2 (x-x_j)} \ ,$$

where we have chosen to consider the approximation based on the Chebyshev Gauss-Lobatto quadrature points

$$x_j = -\cos\left(\frac{\pi}{N}j\right) \quad , \quad j \in [0, N] \ .$$

Differentiation is then accomplished through a matrix vector product as

$$\left.\frac{d\mathcal{I}_N u}{dx}\right|_{x_j} = \sum_{i=0}^{N} \mathrm{D}_{ji} u(x_i) \ ,$$

where the entries of the differentiation matrix are given as

$$\mathrm{D}_{ij} = \begin{cases} -\frac{2N^2+1}{6} & i = j = 0 \\ \frac{\bar{c}_i}{\bar{c}_j}\frac{(-1)^{i+j}}{x_i - x_j} & i \neq j \\ -\frac{x_i}{2(1-x_i^2)} & i = j \in [1, N-1] \\ \frac{2N^2+1}{6} & i = j = N \end{cases} \ . \tag{9.4}$$

Let us consider the accuracy of this approach in the following example.

Example 51. Consider the function, $u(x) \in C^\infty[-1, 1]$, being defined as

$$u(x) = \frac{1}{x + a} \quad , \quad a > 1 \ ,$$

where we compute derivatives using the differentiation matrix, D, implemented exactly as in Eq.(9.4) while higher derivatives are computed by repeatedly multiplying with the differentiation matrix as

$$\boldsymbol{u}^{(1)} = \mathrm{D}\boldsymbol{u} \ , \ \boldsymbol{u}^{(m)} = \mathrm{D}^m \boldsymbol{u} \ .$$

In Table 9.5 we list the maximum pointwise error as obtained for increasing resolution and order of derivative for $a = 1.1$.

This is clearly very disappointing. The direct implementation of the differentiation matrices indicates an accuracy like

$$\max_{x\in[-1,1]} \left| u^{(m)}(x) - \mathcal{I}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M N^{2m+2}) \quad \text{as} \quad N \to \infty \ .$$

Fortunately, there are several things that can be done to improve on this result. Let us first attempt to understand what causes the strong

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.201E+02	0.127E+04	0.545E+05	0.237E+07
16	0.116E+01	0.221E+03	0.218E+05	0.160E+07
32	0.191E-02	0.134E+01	0.441E+03	0.956E+05
64	0.276E-08	0.741E-05	0.917E-02	0.731E+01
128	0.633E-08	0.458E-04	0.161E+00	0.386E+03
256	0.139E-06	0.406E-02	0.589E+02	0.578E+06
512	0.178E-05	0.178E+00	0.983E+04	0.379E+09
1024	0.202E-04	0.837E+01	0.200E+07	0.325E+12
Accuracy	$\mathcal{O}(\varepsilon_M N^4)$	$\mathcal{O}(\varepsilon_M N^6)$	$\mathcal{O}(\varepsilon_M N^8)$	$\mathcal{O}(\varepsilon_M N^{10})$

influence of the finite precision.

All off-diagonal entries of the matrix, D, are given like

$$\mathrm{D}_{ij} \sim C\frac{1}{x_i - x_j} \quad .$$

Close to the boundary, $x = \pm 1$, this term scales like

$$\begin{aligned}\mathrm{D}_{ij} &\sim \frac{1}{x_i - x_j} \sim \frac{1}{1-\cos(\pi/N)} \sim \frac{1}{\mathcal{O}(N^{-2}) + \varepsilon_M} \\ &\sim \frac{\mathcal{O}(N^2)}{1+\mathcal{O}(\varepsilon_M N^2)} \sim \mathcal{O}(N^2)(1-\mathcal{O}(\varepsilon_M N^2)) \\ &\sim \mathcal{O}(N^2) + \mathcal{O}(\varepsilon_M N^4) \quad ,\end{aligned}$$

i.e. the differentiation matrix has a condition number as $\mathcal{O}(\varepsilon_M N^4)$ reflecting the observed scaling in Tabel 9.5. The question to address is what can be done about this. The subtraction of almost equal numbers can be avoided by using trigonometric identities such that the entries of D are initialized like

$$\mathrm{D}_{ij} = \begin{cases} -\frac{2N^2+1}{6} & i=j=0 \\ \frac{\overline{c}_i}{2\overline{c}_j}\frac{(-1)^{i+j}}{\sin\left(\frac{i+j}{2N}\pi\right)\sin\left(\frac{i-j}{2N}\pi\right)} & i \neq j \\ -\frac{x_i}{2\sin^2\left(\frac{\pi}{N}i\right)} & i=j\in[1,N-1] \\ \frac{2N^2+1}{6} & i=j=N \end{cases} \quad .$$

Doing a direct implementation of this matrix reduces the error making the accuracy scale like

$$\max_{x\in[-1,1]} \left| u^{(m)}(x) - \mathcal{I}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M N^{2m+1}) \quad \text{as} \quad N \to \infty \ ,$$

similar to that using the Fast Cosine Transform and the backward recursion. However, it is in fact possible to make the matrix-vector multiply method perform even better. The solution lies hidden in the computation of the elements for $i \sim j \sim N$. In that case we need to compute function like $\sin(\pi - \delta)$, where $\delta \ll \pi$, i.e. this operation is sensitive to round-off error effects. Indeed, by doing an estimate of the condition number like above we obtain

$$\begin{aligned} \mathrm{D}_{ij} &\sim \frac{1}{\sin(\delta)\sin(\pi-\delta)} \sim \frac{1}{\mathcal{O}(N^{-1})(\mathcal{O}(N^{-1})+\varepsilon_M)} \\ &\sim \frac{\mathcal{O}(N^2)}{1+\mathcal{O}(\varepsilon_M N)} \sim \mathcal{O}(N^2) + \mathcal{O}(\varepsilon_M N^3) \ , \end{aligned}$$

i.e. we arrive at a condition number as $\mathcal{O}(\varepsilon_M N^3)$ as expected. This analysis also suggests a way to avoid this problem since it happens only for $i \sim j \sim N$. The remedy is to use the centro-antisymmetry of D such that the entries of the differentiation matrix should be initialized as

$$\mathrm{D}_{ij} = \begin{cases} -\frac{2N^2+1}{6} & i = j = 0 \\ \frac{\overline{c}_i}{2\overline{c}_j} \frac{(-1)^{i+j}}{\sin\left(\frac{i+j}{2N}\pi\right)\sin\left(\frac{i-j}{2N}\pi\right)} & i \in [0, N/2] \neq j \in [0, N] \\ -\frac{x_i}{2\sin^2\left(\frac{\pi}{N}i\right)} & i = j \in [1, N/2] \\ -\mathrm{D}_{N-i,N-j} & i \in [N/2+1, N], j \in [0, N] \end{cases} \ , \qquad (9.5)$$

i.e. only the upper half of the matrix is computed while the lower half is obtained by using the centro-antisymmetry. Let us illustrate the accuracy of the differentiation at this point.

Example 52. Consider the function, $u(x) \in C^\infty[-1,1]$, being defined as

$$u(x) = \frac{1}{x+a} \quad , \quad a > 1 \ ,$$

where we compute derivatives using the differentiation matrix, D, implemented using the trigonometric identities and the centro-antisymmetry, Eq.(9.5). Higher derivatives are computed by repeatedly multiplying

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.201E+02	0.127E+04	0.545E+05	0.237E+07
16	0.116E+01	0.221E+03	0.218E+05	0.160E+07
32	0.191E-02	0.134E+01	0.441E+03	0.956E+05
64	0.262E-08	0.721E-05	0.901E-02	0.721E+01
128	0.203E-10	0.467E-07	0.437E-04	0.196E+00
256	0.554E-10	0.113E-05	0.128E-01	0.109E+03
512	0.354E-09	0.201E-04	0.871E+00	0.304E+05
1024	0.182E-08	0.744E-03	0.153E+03	0.214E+08
Accuracy	$\mathcal{O}(\varepsilon_M N^2)$	$\mathcal{O}(\varepsilon_M N^4)$	$\mathcal{O}(\varepsilon_M N^6)$	$\mathcal{O}(\varepsilon_M N^8)$

with the differentiation matrix as

$$\boldsymbol{u}^{(1)} = \mathrm{D}\boldsymbol{u} \quad , \quad \boldsymbol{u}^{(m)} = \mathrm{D}^m \boldsymbol{u} \quad .$$

In Table 9.6 we list the maximum pointwise error as obtained for increasing resolution and order of derivative for $a = 1.1$.

From Table 9.6 we recover an estimate of the accuracy like

$$\max_{x\in[-1,1]} \left| u^{(m)}(x) - \mathcal{I}_N u^{(m)}(x) \right| \sim \mathcal{O}(\varepsilon_M N^{2m}) \quad \text{as} \quad N \to \infty \ ,$$

which is even better than what is obtained using a standard Cosine transform and the backward recursion. It also illustrates well the care that has to be exercised when initializing the entries of differentiation matrices for polynomial methods.

For methods other than the Chebyshev methods the situation is slightly more bleak. Certainly, the use of trigonometric identities can only be used for the Chebyshev case. The centro-antisymmetry of D, on the other hand, is shared among all the differentiation matrices. However, the effect of using this for more general polynomials remains unknown. Nevertheless, using the even-odd splitting for computing derivatives results in an error that scales somewhat like that seen in Ex. 52 also for Legendre differentiation matrices provided care is exercised in computing the entries of D. This suggests that the use of the centro-antisymmetry, which is implicit in the even-odd splitting, does indeed result in a smaller condition number of the differentiation matrices. We also emphasize that, whenever available, the exact entries of $\mathrm{D}^{(m)}$ should be used rather than computed by multiplying matrices.

For the general polynomial, but not for Chebyshev polynomials as

the above techniques are far superior, one could also use the assumption that the differentiation of a constant has to vanish, i.e.

$$\sum_{j=0}^{N} \mathrm{D}_{ij} = 0 \quad .$$

One may then compute the diagonal elements of the differentiation matrix as

$$\forall i \in [0, N] : \ \mathrm{D}_{ii} = -\sum_{\substack{j=0 \\ j \neq i}}^{N} \mathrm{D}_{ij} \quad ,$$

such that the round-off errors incurred in the computation of the entries are somehow accounted for in the diagonal elements. This technique, however, should only be used when nothing more specific is available.

9.4 Convolution Sums and Dealiasing

The computation of nonlinear terms is a trivial task when using collocation methods as everything is performed in point-space. However, when using a Galerkin or a tau method the situation is quite different. In this section we shall briefly discuss methods for efficiently computing convolution sums appearing from Galerkin or tau approximations of second order nonlinearities and address the issue of aliasing introduced by such techniques.

As we have seen previously, the use of Galerkin methods is essentially restricted to Fourier methods, where, however, they play an important role for e.g. studies of homogeneous, isotropic hydrodynamic turbulence. Let us therefore consider the quadratic nonlinearity

$$w(x) = u(x)v(x) \quad ,$$

where $u(x) \in L^2[0, 2\pi]$ and $v(x) \in L^2[0, 2\pi]$, such that the usual continuous Fourier expansions exist as

$$u(x) = \sum_{n=-\infty}^{\infty} \hat{u}_n \exp(inx) \quad , \quad v(x) = \sum_{n=-\infty}^{\infty} \hat{v}_n \exp(inx) \quad ,$$

where $\hat{u}_n$ and $\hat{v}_n$ represent the continuous Fourier expansion coefficients. If we also introduce the continuous Fourier expansion coefficients

$$\hat{w}_n = \frac{1}{2\pi}\int_{-1}^{1} w(x)\exp(-inx)\,dx \quad ,$$

the three sets of expansion coefficients are related through the convolution

$$\hat{w}_n = \sum_{\substack{k,l=-\infty \\ k+l=n}}^{\infty} \hat{u}_k \hat{v}_l \quad .$$

In the case of $u(x)$, $v(x)$ and $w(x)$ being approximated by finite sums only, the convolution sum becomes

$$n \in [-N/2, N/2-1]: \;\; \hat{w}_n = \sum_{\substack{k,l=-N/2 \\ k+l=n}}^{N/2-1} \hat{u}_k \hat{v}_l \quad .$$

The computational effort, however, involved in the computation of the convolution is $\mathcal{O}(N^2)$, which becomes prohibitive even for moderate N. This should be contrasted with the collocation methods where the computation of the nonlinear term is only $\mathcal{O}(N)$, however, the computation of the derivative at the grid points is of $\mathcal{O}(N\log_2 N)$. Thus, to be competitive, an $\mathcal{O}(N\log_2 N)$ method for computing the convolution sum is required.

Such an approach is known as the transform method and involves the use of the Fast Fourier Transform and therefore also a grid. Hence, even though the Galerkin as such is grid free, we need to introduce a grid for computational efficiency and with the grid comes the aliasing error.

The transform method relies on the speed of the FFT by transforming $\hat{u}_n$ and $\hat{v}_n$ into physical space, i.e. point values of $u(x)$ and $v(x)$, multiplying the two functions in point space and transform the product back to spectral space using the FFT. Thus, this approach requires three FFT and one point space multiplication, in total yielding an $\mathcal{O}(N\log_2 N)$ algorithm. However, this approach also introduces an additional source of error. To see this, assume that we use the even Fourier grid as

$$x_j = \frac{2\pi}{N} j \quad , \quad j \in [0, N-1] \quad ,$$

such that grid point value of $u(x)$ and $v(x)$ are given as

$$u(x_j) = \sum_{n=-N/2}^{N/2-1} \hat{u}_n \exp(inx_j) \quad , \quad v(x_j) = \sum_{n=-N/2}^{N/2-1} \hat{v}_n \exp(inx_j) \quad .$$

Computing the expansion coefficients for the product, $w(x)$, using the FFT yields

$$\begin{aligned}
\hat{w}_n &= \frac{1}{N} \sum_{j=0}^{N-1} u(x_j) v(x_j) \exp(-inx_j) \\
&= \frac{1}{N} \sum_{j=0}^{N-1} \left(\sum_{k=-N/2}^{N/2-1} \hat{u}_k \exp(ikx_j) \right) \left(\sum_{l=-N/2}^{N/2-1} \hat{v}_l \exp(ilx_j) \right) \exp(-inx_j) \\
&= \sum_{k=-N/2}^{N/2-1} \sum_{l=-N/2}^{N/2-1} \hat{u}_k \hat{v}_l \left(\frac{1}{N} \sum_{j=0}^{N-1} \exp(i(k+l-n)x_j) \right) \\
&= \sum_{m=-\infty}^{\infty} \sum_{\substack{k,l=-N/2 \\ k+l=n+mN}}^{N/2-1} \hat{u}_k \hat{v}_l = \hat{w}_n + \sum_{\substack{k,l=-N/2 \\ k+l=n\pm N}}^{N/2-1} \hat{u}_k \hat{v}_l \quad ,
\end{aligned}$$

where the last reduction results from the orthogonality of the discrete exponential function and $m = \{-1, 0, 1\}$ only since $k, l \leq |N/2|$. Hence, not only do we obtain the required expansion coefficients for $\hat{w}_k$, but we get an extra contribution due to the introduction of the grid. This extra contribution is known as the *dynamic aliasing error* and implies that the scheme no longer is a pure Galerkin scheme.

The removal of this extra term has been a source of considerable discussion in the past, i.e. is it necessary to remove this error. At this point in time it is safe to say it is in general not necessary to remove this aliasing if the functions involved are well resolved. However, in special marginally resolved cases and for very sensitive problems where extreme care has to be exercised it may be necessary to remove this error and we shall briefly discuss two ways of doing so.

The transform method is clearly only of interest for methods based on series expansions for which there exists a fast transform, which in essence limits the attention to the trigonometric polynomials and Chebyshev polynomials. Indeed, using the fast Cosine transform, convolution sums for Chebyshev expansions may also be computed efficiently through the transform method. In this case the transformed variable involves the

terms

$$\hat{w}_n = \frac{1}{2} \sum_{m=-\infty}^{\infty} \left(\sum_{\substack{k,l=0 \\ k+l=n+2mN}}^{N} \hat{u}_k \hat{v}_l + \sum_{\substack{k,l=0 \\ |k-l|=n+2mN}}^{N} \hat{u}_k \hat{v}_l \right) \ .$$

The aliasing error only becomes relevant for $m = \{-1, 0, 1\}$ and the methods of removing the aliasing errors developed for the Fourier methods carries directly over to the Chebyshev case.

9.4.1 Dealiasing by Truncation.

Due to its simplicity, dealiasing by truncation is the most widely used. The technique has it origin in using the FFT at M points rather than N point where $M > N$. Let us introduce the M-grid as

$$\tilde{x}_j = \frac{2\pi}{M} j \ \ , \ \ j \in [0, M-1] \ ,$$

such that the discrete expansions of $u(x)$ and $v(x)$ are given as

$$u(\tilde{x}_j) = \sum_{n=-M/2}^{M/2-1} \hat{u}_n \exp(in\tilde{x}_j) \ \ , \ \ v(\tilde{x}_j) = \sum_{n=-M/2}^{M/2-1} \hat{v}_n \exp(in\tilde{x}_j) \ .$$

However, the expansion coefficients, $\hat{u}_n$ and $\hat{v}_n$, are defined as

$$\hat{u}_n = \begin{cases} \hat{u}_n & -N/2 \leq n \leq N/2-1 \\ 0 & \text{otherwise} \end{cases} \ \ , \ \ \hat{v}_n = \begin{cases} \hat{v}_n & -N/2 \leq n \leq N/2-1 \\ 0 & \text{otherwise} \end{cases} \ ,$$

i.e. only the first $\pm N/2$ expansion coefficients are considered while the remaining are set to zero. We then consider the expansion coefficients for $w(x)$ as

$$\begin{aligned} \hat{w}_n &= \frac{1}{M} \sum_{j=0}^{M-1} u(\tilde{x}_j) v(\tilde{x}_j) \exp(-in\tilde{x}_j) \\ &= \sum_{m=-\infty}^{\infty} \sum_{\substack{k,l=-M/2 \\ k+l=n+mM}}^{M/2-1} \hat{u}_k \hat{v}_l = \hat{w}_n + \sum_{\substack{k,l=-N/2 \\ k+l=n\pm M}}^{N/2-1} \hat{u}_k \hat{v}_l \ , \end{aligned}$$

by explicitly using that $\hat{u}_n$ and $\hat{v}_n$ are padded with zeros. We are only interested in $\hat{w}_n$ for $|n| \leq N/2$ and we wish to choose M such that the

aliasing term vanishes for these coefficients. This is ensured for all n by requiring it to be the case for the worst case

$$-\frac{N}{2} - \frac{N}{2} \geq \frac{N}{2} - 1 - M \quad ,$$

yielding the condition on M as

$$M \geq \frac{3}{2}N - 1 \quad ,$$

explaining why this method also is known as the 3/2-rule.

A consequence of using truncation for dealiasing is that a significant amount of extra work has to be done. Indeed, we see that M has to be 50% larger than N to ensure no aliasing and make the method equivalent to a Galerkin method for the first $N/2$ modes.

The technique is most easily implemented by taking M to be a number that allows for using the FFT. Prior to transforming $\hat{u}_n$ and $\hat{v}_n$ into real space, the highest 1/3 of the modes are then forced to zero. Transforming to real space, performing the multiplication and transforming $w(x)$ back to spectral space produces the M expansion coefficients for $w(x)$ and again the highest 1/3 of these modes are forced to zero before continuing. Hence, only 2/3 of the modes are active in the approximation, making the application of the method of truncation rather expensive, in particular in more than one dimension.

We would like to comment that the dealiasing using truncation is nothing more than applying a filter as discussed previously, albeit with a step-function as the filter function. Thus, dealiasing in a collocation method can be implemented using the filter matrix with the special filter function as

$$\sigma\left(\frac{n}{M/2}\right) = \begin{cases} 1 & |n| \leq N/2 \\ 0 & \text{otherwise} \end{cases} \quad .$$

Thus filtering $u(x)$ and $v(x)$ prior to the multiplication and also the final product $w(x)$ ensures that no aliasing error remains and the collocation method is equivalent to the Galerkin method.

9.4.2 Dealiasing by Phase Shift.

A second method of dealiasing employs the properties of Fourier transforms associated with phase shifts. Rather than using a larger grid as in

the method of truncation this method is based on computing the interpolation of $u(x)$ and $v(x)$ at two grids, shifted by half a grid cell. Thus, in addition to computing the usual polynomials at x_j we also compute the polynomials at the grid points

$$\tilde{x}_j = \frac{(2j+1)\pi}{N} \quad j \in [0, N-1] \ ,$$

as

$$\overline{u}(\tilde{x}_j) = \sum_{n=-N/2}^{N/2-1} \hat{u}_n \exp(in\tilde{x}_j) = \exp\left(i\frac{\pi}{N}\right) \sum_{n=-N/2}^{N/2-1} \hat{u}_n \exp(inx_j) \ ,$$

and likewise for $\overline{v}(\tilde{x}_j)$. Note that both these transformations may still be computed using the FFT since the phase shift only results in a multiplicative constant. If we now consider the expansion coefficients for $\overline{w}$ we obtain

$$\begin{aligned}\overline{\hat{w}}_n &= \frac{1}{N} \sum_{j=0}^{N-1} \overline{u}(\tilde{x}_j)\overline{v}(\tilde{x}_j) \exp(-in\tilde{x}_j) \\ &= \sum_{m=-\infty}^{\infty} \exp\left(imN\frac{\pi}{N}\right) \sum_{\substack{k,l=-N/2 \\ k+l=n+mN}}^{N/2-1} \hat{u}_k \hat{v}_l = \hat{w}_n - \sum_{\substack{k,l=-N/2 \\ k+l=n\pm N}}^{N/2-1} \hat{u}_k \hat{v}_l \ .\end{aligned}$$

Hence, the dealiased $\hat{w}_n$ are obtained directed by adding the aliased $\hat{w}_n$ and the shifted $\overline{\hat{w}}_n$ as

$$\hat{w}_n = \frac{1}{2}\left(\hat{w}_n + \overline{\hat{w}}_n\right) \ .$$

Using the method of phase shifting thus requires an additional FFT as also $\overline{\hat{w}}_n$ needs to be computed besides the aliased $\hat{w}_n$. However, the transforms are only of length N rather than $3/2N$ as for the truncation method. Nevertheless, in one dimension the method of truncation is cheaper than phase shifting, while in more than one dimension it becomes harder to decisively choose among the two different approaches.

9.5 On the Use of Mappings

Change of variables remains a very important tool in all branches of physics and engineering and plays an equally important role in the ap-

plication of spectral methods, in particular those based on expansions in orthogonal polynomials.

The variety of useful mappings is large and we will only cover those most often used, with the emphasis on one-dimensional methods. In most cases, mappings are used to allow for treating problems in geometries different from the standard interval. However, in the last section we will discuss the use of mappings for improving on the accuracy of high-order derivatives.

Let us consider the function, $u(x) \in L^2[a,b]$, where $a < b$ while a and/or b may both be infinite. If we make a change of variables through the mapping function, $\psi(\xi)$, as

$$\psi(\xi) : \mathsf{I} \rightarrow [a,b] \;\text{ as }\; x = \psi(\xi) \quad ,$$

where $\mathsf{I} = [\xi_{\min}, \xi_{\max}]$ represents the interval $[0, 2\pi]$ when dealing with Fourier methods while it becomes $[-1,1]$ in the case of polynomial expansions, we have the differential

$$dx = \psi'(\xi) d\xi \quad .$$

Hence the magnitude of ψ' supplies a measure of how the nodes are distorted relative to the standard grid given in I. For ψ' being less than one the grid is compressed whereas it is dilated when ψ' is larger than one. This provides a rough guide to which mapping may be appropriate for a particular problem.

When using mapping in connection with spectral methods one must compute derivatives in x, while the methods of computing the derivatives hitherto were computed using the standard interval, I, only, i.e. with respect to ξ. However, using the chain rule for differentiation we directly obtain

$$\frac{d}{dx} = \frac{1}{\psi'}\frac{d}{d\xi} \quad , \quad \frac{d^2}{dx^2} = \frac{1}{(\psi')^3}\left(\psi' \frac{d^2}{d\xi^2} - \psi'' \frac{d}{d\xi}\right) \quad ,$$

and in a similar fashion for higher derivatives.

The simplest example of a suitable mapping function is for a situation where one needs to map the finite interval, $[a,b]$, onto I, in which case we have the well known result

$$x = \psi(\xi) = a + \frac{\xi_{\max} + \xi}{\xi_{\max} - \xi_{\min}}(b-a) \quad , \quad \psi'(\xi) = \frac{b-a}{2} \quad .$$

As expected we find that all parts of the interval is mapped with the same factor, $\psi'(\xi)$.

Before discussing more general mapping functions let us briefly touch on the subject of implementation. Let us for simplicity restrict the attention to the case of polynomial expansions but note that everything carries straight over to Fourier methods.

In the case of Galerkin or tau methods, we consider approximations of the form

$$\mathcal{P}_N u(\psi(\xi)) = \sum_{n=0}^{N} \hat{u}_n P_n^{(\alpha)}(\xi) \quad ,$$

where

$$\hat{u}_n = \frac{1}{\gamma_n} \int_{-1}^{1} u(\psi(\xi)) P_n^{(\alpha)}(\xi) w(\xi)\, d\xi \quad .$$

Within this formulation, we wish now to obtain an approximation of the derivative of the mapped function, $u(x)$, as

$$\mathcal{P}_N u^{(1)}(\psi(\xi)) = \frac{1}{\psi'(\xi)} \sum_{n=0}^{N} \hat{u}_n^{(1)} P_n^{(\alpha)}(\xi) \quad ,$$

where $\hat{u}_n^{(1)}$ signifies the expansion coefficients for the first derivative obtained through the backward recursion. Now, in the case of Galerkin and tau methods the unknowns are the expansion coefficients, $\hat{u}_n$, for which we need to obtain equations. Hence, we also need to expand the mapping function, $\psi'(\xi)$, as

$$\mathcal{P}_N \frac{1}{\psi'(\xi)} = \sum_{n=0}^{N} \hat{\psi}_n P_n^{(\alpha)}(\xi) \quad ,$$

such that

$$\mathcal{P}_N u^{(1)}(\psi(\xi)) = \sum_{n,l=0}^{N} \hat{\psi}_l \hat{u}_n P_l^{(\alpha)}(\xi) P_n^{(\alpha)}(\xi) \quad ,$$

being a convolution sum. Although the expression for the convolution of the polynomials in general is available it is very complicated with the exception of a few special cases, e.g. for Chebyshev polynomials. However, the expression for the expansion coefficients in general becomes

very complicated and general mappings are, for this reason, rarely used in Galerkin and tau methods except in cases where the mapping is particularly simple, e.g. the linear mapping where the mapping derivative becomes a constant and the convolution reduces to a multiplication. Another example where the mapping is reasonably simple is the case where the mapping function, $1/\psi'(\xi)$, is a non-singular rational function in ξ, in which case the convolution operator becomes a banded and very sparse matrix operator as a consequence of the three-term recurrence relations valid for orthogonal polynomials.

Let us now turn towards the use of mapping in Collocation methods, where it becomes much simpler. In this case we seek approximations of the form

$$\mathcal{I}_N u(\psi(\xi)) = \sum_{n=0}^{N} \tilde{u}_n P_n^{(\alpha)}(\xi) = \sum_{j=0}^{N} u(\psi(\xi_j)) l_j(\xi) \quad ,$$

where ξ_j signifies the standard grid in I, the discrete expansion coefficients, $\tilde{u}_n$, are found using a quadrature formula and $l_j(\xi)$ represents the interpolating Lagrange polynomial associated with the grid points. In this case we wish to compute derivatives of $u(x)$ at the grid points as

$$\mathcal{I}_N u^{(1)}(\psi(\xi_i)) = \frac{1}{\psi'(\xi_i)} \sum_{n=0}^{N} \tilde{u}_n^{(1)} P_n^{(\alpha)}(\xi_i) = \frac{1}{\psi'(\xi_i)} \sum_{j=0}^{N} u(\psi(\xi_j)) \mathrm{D}_{ij} \quad ,$$

where D represents the differentiation matrix associated with $l_j(\xi)$ and the grid points, ξ_i. However, since everything is done in point-space the mapping just corresponds to multiplying with a constant at each grid point. Hence, introducing the diagonal matrix

$$\mathrm{M}_{ii}^{(m)} = \psi^{(m)}(\xi_i) \quad ,$$

the mapping is accomplished by multiplying the inverse of this matrix onto the solution vector following the computation of the derivative at the grid points. Indeed, when using the m-order differentiation matrix, $\mathrm{D}^{(m)}$, for the computation of the derivative we simply obtain

$$\boldsymbol{u}^{(1)} = \frac{1}{\mathrm{M}^{(1)}} \mathrm{D} \boldsymbol{u} \quad , \quad \boldsymbol{u}^{(2)} = \frac{1}{(\mathrm{M}^{(1)})^3} \left(\mathrm{M}^{(1)} \mathrm{D}^{(2)} - \mathrm{M}^{(2)} \mathrm{D}^{(1)} \right) \boldsymbol{u} \quad ,$$

where $\boldsymbol{u}$ represents the solution vector at the grid points and similarly for $\boldsymbol{u}^{(1)}$ and $\boldsymbol{u}^{(2)}$ for the first and second derivative at the nodal points,

respectively. Clearly, mapped higher derivatives may be obtained in similar manner illustrating the particular ease by which mappings can be introduced in this formulation. One should also that since $\mathrm{M}^{(m)}$ is diagonal very little computational overhead is introduced.

9.5.1 Local Refinement Using Fourier Methods.

We shall first consider the use of mappings in connection with Fourier collocation methods. However, prior to discussing useful mapping functions, let us briefly concern ourselves with general properties that have to be shared among the mapping functions such that the mapped functions retain the spectral accuracy.

If we consider the general function, $u(x) \in L^2[a,b]$ and the mapping, $\psi(\xi) : \mathrm{I} \to [a,b]$, the continuous expansion coefficients become

$$\begin{aligned} 2\pi\hat{u}_n &= \int_0^{2\pi} u(\psi(\xi)) \exp(-in\xi) d\xi \\ &= \frac{-1}{in} \left[u(\psi(2\pi)) - u(\psi(0)) \right] + \frac{1}{in} \int_0^{2\pi} \psi'(\xi) u'(\psi(\xi)) \exp(-in\xi) d\xi \quad . \end{aligned}$$

Hence, to maintain spectral accuracy we have to put requirements on $\psi(\xi)$ similar to those on $u(x)$, i.e. $\psi(\xi)$ and its derivatives has to be periodic and smooth to allow for spectral accuracy.

This said, let us now consider two mappings useful in connection with Fourier methods. Many problems have solutions which are localized in space, however, remain periodic. For such problems one may apply a standard Fourier method, although it seems natural to cluster the grid points around the steep gradients of the solution. This can be done by mapping the equidistant grid to increase the local resolution.

One choice of a mapping function having this effect on a periodic function defined on the interval $\xi \in [-\pi, \pi]$, is the Arctan-mapping given as

$$x = \psi(\xi) = 2 \arctan(L \tan \frac{\xi - \xi_0}{2}) \quad , \quad \psi'(\xi) = \frac{L(1 + \tan^2 \frac{\xi - \xi_0}{2})}{1 + L^2 \tan^2 \frac{\xi - \xi_0}{2}} \quad ,$$

where $L \leq 1$ is a control parameter for the amount of clustering that is required around ξ_0. Clearly, for $L = 1$ the mapping reduced to a unity mapping. The best way to understand the effect of this mapping is to recall that

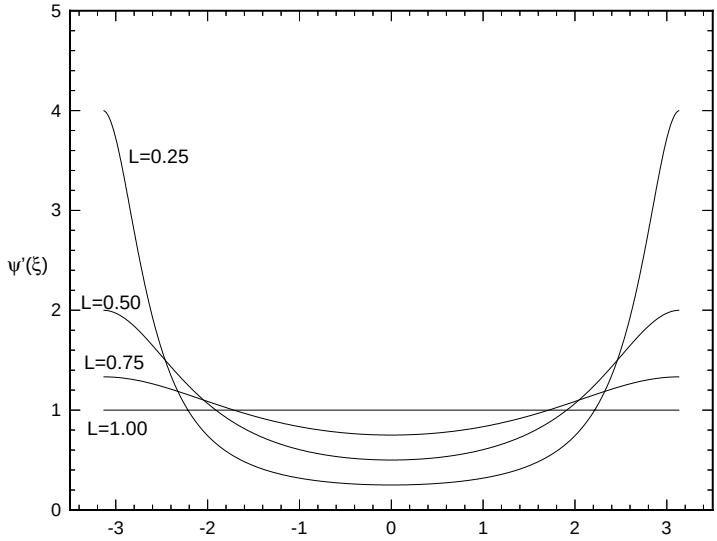

figure 9.1. Illustration of clustering of grid points around $\xi_0 = 0$ using the Arctan-mapping for different values of the mapping parameter, L.

$$dx = \psi'(\xi) d\xi \quad ,$$

i.e. since $d\xi$ is a constant on the equidistant grid we obtain clustering where $\psi'(\xi) < 1$ and stretching of the grid where $\psi'(\xi) > 1$. This is illustrated in Fig. 9.1 where we plot the value of $\psi'(\xi)$ for different values of the mapping parameter, L.

We observe a clear clustering of the grid points around $\xi_0 = 0$ and find that the size L controls the amount of clustering with increasing clustering around ξ_0 for decreasing L.

Since $\psi'(\xi)$ consists of trigonometric functions periodicity of $u(x)$ is preserved through the mapping. This holds for arbitrary order of differentiation. Moreover, the mapping function is smooth and introduces no singularities in the domain. Consequently, we may expect that the mapping preserves spectral accuracy of the approximation of a smooth function.

An alternative to this mapping, which, however, has similar properties, is given as

$$x = \psi(\xi) = \arctan\left(\frac{(1-\beta^2)\sin(\xi - \xi_0)}{(1+\beta^2)\cos(\xi - \xi_0) + 2\beta}\right) \quad ,$$

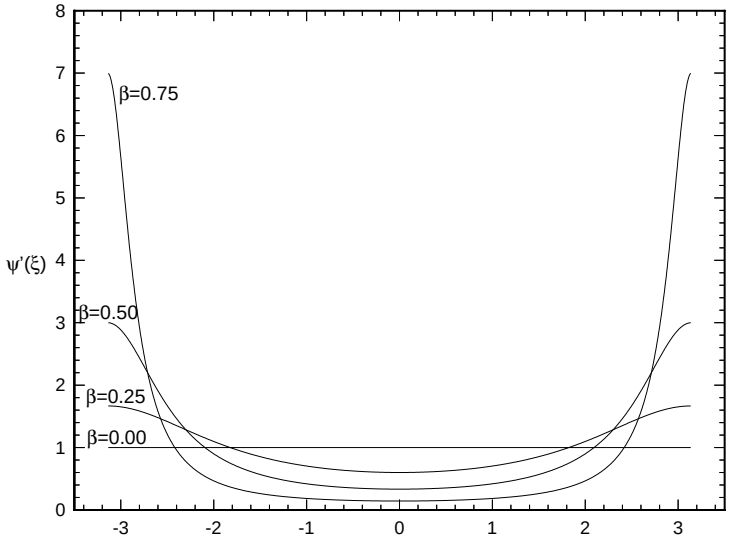

figure 9.2. Illustration of clustering of grid points around $\xi = 0$ different values of the mapping parameter, β.

$$\psi'(\xi) = \frac{1-\beta^2}{1+\beta^2+2\beta\cos(\xi-\xi_0)} \ ,$$

where $|\beta| < 1$ controls the clustering around $\xi_0 \in [-\pi, \pi]$. For reasons of comparison we plot in Fig. 9.2 the mapping derivative for several values of β. Note that the mapping becomes singular for $\beta = 1$, while $\beta = 0$ corresponds to a unity mapping. This mapping also preserves periodicity and spectral accuracy of the approximation. Comparing the two schemes as illustrated in Fig. 9.1 and Fig. 9.2 we observe that the latter mapping leads to a less localized clustering around ξ_0 which in a many cases is a desirable property.

9.5.2 Mapping Functions for Polynomial Methods.

Considering mapping functions for polynomial methods we shall concern ourselves with problems utilizing the ultraspherical polynomials as the Laguerre and Hermite polynomials are restricted to problems on an infinite interval and only the linear mapping is of interest in these cases.

Let us therefore consider a function, $u(x) \in L^2_w[a,b]$, expanded in ultraspherical polynomials as

$$\mathcal{I}_N u(\psi(\xi)) = \sum_{n=0}^{N} \tilde{u}_N P_n^{(\alpha)}(\xi) = \sum_{j=0}^{N} u(\psi(\xi_j)) l_j(\xi) \quad .$$

Performing the usual integration by parts procedure to study the rate of decay of the expansion coefficients, it is easily established that $\psi(\xi)$ must be a smooth function to maintain spectral convergence and $\psi(\pm 1)$ must be bounded to avoid effects from the boundary.

In the following we shall discuss mappings that allowing for the use of ultraspherical polynomials, and in particular Chebyshev polynomials, for the approximation of problems in the semi-infinite and infinite interval. However, we shall also discuss a different use of mappings that results in an increased accuracy by which derivatives may be computed.

9.5.2.1 Treatment of Semi-Infinite Intervals.

The straightforward way of approximating problems in the semi-infinite interval is to use expansions of Laguerre polynomials. However, the lack of fast Laguerre transforms and the poor convergence properties of these polynomials suggests that alternative methods should be sought after.

The existence of the fast transform methods for Chebyshev methods makes the use of this polynomial family very appealing. However, attention has to be paid to the approximation of the mapped function since we have to impose a singular mapping in order to map infinity into a finite value. An appropriate guideline is that uniform spectral convergence may be maintained if the function, $u(x)$ and the mapping function $x = \psi(\xi)$, both are sufficiently smooth and the function, $u(x)$, decays fast enough without severe oscillations towards infinity.

A widely used mapping is the exponential mapping, $\psi(\xi) : \mathsf{I} \rightarrow [0, \infty[$, as

$$x = \psi(\xi) = -L \ln\left(\frac{1-\xi}{2}\right) \quad , \quad \psi'(\xi) = \frac{L}{1-\xi} \quad ,$$

where L is a scale length of the mapping. We note that the grid is distorted with a linear rate towards infinity and that no singular behavior at $x = 0$ is introduced. This mapping has been given significant attention due to its rather complicated behavior. It has been shown that only for functions that decay at least exponentially towards infinity may one expect to maintain the spectral convergence. This result, however, is based on asymptotic arguments and good results have been

reported by other researchers. The reason is the logarithmic behavior which results in a strong stretching of the grid. This may behave well in many cases, while in other situations it may result in a slowly convergent approximation.

An alternative to the exponential map is the algebraic mapping given as

$$x = \psi(\xi) = L\frac{1+\xi}{1-\xi} \quad , \quad \psi'(\xi) = \frac{2L}{(1-\xi)^2} \quad ,$$

where L again plays the role of a scale length. This mapping has been studied in great detail and, used in Chebyshev approximations, the mapped basis functions has been dubbed *rational Chebyshev polynomials* defined as

$$TL_n(x) = T_n\left(\frac{x-L}{x+L}\right) \quad .$$

This family is defined for $x \in [0, \infty[$ and orthogonality as well as completeness may be established. Consequently, we may expect spectral accuracy for approximation of smooth functions. On the other hand, this is not a great surprise as the rational Chebyshev polynomials are simply mapped Chebyshev polynomials. The advantage of using the algebraic mapping is that the function, $u(x)$, only needs to decay algebraicly towards infinity or asymptote towards a constant value in order for the approximation to maintain spectral accuracy. This is contrary to the exponential mapping which requires at least exponential decay. Several authors have found that algebraic mappings are more accurate and robust than exponential mappings, which is the reason for their wide use.

One should observe that when applying a singular mapping it is not always convenient to chose the Gauss-Lobatto points as collocation points as they include the singular point. The proper choice may be the Gauss-Radau points for the polynomial family.

As an alternative to using a singular mapping, one may truncate the domain and apply a mapping. At first it may seem natural to just apply a linear mapping after the truncation. However, this has the effect of wasting a significant amount of resolution towards infinity where only little is needed. If this is not the case, truncation becomes obsolete.

The idea behind domain truncation is that if the function decays exponentially fast towards infinity, then we will only make an exponentially

small error by truncating the interval. This approach yields spectral convergence of the approximation for increasing resolution.

An often used mapping is the logarithmic mapping function, $\psi(\xi) : \mathrm{I} \rightarrow [0, L_{\max}]$, defined as

$$x = \psi(\xi) = L_{\max} \frac{\exp(a(1-\xi)) - \exp(2a)}{1 - \exp(2a)} \quad , \quad \psi'(\xi) = -a\psi(\xi) \ ,$$

where a is a tuning parameter.

One thing should be noted though. There is a complication about using domain truncation in that for increasing resolution we need to increase the domain size in order to avoid that the error introduced by truncating the domain will dominate over the error of the approximation.

9.5.2.2 Treatment of Infinite Intervals.

When approximating functions defined on the infinite interval it may seem natural to employ expansions based on Hermite polynomials. However, no fast Hermite transforms are known and the convergence rate of Hermite expansions is rather slow suggesting that alternatives could be useful.

As on semi-infinite intervals, we wish to develop singular mappings which may be used to map the infinite interval into the standard interval such that ultraspherical polynomials can be applied for approximating the function. Similar to the guidelines used for choosing the mapping function on the semi-infinite interval we can expect that spectral convergence is conserved under the mapping provided the function, $u(x)$, is exponentially decaying and non-oscillatory when approaching infinity. Clearly, the mapping function needs to be singular at both endpoints to allow for mapping of the infinite interval onto the finite standard interval.

As for the semi-infinite case, we may also construct an exponential mapping function, $\psi(\xi) : \mathrm{I} \rightarrow]-\infty, \infty[$ as

$$x = \psi(\xi) = L \tanh^{-1} \xi \quad , \quad \psi'(\xi) = \frac{L}{1 - \xi^2} \ ,$$

where L plays the role of a scale length. This mapping requires exponential decay of the function towards infinity to yield spectral accuracy.

Alternatively, one may use an algebraic mapping as

$$x = \psi(\xi) = L \frac{\xi}{\sqrt{1 - \xi^2}} \quad \psi'(\xi) = \frac{L}{\sqrt{(1 - \xi^2)^3}} \ ,$$

where L again plays the role of a scale length. This mapping has been given significant attention and, used in Chebyshev approximations, a special symbol has been introduced for the *rational Chebyshev polynomials* as

$$TB_n(x) = T_n\left(\frac{x}{\sqrt{L^2+x^2}}\right) \quad ,$$

for which one may prove orthogonality as well as completeness. The advantage of applying this mapping is that spectral accuracy of the approximation may be obtained even when the function decays only algebraicly or asymptotes towards a constant value at infinity.

We note that the proper choice of collocation points on the infinite interval may, in certain cases, not be the usual Gauss-Lobatto points but rather the Gauss quadrature points.

9.5.2.3 Mappings for Accuracy Improvement.

As a final example of the use of mappings we shall now turn to a slightly different, however important, problem of pseudospectral methods. In many problems in physics the partial differential equation includes derivatives of high order, e.g., 3rd order derivatives in the Korteweg-de Vries equation and 4th order derivatives in the Kuramoto-Shivashinsky equation. Additionally, such equations often introduces very complex behavior, thus requiring a large number of modes in the polynomial expansion. For problems of this type the effect of roundoff error becomes a significant problem as the polynomial differential operators are ill conditioned as discussed in detail in Sec. 9.3.2. Even for moderate values of m and N can this problem be disastrous and ruin the numerical scheme.

To overcome this problem, at least partially, one may apply the following mapping, $\psi(\xi) : \mathsf{I} \to \mathsf{I}$, as

$$x = \psi(\xi) = \frac{\arcsin(\alpha\xi)}{\arcsin\alpha} \quad , \quad \psi'(\xi) = \frac{\alpha}{\arcsin\alpha}\frac{1}{\sqrt{1-(\alpha\xi)^2}} \quad ,$$

where α controls the mapping. It may be shown that the error, ε, introduced by applying the mapping, which is singular for $\xi = \pm\alpha^{-1}$, is related to α as

$$\alpha = \cosh^{-1}\left(\frac{|\ln\varepsilon|}{N}\right) \quad ,$$

i.e. by choosing $\varepsilon \sim \varepsilon_M$ the error introduced by the mapping is guaran-

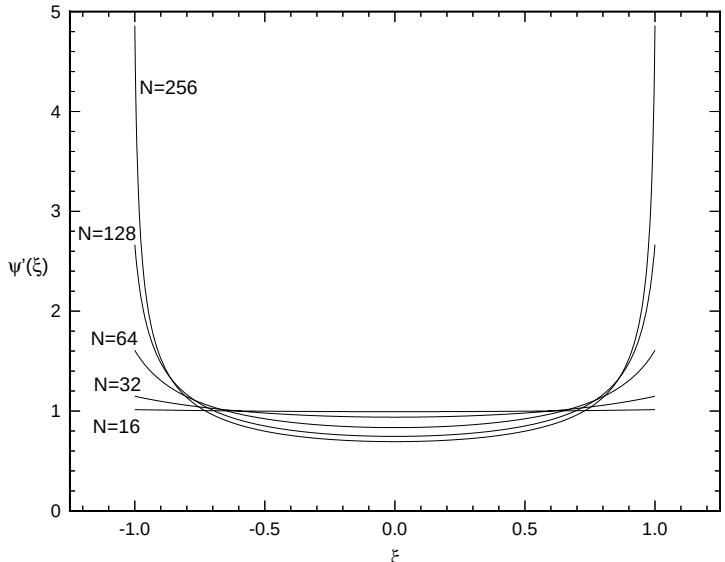

figure 9.3. Illustration of the effect of the mapping used for accuracy improvement when evaluating spatial derivatives at increasing resolution.

N	$m=1$	$m=2$	$m=3$	$m=4$
8	0.155E-03	0.665E-02	0.126E+00	0.142E+01
16	0.316E-12	0.553E-10	0.428E-08	0.207E-06
32	0.563E-13	0.171E-10	0.331E-08	0.484E-06
64	0.574E-13	0.159E-09	0.174E-06	0.111E-03
128	0.512E-12	0.331E-08	0.124E-04	0.321E-01
256	0.758E-12	0.708E-08	0.303E-03	0.432E+01
512	0.186E-10	0.233E-05	0.143E+00	0.587E+04
1024	0.913E-10	0.361E-04	0.756E+01	0.109E+07

teed to be harmless.

The effect of the mapping is to stretch the grid close to the boundary points. This is easily realized by considering the two limiting values of α;

$$\begin{array}{ll} \alpha \to 0 & \Delta_{\min} x \to 1 - \cos\frac{\pi}{N} \\ \alpha \to 1 & \Delta_{\min} x \to \quad \frac{2}{N} \end{array} ,$$

where $\Delta_{\min} x$ represents the minimum grid spacing. We observe that for α approaching one the grid is mapped to an equidistant grid. In the opposite limit, the grid is equivalent to the well known Chebyshev Gauss-Lobatto grid. One should note that the limit of one is approached when increasing N, i.e. is it advantageous to evaluate high order deriva-

N	α	$m=1$	$m=2$	$m=3$	$m=4$
8	0.0202	0.154E-03	0.659E-02	0.124E+00	0.141E+01
16	0.1989	0.290E-13	0.383E-11	0.236E-09	0.953E-08
32	0.5760	0.211E-13	0.847E-11	0.231E-08	0.360E-06
64	0.8550	0.180E-12	0.225E-09	0.118E-06	0.436E-04
128	0.9601	0.138E-12	0.227E-09	0.334E-06	0.282E-03
256	0.9898	0.549E-12	0.201E-08	0.262E-05	0.521E-02
512	0.9974	0.949E-11	0.857E-07	0.467E-03	0.180E+01
1024	0.9994	0.198E-10	0.379E-06	0.433E-02	0.344E+02

tives with high resolution at an almost equidistant grid. In Fig. 9.3 we plot the mapping derivative for different resolution with the optimal value of α. This clearly illustrates that the mapping gets stronger for increasing resolution.

Example 53.

Let us consider the function

$$u(x) = \sin(2x) \quad , \quad x \in [-1, 1] \ .$$

We wish to evaluate the first four derivatives of this functions using a standard Chebyshev collocation method with the entries given in Eq.(9.5). In Table 9.7 we list the maximum pointwise error that is obtained for increasing resolution.

We clearly observe the effect of the round-off error and it is obvious that only very moderate resolution can be used in connection with the evaluation of high order derivatives.

As previously, we apply the singular mapping in the hope that the accuracy of the derivatives improve. In Table 9.8 we list the maximum pointwise error for derivatives with increasing resolution. For information we also list the optimal value for α as found for a machine accuracy of $\varepsilon_M \simeq 1.0E-16$.

The effect of applying the mapping is to gain at least an order of magnitude in accuracy and significantly more for high derivatives and large N.

10

Discrete Stability and Time Integration

Until now we have concerned ourselves with the spectral approximation of the solution and the operator in the general initial value problem

$$\begin{aligned}
\frac{\partial u(x,t)}{\partial t} &= \mathcal{L}u(x,t) \ , \quad x \in [-1,1] \ , \ t \geq 0 \ , \\
\mathcal{B}_- u(-1,t) &= g(t) \ , \qquad t \geq 0 \\
\mathcal{B}_+ u(1,t) &= h(t) \ , \qquad t \geq 0 \\
u(x,0) &= f(x) \ , \qquad x \in [-1,1] \ , \ t = 0 \ .
\end{aligned}$$

where $\mathcal{B}_\pm$ represents the boundary operator at $x = \pm 1$, $g(t)$ and $h(t)$ the possibly time dependent boundary conditions and $f(x)$ signifies the initial condition.

Assuming that the boundary operator, $\mathcal{B}_\pm$, is included in the spatial operator, $\mathcal{L}$, we consider in this chapter the properties of the semi-discrete approximation

$$\frac{du_N}{dt} = \mathcal{L}_N(u_N(x,t),x,t) \ ,$$

with appropriate initial conditions, and shall discuss a number of problems central to the solution of the semi-discrete set of coupled ordinary differential equations.

We shall confine the theoretical discussion to the case of linear operators yielding the semi-discrete, linearized and localized problem

$$\frac{du_N}{dt} = \mathcal{L}_N u_N(x,t) \ .$$

We recall that in the semi-discrete situation, $u_N(x,t)$, represents a vector of length $N+1$, containing the expansion coefficients when a Galerkin or a tau method is used and the grid point values in case a collocation method is applied, while $\mathcal{L}_N$ is a matrix of order $N+1$. Thus, in the following we write the semi-discrete problem as

$$\frac{d\boldsymbol{u}}{dt} = \mathrm{L}\boldsymbol{u}(x,t) \quad ,$$

also known as the method of lines. This equation may naturally be solved exactly as

$$\boldsymbol{u}(x,t) = \exp\left[\mathrm{L}t\right]\boldsymbol{u}(x,0) = \exp\left[\mathrm{L}t\right]\boldsymbol{f}(x) \quad ,$$

through the introduction of the matrix exponential. However, it is, with the exception of very simple operators, L, impracticable to apply the matrix exponential for solving the time-dependent problem. Thus, one often introduces an approximation to the matrix exponential with the most frequent schemes being based on an explicit or implicit finite difference scheme to approximate the solution over some time step Δt. Essentially, the exponential, $\exp(z)$, is approximated either as a finite Taylor series

$$\exp(\mathrm{L}\Delta t) \simeq \mathrm{K}(\mathrm{L},\Delta t) = \sum_{i=0}^{m} \frac{(\mathrm{L}\Delta t)^i}{i!} \quad ,$$

or through a Pade approximation

$$\exp(\mathrm{L}\Delta t) \simeq \mathrm{K}(\mathrm{L},\Delta t) = \frac{\sum_{i=0}^{m} a_i(\mathrm{L}\Delta t)^i}{\sum_{i=0}^{n} b_i(\mathrm{L}\Delta t)^i} \quad ,$$

with the expansion coefficients, a_i and b_i, being found such that the approximation agrees with the Taylor expansion. Certainly, for Δt being sufficiently small such an approximation is expected to be valid.

At first it may appear strange that, while using a spectrally accurate spatial approximation in space, we choose to use a finite difference approximation in time with the plausible result that the global error is going to be dominated by this latter term and, hence, the total scheme becomes only accurate to some algebraic order of Δt. However, there are several reasons for using such an approach. The simplicity of the finite difference approximations to the matrix exponential is certainly appealing. Moreover, using explicit methods the maximum time step is

typically restricted in size and this bound often depends in a nonlinear way of the spatial discretization thereby causing the temporal error to become much smaller that the spatial error thereby recovering the spectral accuracy. Additionally, since we are often solving problems with several spatial variables, any reduction in memory requirements is most pronounced in space as we only have one temporal variable. We shall note that when applying a fully implicit scheme, which in some cases have no restriction on the maximum allowable time step, significant care has to be exercised to ensure that the time differencing error does not dominate over the spatial approximation error. We should also note that methods, being spectrally accurate in time, have been developed but they are generally available only for the solution of simple linear problems or require the solution of large, non-symmetric sparse linear problems.

Let us now assume that the matrix exponential is approximated in some appropriate way such that advancing from t to $t + \Delta t$ amount to

$$\boldsymbol{u}(x, t + \Delta t) = \boldsymbol{u}^{n+1} = \mathrm{K}(\mathrm{L}, \Delta t)\boldsymbol{u}^n \quad ,$$

where $t = n\Delta t$, with Δt being the time step, and the matrix $\mathrm{K}(\mathrm{L}, \Delta t)$ represents the approximation to the matrix exponential. Applying $\mathrm{K}(\mathrm{L}, \Delta t)$ repeatedly yields the solution at $t = n\Delta t$ as

$$\boldsymbol{u}^{n+1} = [\mathrm{K}(\mathrm{L}, \Delta t)]^n \, \boldsymbol{f} \quad .$$

Following the discussion on the Lax-Richtmeyer Equivalence Theorem we say that the fully discrete scheme is strongly stable provided

$$\| \, [\mathrm{K}(\mathrm{L}, \Delta t)]^n \, \|_{L^2_w} \leq K(\Delta t) \quad ,$$

where the matrix norm is defined in the usual manner. As sufficient, but not necessary, condition for strong stability is that

$$\|\mathrm{K}(\mathrm{L}, \Delta t)\|_{L^2_w} \leq 1 + \kappa \Delta t \quad ,$$

for some bounded κ and all sufficiently small values of Δt.

In the special case where $\mathrm{K} = \mathrm{K}(\mathrm{L}, \Delta t)$ is a normal matrix, i.e. $\mathrm{K}^T\mathrm{K} = \mathrm{K}\mathrm{K}^T$, strong stability is ensured in L^2 through the von Neumann stability condition

$$\max |\lambda_K| \leq 1 + \kappa \Delta t \quad ,$$

where λ_K represents the eigenvalues of $\mathrm{K}(\mathrm{L}, \Delta t)$. In case $\mathrm{K}(\mathrm{L}, \Delta t)$ is non-normal, which is the case for most spectral approximations, von Neumann stability is still a necessary condition for strong stability but no longer sufficient.

Stability within this framework is also known as Lax stability, since it represents the fully discrete version of the Equivalence theorem ensuring convergence provided stability and consistency is given. However, although playing an important role in the analysis of numerical methods, is it impractical for long-time integration since it allows for exponentially growing solutions and only ensures that the spatial and temporal approximation can be refined sufficiently to ensure convergence at any give time.

A more practical notion of stability is known as asymptotic stability which amounts to requiring that for sufficiently small Δt we may obtain that

$$\|\mathrm{K}(\mathrm{L}, \Delta t)\|_{L^2_w} \leq 1 \quad ,$$

i.e. $\kappa = 0$.

10.1 Eigenvalue Spectra of Fundamental Operators

Although in general only supplying necessary conditions, it is clear from the above discussion that the eigenvalues of $\mathrm{K}(\mathrm{L}, \Delta t)$ play an important role in the understanding of the stability of the fully discrete approximation to the initial boundary value problem.

Let us therefore consider the similarity transform of $\mathrm{K}(\mathrm{L}, \Delta t)$ as

$$\mathrm{K}(\mathrm{L}, \Delta t) = \mathrm{S}^{-1} \Lambda_K \mathrm{S} \quad ,$$

where Λ_K represents the diagonal eigenvalue matrix while S and S^{-1} are the matrices of the left and right eigenvectors of $\mathrm{K}(\mathrm{L}, \Delta t)$, respectively. Provided $\mathrm{K}(\mathrm{L}, \Delta t)$ is normal S and S^{-1} are bounded independent of Δt and L, establishing the sufficiency of the von Neumann stability in this case. In the more general case where $\mathrm{K}(\mathrm{L}, \Delta t)$ is non-normal we can no longer guarantee boundedness of S and S^{-1} and conditions beyond the von Neumann stability has to be considered.

However, if we restrict the attention to the von Neumann criteria it is clear that

$$\Lambda_K = \mathrm{K}(\Lambda_L, \Delta t) \quad ,$$

using the Cayley-Hamilton theorem for the matrix polynomial, K, where Λ_L represents the eigenvalues of the discrete approximation of the spatial operator, $\mathcal{L}$. Hence, the eigenvalues of the discrete approximation of various operators plays an important role as they appear directly in Λ_K and in the following we shall discuss the eigenvalue spectrum of various discrete operators obtained using Fourier or polynomial methods.

10.1.1 Fourier Methods.

Obtaining the eigenvalue spectrum for Fourier approximations of linear operators is simple. If we consider the Galerkin approximations, we immediately get

$$\mathrm{L}^{(1)} = \mathrm{diag}(-iN/2, \ldots, -i, 0, i, \ldots, N/2) \ ,$$

and

$$\mathrm{L}^{(2)} = \mathrm{diag}(-N^2/4, \ldots, -1, 0, -1, \ldots, -N^2/4) \ ,$$

for the first and second order operator, respectively, directly supplying the eigenvalue spectrum. Clearly, the spectrum for higher order derivatives are obtained in an equivalent manner. We note in particular that the maximum eigenvalue for the m-order differentiation is given as

$$\max |\lambda^{(m)}| = \left(\frac{N}{2}\right)^m \ .$$

Since $\mathrm{L}^{(m)}$ is normal the von Neumann stability criteria is both necessary and sufficient to ensure stability.

The situation for the Fourier collocation method is the almost the same, although slightly more complicated. Introducing the differentiation matrix, $\mathrm{D}^{(m)}$, or alternatively using the discrete expansion coefficients, $\tilde{u}_n$, for the computation of L we find that the eigenfunctions for $\mathrm{D}^{(m)}$ are given as $\psi_n(x) = \exp(inx)$ since

$$\forall n \in [-N/2+1, N/2-1] : \mathrm{D}^{(m)}\psi_n(x) = (in)^m \psi_n(x) \ .$$

However, this leaves us with the final two eigenfunctions for $n = \pm N/2$ for which we obtain for, e.g. the first order operator D, that

$$\mathrm{D}\psi_{\pm N/2}(x) = \mathcal{I}_N \frac{d}{dx} \mathcal{I}_N \left[\cos\left(\frac{N}{2}x\right) \pm i \sin\left(\frac{N}{2}x\right)\right] = 0 \ .$$

Hence, these two most extreme modes picks up two extra zero eigenvalues making the spectrum different from that of the Galerkin method due to the use of the slightly different space in which we seek solutions. However, in practice the Galerkin and the collocation methods perform in exactly the same way and it is safe to assume that the eigenvalue spectra for the different methods are identical in any practical situation.

We note that in case the odd number of grid points, y_j, is used that eigenvalue spectrum becomes identical to that of the Galerkin method.

10.1.2 Polynomial Methods.

Contrary to what we found for the Fourier methods, it is no longer possible to obtain simple and analytic expressions for the eigenvalue spectrum when using polynomial methods for the approximation of the spatial operators. In the following we shall discuss the eigenvalue spectra appearing from the approximation of the first and second order spatial derivatives using Legendre and Chebyshev polynomials, with the emphasis on approximate operators appearing from tau and collocation projections.

10.1.2.1 Spectrum of the Advective Operator.

Let us first consider the behavior of the eigenvalue spectrum of the advective operator

$$\mathcal{L}u = \frac{du}{dx} \ ,$$

subject to homogeneous Dirichlet boundary conditions.

Let us for simplicity first consider the Chebyshev collocation approximation

$$\mathrm{L}\boldsymbol{u} = \mathrm{D}\boldsymbol{u} \ ,$$

where D represents the Chebyshev-Gauss-Lobatto differentiation matrix with the boundary conditions being imposed by removing the last row and column of D. As D has no significant symmetry we can not obtain the eigenvalue spectrum by analytical means, leaving us with no alternative to the numerical computation of the spectrum. In Fig. 10.1 we illustrate the corresponding spectrum for various orders of discretization.

We observe that the real parts are strictly negative while the maximum eigenvalue scales as

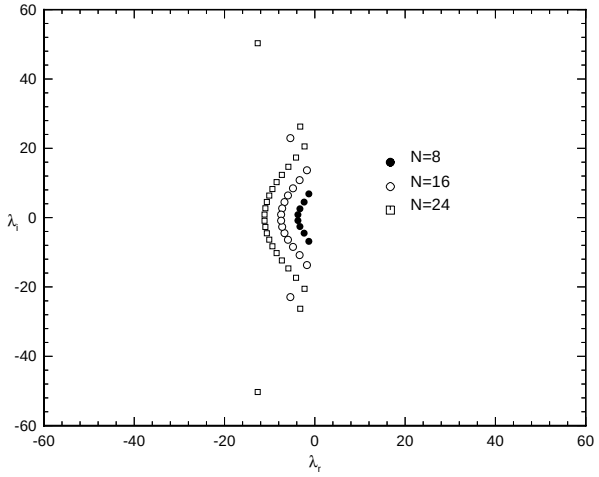

figure 10.1. Eigenvalue spectra of a Chebyshev-Gauss-Lobatto advection operator for increasing resolution, N.

$$\max|\lambda^{(1)}| = \mathcal{O}(N^2) \ ,$$

with the estimate being asymptotically sharp. We note that, contrary to the case for Fourier methods, the eigenvalue grows like N^2 rather that linear in N.

The spectrum for the Chebyshev tau approximation of the advection operator is qualitatively the same as for the collocation approximation although numerical studies show the numerical value of the maximum eigenvalue is slightly smaller than obtained in the collocation approximation.

The situation when using Legendre polynomials to construct the approximation is slightly different. In Fig. 10.2 we show the spectrum of the Legendre-Gauss-Lobatto differentiation matrix and observe that although the spectrum is qualitatively different from the Chebyshev-Gauss-Lobatto case in Fig. 10.1 we recover a similar scaling of the maximum eigenvalue as

$$\max|\lambda^{(1)}| = \mathcal{O}(N^2) \ ,$$

with all real parts of the spectrum being strictly negative, albeit with a very small real part for the extreme eigenvalues.

The situation for the Legendre tau approximation is slightly different, at least in theory. Indeed, it is possibly to show analytically that

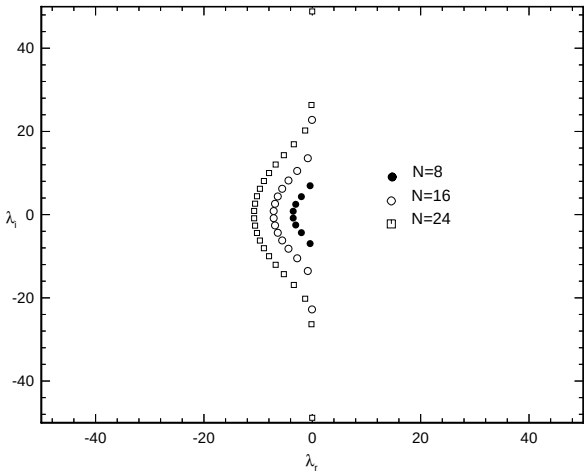

figure 10.2. Eigenvalue spectra of a Legendre-Gauss-Lobatto advection operator for increasing resolution, N.

asymptotically the maximum eigenvalue scales likes

$$\max |\lambda^{(1)}| = \mathcal{O}(N) \ ,$$

in the Legendre tau approximation of the advective operator. On the other hand, numerical studies suggest a quadratic dependence of N. This discrepancy between theory and computation is caused by ill-conditioning of the tau operator, which makes it extremely sensitive to round-off errors, the effect of this being that in effect the maximum eigenvalue scales like N^2 in an actual implementation of the Legendre tau approximation.

One could think that the quadratic dependence of maximum eigenvalue on N is related to the minimal grid size in the Gauss-Lobatto grid, which indeed is of $\mathcal{O}(N^{-2})$ for ultraspherical polynomials. However, although tempting to make such a connection, care has to exercised as the grid scaling is a consequence, rather that a source, of the application of the ultraspherical polynomials with the associated eigenvalue, $\lambda_n \sim n^2$, from the Sturm-Liouville problem. Nevertheless, the minimum grid spacing in practical implementation of polynomial collocation methods supplies a very good estimate of the inverse of the maximum eigenvalue.

The ill-conditioning of the discrete approximations to the advection operators, as we also experienced in the previous chapter, may cause the eigenvalues to behave differently from what we have discussed so far. This supplies yet another reason for exercising great care when

Approximation	$\max \lvert\lambda^{(2)}\rvert/N^4$	$\max \lvert\lambda^{(2)}\rvert/N^4$
Chebyshev Tau	0.300	0.047
Chebyshev Collocation	0.047	0.014
Legendre Tau	0.110	0.026
Legendre Collocation	0.026	0.006

implementing the differentiation matrices in order to minimize the effects of the finite precision.

10.1.2.2 Spectrum of the Diffusive Operator.

Let us now consider the eigenvalue spectrum of the diffusive operator

$$\mathcal{L}u = \frac{d^2u}{dx^2} \quad .$$

We first consider the case of homogeneous Dirichlet boundary conditions enforced by setting the entries of the first and last row and column of the differentiation matrices to zero in the collocation method while the boundary conditions are implemented in the usual manner in the tau method by having the final two rows enforcing the boundary conditions.

We first note that there is only little quantitative difference between the Legendre and Chebyshev methods as well as the tau and collocation approach. Indeed, the eigenvalue spectrum is in all cases strictly negative, real and distinct and bounded by the two constants, c_1 and c_2, as

$$-c_1 N^4 \leq \lambda \leq c_2 < 0 \quad ,$$

i.e. the maximum eigenvalue scales as

$$\max |\lambda^{(2)}| = \mathcal{O}(N^4) \quad ,$$

asymptotically. This result can be confirmed through numerical experiments, and in Table 10.1 we give the asymptotic values of c_1 for reference. We note that the maximum eigenvalue is always smaller using collocation methods as compared to the use of tau methods independent of the choice of polynomials.

The situation for Neumann boundary conditions is very similar to that of Dirichlet conditions. Neumann conditions are implemented in the collocation methods by exchanging the first and last row of the second

order differentiation matrix with those of the first order differentiation matrix.

As for the Dirichlet boundary conditions we obtain a scaling as

$$-c_1 N^4 \leq \lambda \leq c_2 < 0 \ ,$$

with the exception of the zero eigenvalue introduced by the Neumann boundary condition. Thus, the maximum eigenvalue scales as

$$\max |\lambda^{(2)}| = \mathcal{O}(N^4) \ ,$$

asymptotically and in Table 10.1 we give the asymptotic values of c_1. Enforcing the Neumann conditions in a different manner, e.g. implicitly, yields similar values of the maximum eigenvalue and the growth remains to scale with N^4.

10.2 Standard Time Integration Schemes

The actual choice of the time integration method is influenced by several factors such as required accuracy, available memory and computer speed. In the following we shall briefly discuss the most commonly used time integration methods for integrating time dependent partial differential equations with the spatial operators being approximated using spectral methods and discuss the implications of the strong N dependence of the eigenvalue spectrum on the maximum allowable time step.

We consider the initial boundary value problem approximated using spectral methods as

$$\frac{d\boldsymbol{u}^n}{dt} = \mathrm{L}(\boldsymbol{u}^n, n\Delta t) = \mathrm{L}^n \ ,$$

where $\boldsymbol{u}^n$ represents the solution vector at $t = n\Delta t$ with Δt being the actual time step and L^n is the spectral approximation to the operator at $t = n\Delta t$. We assume that the boundary operator is included in L^n and that proper initial conditions are supplied.

10.2.1 Multi-step Schemes.

The general multi-step scheme is of the form

$$\sum_{i=0}^{p} \alpha_i \boldsymbol{u}^{n-i} = \Delta t \sum_{i=0}^{p} \beta_i \mathrm{L}^{n-i} \quad ,$$

where p refers to the order of the scheme. We may distinguish between implicit and explicit methods by realizing that for $\beta_0 = 0$ we obtain the solution at $t = n\Delta t$ from knowledge of the solution at previous time steps only, i.e. the scheme is explicit. Multi-step schemes in general require that solutions at one or more previous time-steps are retained, thus making such schemes memory intensive, in particular when addressing multi-dimensional problems. However, only one evaluation of L^n is required to advance one time step, thereby reducing the computational workload. The importance of memory over computational speed is problem dependent and it is very hard to give general guidelines.

Initially, one may not have the solution at the required number of time steps backward in time. Thus, it is necessary to start out with a few 1st order steps while retaining the results and, once sufficient steps is known backward in time, a high order time differencing scheme can be applied. A suitable choice for performing the initial steps is the forward Euler method but other may also be applied. Since one is only doing very few steps with this initial method, the question of stability may in fact be neglected and even unconditionally unstable schemes can be used for initializing multi-step schemes.

Let us as a first example consider the classic 2nd order explicit Leap-Frog Scheme being defined as

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^{n-1} + 2\Delta t \mathrm{L}^n \quad .$$

Since the stability of the scheme is related to the eigenvalue spectrum of L^n, we shall consider the stability of the linearized problem

$$\mathrm{L}^n = \lambda \boldsymbol{u}^n \quad , \tag{10.1}$$

to obtain necessary conditions for stability of the Leap-Frog scheme. The analysis of a multi-step scheme is done by writing the scheme in matrix form as

$$\begin{pmatrix} \boldsymbol{u}^{n+1} \\ \boldsymbol{u}^{n} \end{pmatrix} = \begin{bmatrix} 2\lambda\Delta t & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{u}^{n} \\ \boldsymbol{u}^{n-1} \end{pmatrix} \quad ,$$

where the matrix operator plays the role of $\mathrm{K}(\mathrm{L}, \Delta t)$ introduced previ-

ously. In order for the solution to remain bounded we must require that the eigenvalues, μ, of the matrix operator advancing the solution by one time step are less than or equal to one. The eigenvalues are easily found as

$$\mu = \lambda\Delta t \pm \sqrt{1 + \lambda^2\Delta t^2} \ ,$$

subject to the constraint $|\mu| \leq 1$. This implies that $|\lambda\Delta t| \leq 1$ and $\lambda\Delta t$ must be purely imaginary. Hence, the Leap-Frog scheme is only stable when advancing purely advective problems being approximated using a Fourier method, i.e. its use is rather limited. However, used in such a situation, the time step has to be restricted as

$$\Delta t \leq \frac{1}{\max|\lambda^{(1)}|} = \frac{2}{N} = \frac{1}{\pi}\frac{1}{\Delta x} \ ,$$

i.e. the time step must decay linearly with the resolution in order to maintain stability, much like the case of finite difference methods.

In the following we shall give numerous alternatives to the Leap-Frog scheme, in particular methods suitable for the integration of approximations based on polynomials as in this case it should be clear that the Leap-Frog scheme in inappropriate. We shall not perform a detailed stability analysis of all the schemes as such an analysis is fairly trivial, following the exact same approach as discussed above, and can be found in most text books on the solutions of ordinary differential equations.

10.2.1.1 Adams Methods.

A popular choice of explicit time integration schemes are the explicit Adam-Bashforth Methods of which the first reads

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \Delta t \mathrm{L}^n \ ,$$

also known as the 1st order forward Euler scheme. Performing the stability analysis using Eq.(10.1) yields the stability condition

$$|1 + \lambda\Delta t| \leq 1 \ ,$$

which represents the unit circle, centered at $\lambda\Delta t = -1$, i.e. the scheme is stable as long as $\lambda\Delta t$ is inside this stability region. Consequently, using the forward Euler method for integrating an advective periodic problem, approximated using a Fourier method, is inherently unstable, since the imaginary axis is a marginal member of the stability region

figure 10.3. Stability regions for Adam-Bashforth methods of order 2, 3 and 4.

only. However, approximating the operator using a polynomial basis leads to eigenvalues with non-zero real and imaginary parts and, provided Δt is chosen sufficiently small, one may obtain stable schemes by ensuring that $\lambda \Delta t$ remains inside the stability region. This, however, poses very strict constraints on Δt and should not be used except when no alternative is available.

The 2nd order Adam-Bashforth scheme reads

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{2}(3\mathrm{L}^n - \mathrm{L}^{n-1}) \quad ,$$

the still more accurate 3rd order scheme is given as

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{12}(23\mathrm{L}^n - 16\mathrm{L}^{n-1} + 5\mathrm{L}^{n-2}) \quad ,$$

and the 4th order scheme becomes

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{24}(55\mathrm{L}^n - 59\mathrm{L}^{n-1} + 37\mathrm{L}^{n-2} - 9\mathrm{L}^{n-3})$$

The stability regions for these methods may be obtained by the exact

figure 10.4. Stability regions for Adam-Moulton methods of order 3 and 4.

same method as applied for analyzing the Leap-Frog scheme and in Fig. 10.3 we show the stability regions for these high order schemes. Note that the stability region decreases for increasing order of the method, thus placing stronger constraints on the maximum allowable time step. One should also note that while the 1st and 2nd order schemes do not include any part of the imaginary axis, this is no longer true for higher order methods, thus rendering these methods well suited for approximations based on Fourier as well as polynomial methods.

A family closely related to the explicit Adam-Bashforth methods is known as the implicit Adam-Moulton Methods, which, as it first member, contains the scheme

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \Delta t \mathrm{L}^{n+1} \quad ,$$

also know as the 1st order backward Euler method. Performing a stability analysis for this scheme results in a condition as

$$\frac{1}{|1 - \lambda \Delta t|} \leq 1 \quad ,$$

implying that the scheme is stable provided $\lambda\Delta t$ is outside a unit circle centered at $\lambda\Delta t = 1$, i.e. the scheme is A-stable. Since the eigenvalue spectra of the approximated spectral operators all have strictly negative real parts, it is clear that the backward Euler method is unconditionally stable, i.e. independent of the size of Δt. However, one should remember that the error is $\mathcal{O}(\Delta t)$ putting constraints on the time step with respect to accuracy. A-stability is also a property of the 2nd order Adam-Moulton method, also known as the Crank-Nicolson method, being

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{2}(\mathrm{L}^{n+1} + \mathrm{L}^n) \ ,$$

which is 2nd order accurate and widely used for solving diffusion problems. High order Adam-Moulton methods may also be obtained to 3rd order as

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{12}(5\mathrm{L}^{n+1} + 8\mathrm{L}^n - \mathrm{L}^{n-1}) \ ,$$

and the 4th order scheme is given as

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{24}(9\mathrm{L}^{n+1} + 19\mathrm{L}^n - 5\mathrm{L}^{n-1} + \mathrm{L}^{n-2}) \ .$$

Contrary to the lower order members of the family, the 3rd and 4th order schemes are only conditionally stable with the stability regions being shown in Fig. 10.4. One should note that the high order methods does not include the imaginary axis except for the origin, thus rendering them ill suited for Fourier and Legendre approximated advection problems.

Comparing the stability regions of the explicit Adam-Bashforth methods in Fig. 10.3 and those of the implicit Adam-Moulton methods in Fig. 10.4 it is clear that the latter has a stability region being roughly ten timers larger than the former. Moreover, the implicit methods have a smaller truncation error, making them more accurate, however at the expense of requiring the solution of an implicit set of equations.

A common way of combining the advantages of an explicit scheme with the higher accuracy and increased stability region of an implicit scheme is to use the Adam-Bashforth methods as a predictor to the Adam-Moulton methods, yielding the Adam-Bashforth Predictor-Corrector Methods of which the 2nd order scheme is

figure 10.5. Stability regions for Adam-Bashforth Predictor-Corrector methods of order 2, 3 and 4.

$$\begin{aligned} \boldsymbol{u}^* &= \boldsymbol{u}^n + \frac{\Delta t}{2}(3\mathrm{L}^n - \mathrm{L}^{n-1}) \ , \\ \boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{2}(\mathrm{L}^* + \mathrm{L}^n) \ , \end{aligned}$$

where $\mathrm{L}^* = \mathrm{L}(\boldsymbol{u}^*, n\Delta t)$, i.e. the solution obtained using the predictor is taking the role of the solution in the implicit corrector, yielding a two-step explicit scheme. The widely used 3rd order predictor-corrector scheme is given as

$$\begin{aligned} \boldsymbol{u}^* &= \boldsymbol{u}^n + \frac{\Delta t}{12}(23\mathrm{L}^n - 16\mathrm{L}^{n-1} + 5\mathrm{L}^{n-2}) \ , \\ \boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{12}(5\mathrm{L}^* + 8\mathrm{L}^n - \mathrm{L}^{n-1}) \ , \end{aligned}$$

while a 4th order scheme is obtained directly by combining the explicit and implicit Adams methods discussed above.

Compared to the one-step explicit schemes, the predictor-corrector

methods have a larger stability region, see Fig. 10.5, and are more accurate although less than for the purely implicit scheme. We observe in Fig. 10.5 that only the stability region of the 3rd order predictor-corrector contains part of the imaginary axis, explaining the wide use of this particular method for Fourier as well as polynomial based methods.

Contrary to the explicit one-step schemes, the predictor-corrector schemes require two evaluations of L, which may be costly. This tradeoff between higher accuracy and more computations makes it hard to generally state that the predictor-corrector methods are the best choice. Indeed, for many problems the explicit Adams-Bashforth methods may be a better choice. However, it is possible to change these two-step schemes such that only one computation of L is required, known as partially corrected schemes contrary to the standard fully corrected schemes. Indeed, the partially corrected two-step schemes have considerably smaller errors than the explicit one-step schemes, although also a slightly smaller stability region that the fully corrected scheme. The partially corrected 3rd order predictor-corrector scheme is given as

$$\begin{aligned}\tilde{\boldsymbol{u}}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{12}(23\tilde{\mathrm{L}}^n - 16\tilde{\mathrm{L}}^{n-1} + 5\tilde{\mathrm{L}}^{n-2}) \ , \\ \boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{12}(5\tilde{\mathrm{L}}^{n+1} + 8\tilde{\mathrm{L}}^n - \tilde{\mathrm{L}}^{n-1}) \ ,\end{aligned}$$

where $\tilde{\mathrm{L}}^n = \mathrm{L}(\tilde{\boldsymbol{u}}^n, n\Delta t)$, i.e. only after the predictor step does one need to compute the time derivative thereby making the required computational work similar to that of a one-step scheme, albeit with higher accuracy.

10.2.1.2 Backward Differentiation Schemes.

As an alternative to the Adam schemes, one can use the Backward Differentiation Formulas (BDF), being implicit schemes, the first of which is the backward Euler scheme. The 2nd order scheme is given as

$$3\boldsymbol{u}^{n+1} - 4\boldsymbol{u}^n + \boldsymbol{u}^{n-1} = 2\Delta t \mathrm{L}^{n+1} \ .$$

Comparing with the Adam-Moulton methods we observe that the solution, $\boldsymbol{u}^n$, at the previous time steps rather than its time derivative, L^n, needs to be stored to advance in time.

For some types of problems this has a significant advantage and BDF-methods are widely used for solving e.g. stiff systems of diffusive prob-

figure 10.6. Stability regions for BDF methods of order one to six.

lems. The 3rd order scheme yields

$$11\boldsymbol{u}^{n+1} - 18\boldsymbol{u}^{n} + 9\boldsymbol{u}^{n-1} - 2\boldsymbol{u}^{n-2} = 6\Delta t \mathrm{L}^{n+1} \quad ,$$

while the 4th order scheme is given as

$$25\boldsymbol{u}^{n+1} - 48\boldsymbol{u}^{n} + 36\boldsymbol{u}^{n-1} - 16\boldsymbol{u}^{n-2} + 3\boldsymbol{u}^{n-3} = 12\Delta t \mathrm{L}^{n+1} \quad .$$

In Fig. 10.6 we show the stability regions of the BDF methods and note that the first and second order schemes are A-stable while this property is lost for higher order schemes. A scheme is stable if $\lambda\Delta t$ is outside the stability region for all eigenvalues of the operator, since we are dealing with implicit methods.

10.2.1.3 Semi-Implicit Schemes.

The advection-diffusion equation, or indeed the incompressible Navier-Stokes equation, plays a very important role in many branches of science and engineering and special schemes tailored for the efficient solution of such problems has received significant attention. Purely explicit schemes may be very expensive due to the requirement of a very small time-step,

this particularly being true when using polynomial methods, where the eigenvalues of the diffusive operator grows like N^4. On the other hand, the advective part of the equation is often non-linear making a purely implicit solution difficult and costly.

If we consider the general discretized advection-diffusion equation

$$\frac{d\boldsymbol{u}}{dt} = \mathrm{L}(\boldsymbol{u}, t) = -\mathrm{F}(\boldsymbol{u}, t) + \mathrm{G}(\boldsymbol{u}, t) \ ,$$

then the advective operator, $\mathrm{F}(\boldsymbol{u}, t)$, is often non-linear while the diffusive operator, $\mathrm{G}(\boldsymbol{u}, t)$, is linear in $\boldsymbol{u}$. This observation has lead to the introduction of the Semi-Implicit Schemes where the non-linear part is advanced in time using an explicit scheme while the linear part is dealt with using an implicit scheme, thereby, at least partially, avoiding the influence of the diffusive operator on the time step restriction.

The most straightforward semi-implicit scheme is obtained by using the 2nd order Adam-Bashforth scheme for the non-linear part and the Crank-Nicholson scheme for the diffusive part as

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \frac{\Delta t}{2}(3\mathrm{F}^n - \mathrm{F}^{n-1}) + \frac{\Delta t}{2}(\mathrm{G}^{n+1} + \mathrm{G}^n) \ ,$$

yielding a scheme in which only the linear part of the equation requires to be solved implicitly. Since the Crank-Nicholson scheme is A-stable, the total scheme has stability region as the explicit Adam-Bashforth method, however, only for the advective part thereby avoiding the effects of the diffusive operator on the time step. It is certainly possible to use a higher order scheme for advective part, however, the scheme will remain second order. To achieve higher order more elaborate schemes, known as Stiffly Stable Schemes, needs to be considered.

The stiffly stable schemes are obtained by combining the BDF methods with an explicit multi-step scheme, specially tailored for stability and accuracy, for advancing the non-linear part of the equation. The 1st order scheme, being nothing more than a combination of the forward Euler scheme for the non-linear part with the backward Euler scheme for the linear part, yields

$$\boldsymbol{u}^{n+1} - \boldsymbol{u}^n = \Delta t \mathrm{F}^n + \Delta t \mathrm{G}^{n+1} \ .$$

The 2nd order scheme is given as

$$3\boldsymbol{u}^{n+1} - 4\boldsymbol{u}^n + \boldsymbol{u}^{n-1} = 2\Delta t(2\mathrm{F}^n - \mathrm{F}^{n-1}) + 2\Delta t \mathrm{G}^{n+1} \ ,$$

while the even more accurate, and widely used, 3rd order scheme is given as

$$11\boldsymbol{u}^{n+1}-18\boldsymbol{u}^{n}+9\boldsymbol{u}^{n-1}-2\boldsymbol{u}^{n-2} = 6\Delta t(3\mathrm{F}^{n}-3\mathrm{F}^{n-1}+\mathrm{F}^{n-2})+6\Delta t\mathrm{G}^{n+1} \quad .$$

The connection with the BDF schemes is clear and we note that these schemes all are one-step schemes. It is generally found that the stability is governed by the nonlinear term with a time-step restriction being close to that of the explicit Adam-Bashforth methods although slightly smaller time step in general is required.

10.2.2 Runge-Kutta Schemes.

A popular and highly efficient alternative to the multi-step schemes is known as the Runge-Kutta Methods, given on the general form

$$\begin{aligned}
\boldsymbol{k}_1 &= \mathrm{L}(\boldsymbol{u}^n, n\Delta t) = \mathrm{L}^n \\
\boldsymbol{k}_i &= \mathrm{L}(\boldsymbol{u}^n + \Delta t \sum_{j=1}^{i-1} a_{ij}\boldsymbol{k}_i, (n+c_i)\Delta t) \\
\boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \Delta t \sum_{i=1}^{s} b_i \boldsymbol{k}_i \quad .
\end{aligned}$$

Such a scheme is termed an s-stage explicit Runge-Kutta scheme, where the choice of the constants a_{ij}, c_i and b_i determines the accuracy and efficiency of the overall scheme.

For linear operators, L, the s-stage Runge-Kutta schemes are nothing else than the Taylor expansion of the matrix exponential to order s. The main difference between the Runge-Kutta schemes and the multi-step schemes is that the former require more evaluations of L to advance a time step while no information from previous time-steps is required as is the case for multi-step schemes.

10.2.2.1 Standard Schemes.

A popular 2nd order scheme is found for $c_2 = a_{21} = \frac{1}{2}$ and $b_2 = 1$ and zero otherwise. This scheme is known as the midpoint method and yields

$$\boldsymbol{k}_1 = \mathrm{L}(\boldsymbol{u}^n, n\Delta t) = \mathrm{L}^n$$

$$\begin{aligned}
\boldsymbol{k}_2 &= \mathrm{L}(\boldsymbol{u}^n + \tfrac{1}{2}\Delta t \boldsymbol{k}_1, (n+\tfrac{1}{2})\Delta t) \\
\boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \Delta t \boldsymbol{k}_2 \quad ,
\end{aligned}$$

being a 2-stage scheme. An alternative 2nd order accurate, 2-stage scheme is known as the Heun method

$$\begin{aligned}
\boldsymbol{k}_1 &= \mathrm{L}(\boldsymbol{u}^n, n\Delta t) = \mathrm{L}^n \\
\boldsymbol{k}_2 &= \mathrm{L}(\boldsymbol{u}^n + \Delta t \boldsymbol{k}_1, (n+1)\Delta t) \\
\boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{2}(\boldsymbol{k}_1 + \boldsymbol{k}_2) \quad .
\end{aligned}$$

Stability of these schemes is established by considering the scalar equation

$$\frac{du}{dt} = \lambda u \quad ,$$

for which the general s-stage scheme can be expressed as a truncated Taylor expansion of the exponential functions as

$$u^{n+1} = \sum_{i=1}^{s} \frac{(\lambda \Delta t)^i}{i!} u^n \quad ,$$

and, consequently, stability is obtained provided

$$\left| \sum_{i=1}^{s} \frac{(\lambda \Delta t)^i}{i!} \right| \leq 1 \quad ,$$

for which the stability region for $\lambda \Delta t$ may be obtained.

A popular 3rd order 3-stage scheme is the Heun scheme given as

$$\begin{aligned}
\boldsymbol{k}_1 &= \mathrm{L}(\boldsymbol{u}^n, n\Delta t) = \mathrm{L}^n \\
\boldsymbol{k}_2 &= \mathrm{L}(\boldsymbol{u}^n + \tfrac{1}{3}\Delta t \boldsymbol{k}_1, (n+\tfrac{1}{3})\Delta t) \\
\boldsymbol{k}_3 &= \mathrm{L}(\boldsymbol{u}^n + \tfrac{2}{3}\Delta t \boldsymbol{k}_2, (n+\tfrac{2}{3})\Delta t) \\
\boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{4}(\boldsymbol{k}_1 + 3\boldsymbol{k}_3) \quad .
\end{aligned}$$

Note, that although being a 3rd order scheme only two storage levels is required.

The classic 4th order accurate, 4-stage scheme is given as

figure 10.7. Stability regions for Runge-Kutta methods.

$$
\begin{aligned}
\boldsymbol{k}_1 &= \mathrm{L}(\boldsymbol{u}^n, n\Delta t) = \mathrm{L}^n \\
\boldsymbol{k}_2 &= \mathrm{L}(\boldsymbol{u}^n + \tfrac{1}{2}\Delta t \boldsymbol{k}_1, (n+\tfrac{1}{2})\Delta t) \\
\boldsymbol{k}_3 &= \mathrm{L}(\boldsymbol{u}^n + \tfrac{1}{2}\Delta t \boldsymbol{k}_2, (n+\tfrac{1}{2})\Delta t) \\
\boldsymbol{k}_4 &= \mathrm{L}(\boldsymbol{u}^n + \Delta t \boldsymbol{k}_3, (n+1)\Delta t) \\
\boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \frac{\Delta t}{6}(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4) \quad ,
\end{aligned}
$$

requiring four storage levels.

In Fig. 10.7 we display the stability regions for the three Runge-Kutta methods

One should note that contrary to the fully explicit Adam-Bashforth methods, the stability regions expand with increasing order of the Runge-Kutta method. However, one should also remember that increasing the order of the scheme also increases the amount of memory and computation required to complete the step.

Observe also, that the 2nd order Runge-Kutta scheme is only marginally stable at the imaginary axis, thus rendering it ill suited for Fourier and

Legendre based schemes for approximating advective operators.

10.2.2.2 Low-Storage Methods.

The requirement for several storage levels, e.g. four for the classical 4th order Runge-Kutta method, leads to excessive memory requirement in particular when dealing with multi-dimensional problems. However, by defining the constants of the Runge-Kutta method properly it is possible to arrive at methods that require only two storage levels, however, at the expense of performing one extra evaluation of L. The introduction of the additional step introduces extra degrees of freedom in the design of the scheme such that the resulting schemes also have a larger stability region making the work per time unit about the same at the classical methods, albeit with less memory requirements.

The s-stage Low-Storage Method is given on the form

$$\begin{aligned} & \boldsymbol{u}_0 = \boldsymbol{u}^n \\ \forall j \in [1,s]: \quad & \begin{cases} \boldsymbol{k}_j = a_j \boldsymbol{k}_{j-1} + \Delta t \mathrm{L}(\boldsymbol{u}_j, (n+c_j)\Delta t) \\ \boldsymbol{u}_j = \boldsymbol{u}_{j-1} + b_j \boldsymbol{k}_j \end{cases} \\ & \boldsymbol{u}^{n+1} = \boldsymbol{u}_s \quad , \end{aligned}$$

where the constants a_j, b_j and c_j are determined to yield the desired order, $s-1$, of the scheme. For the scheme to be self-starting we require that $a_1 = 0$. Note that we need only two storage levels containing, $\boldsymbol{k}_j$ and $\boldsymbol{u}_j$, to advance the solution.

A 4-stage 3rd order Runge-Kutta scheme is obtained using the constants

$$\begin{aligned} a_1 &= 0 \qquad b_1 = \tfrac{1}{3} \; c_1 = 0 \\ a_2 &= -\tfrac{11}{15} \; b_2 = \tfrac{5}{6} \; c_2 = \tfrac{1}{3} \\ a_3 &= -\tfrac{5}{3} \quad b_3 = \tfrac{3}{5} \; c_3 = \tfrac{5}{9} \\ a_4 &= -1 \quad b_4 = \tfrac{1}{4} \; c_4 = \tfrac{8}{9} \end{aligned}$$

The constants for a 5-stage 4th order Runge-Kutta scheme is given as

$$
\begin{array}{lll}
a_1 = 0 & b_1 = \frac{1432997174477}{9575080441755} & c_1 = 0 \\
a_2 = -\frac{567301805773}{1357537059087} & b_2 = \frac{5161836677717}{13612068292357} & c_2 = \frac{1432997174477}{9575080441755} \\
a_3 = -\frac{2404267990393}{2016746695238} & b_3 = \frac{1720146321549}{2090206949498} & c_3 = \frac{2526269341429}{6820363962896} \\
a_4 = -\frac{3550918686646}{2091501179385} & b_4 = \frac{3134564353537}{4481467310338} & c_4 = \frac{2006345519317}{3224310063776} \\
a_5 = -\frac{1275806237668}{842570457699} & b_5 = \frac{2277821191437}{14882151754819} & c_5 = \frac{2802321613138}{2924317926251}
\end{array}
\quad .
$$

These constants are accurate up to 26 digits, being sufficient for most implementations.

The stability regions for these low storage methods are similar to those of the classical methods, although slightly larger. In particular, both low storage schemes contain the imaginary axis as part of the stability region.

11

Penalty Methods

12

Domain Decomposition Methods

A

Norms, Spaces, and Inequalities

In the following we shall briefly review a number of definition and concepts from approximation theory in linear spaces. We do not attempt to give a complete picture of the underlying theories, but rather to provide a simple and very selective introduction to issues of relevance to the development and analysis of spectral methods.

For a more thorough introduction to linear approximation theory and the modern terminology from functional analysis we refer to one of the many excellent texts on the topic, e.g., [**?**].

A.1 Normed Linear Spaces

Consider a normed linear space, V, endowed with the norm $\|\cdot\|$, of functions, u. A function u is said to belong to V if $\|u\|$ is bounded. This, on the other hand, also defines V in terms of all functions, u, as

$$\mathsf{V} = \{u \mid \|u\| < \infty\} \ .$$

We assume, for simplicity, that V is defined over the field of real numbers, R. However, most of the subsequent results also hold for the field of complex numbers.

The norm, $\|\cdot\|$, enables the definition of linear spaces in which all elements $u \in \mathsf{V}$, subject to $\|u\| \neq 0$, are normalized such that $u/\|u\|$ is unity. Such spaces are termed normed linear spaces and, by the defini-

tion of a norm, we immediately recover

$$\begin{array}{ll} \text{Positive definite} & - \|u\| \geq 0 \ , \ \ \|u\| = 0 \Leftrightarrow u \equiv 0 \\ \text{Triangular inequality} & - \|u+v\| \leq \|u\| + \|v\| \\ \text{Homogeneous} & - \|au\| = |a|\|u\| \end{array}$$

Additional useful bounds are given as

$$2\|uv\| \leq \|u\|^2 + \|v\|^2 \ ,$$

$$| \, \|u\| - \|v\| \, | \leq \|u \pm v\| \leq \|u\| + \|v\| \ ,$$

for $u, v \in \mathsf{V}$.

A.2 Banach Spaces

Let us use the norm, $\|\cdot\|$, to define a metric, $d(u,v) = \|u - v\|$, such that the distance between elements in V can be measured. Let us furthermore introduce the Cauchy sequence, $\{u_n\}_{n=0}^{\infty} \in \mathsf{V}$, defined through the condition

$$\forall \varepsilon > 0 \, \exists N \, \forall m, n > N \, : \, d(u_n, u_m) < \varepsilon \ ,$$

or, equivalently,

$$\lim_{m,n \to \infty} d(u_n, u_m) = 0 \ .$$

If, indeed, any Cauchy sequence in V is convergent to an element in V, the normed linear space, $(\mathsf{V}, \|\cdot\|)$, is termed complete.

The completeness of $(\mathsf{V}, \|\cdot\|)$ has the important consequence that any vector, $u \in \mathsf{V}$, can be approximated arbitrarily close by elements contained entirely in V, i.e., there exists a Cauchy sequence for which $\lim_{n \to \infty} d(u, u_n) = 0$, where closeness is measured using the norm, $\|\cdot\|$.

Complete normed linear spaces are also known as Banach spaces and play an important role in many areas of analysis, e.g., a prominent member is any finite dimensional normed space, e.g., R^N, endowed with the p-norm

$$\|u\|_p = \left(\sum_{i=0}^{N} |u_i|^p \right)^{1/p} \ .$$

A.2.1 The Spaces of Continuous Functions, $C^m[\mathsf{D}]$.

A particularly important example of a Banach space is the space of continuous and continuously differentiable functions, denoted by $C^m[\mathsf{D}]$ where D is a bounded subset of R. If we, for simplicity, restrict ourselves to functions of one variable and define

$$u^{(n)} = \mathcal{D}^{(n)}u = \frac{d^n}{dx^n}u \quad ,$$

then the space $C^m[\mathsf{D}]$ defined on $\mathsf{D} \subset \mathsf{R}$ is a Banach space with the norm

$$\|u\| = \sum_{n \leq m} \max_{x \in \mathsf{D}} |\mathcal{D}^{(n)}u| \quad .$$

Hence, $C^m[\mathsf{D}]$ is the space of function, defined on D that have at least m continuous derivatives. $C^m[\mathsf{D}]$ is clearly complete since every sequence will converge to a continuous function, i.e., an element in the space itself.

We note that the restriction of the above definitions of $C^m[\mathsf{D}]$ to functions of one variable by no means is necessary and generalizations to multiple dimensions is straightforward.

A.2.2 The $L^p[\mathsf{D}]$ and $L^p_w[\mathsf{D}]$ Spaces.

The $L^p[\mathsf{D}]$ spaces are defined as consisting of functions, $u(\boldsymbol{x})$, for which $|u(\boldsymbol{x})|^p$ is Lebesgue integrable, and endowed with the norm

$$\|u\|_{L^p[\mathsf{D}]} = \left(\int_{\mathsf{D}} |u(\boldsymbol{x})|^p d\boldsymbol{x} \right)^{1/p} \quad ,$$

where $1 \leq p < \infty$

A generalization of the $L^p[\mathsf{D}]$ spaces involves the weighted $L^p_w[\mathsf{D}]$ space, endowed with the norm

$$\|u\|_{L^p_w[\mathsf{D}]} = \left(\int_{\mathsf{D}} |u(\boldsymbol{x})|^p w(\boldsymbol{x}) d\boldsymbol{x} \right)^{1/p} \quad ,$$

where $1 \leq p < \infty$ and $w(\boldsymbol{x}) \in L^1[\mathsf{D}]$ is a strictly positive weightfunction.The weighted $L^p_w[\mathsf{D}]$ spaces play a central role in the analysis of spectral methods based on polynomials.

The definition of the $L^\infty[\mathsf{D}]$ is enabled through the norm

$$\|u\|_{L^\infty[\mathsf{D}]} = \sup_{x\in\mathsf{D}} |u(x)| \quad .$$

Note that there is no such thing as a weighted L^∞ norm.

The generalized triangle inequality in $L^p_w[\mathsf{D}]$, known as Minkowskis inequality and valid for $1 < p \leq \infty$, reads

$$\|u+v\|_{L^p_w[\mathsf{D}]} \leq \|u\|_{L^p_w[\mathsf{D}]} + \|v\|_{L^p_w[\mathsf{D}]} \quad ,$$

where $u, v \in L^p_w[\mathsf{D}]$.

A result, exclusive to $L^p_w[\mathsf{D}]$ spaces and known as the Hölder inequality, takes the form

$$\left| \int_{\mathsf{D}} u(\boldsymbol{x}) v(\boldsymbol{x}) w(\boldsymbol{x})\, d\boldsymbol{x} \right| \leq \|u\|_{L^p_w[\mathsf{D}]} \, \|v\|_{L^q_w[\mathsf{D}]} \, \|w\|_{L^r_w[\mathsf{D}]}$$

provide that

$$1 = \frac{1}{p} + \frac{1}{q} + \frac{1}{r} \quad ,$$

where $p, q, r > 1$.

It is also worth recalling a few relations between the different $L^p_w[\mathsf{D}]$ spaces. In particular, they form a sequence of Banach spaces since

$$1 \leq q < p \leq \infty \;:\; L^p_w[\mathsf{D}] \subset L^q_w[\mathsf{D}] \quad ,$$

i.e., $L^1_w[\mathsf{D}]$ is the largest and $L^\infty[\mathsf{D}]$ the smallest of the spaces.

The norms associated with the $L^p_w[\mathsf{D}]$ spaces are all equivalent for the same weight, $w(x)$, as

$$\forall 1 \leq p, q \leq \infty \; \exists C_1, C_2 \;:\; C_1 \|u\|_{L^p_w[\mathsf{D}]} \leq \|u\|_{L^q_w[\mathsf{D}]} \leq C_2 \|u\|_{L^p_w[\mathsf{D}]} \quad ,$$

provided only that u are in $L^p_w[\mathsf{D}]$ as well as $L^q_w[\mathsf{D}]$.

A.2.3 Dense Subspaces

Assume that $(\mathsf{V}, \|\cdot\|)$ is a Banach space and let $\mathsf{S} \subset \mathsf{V}$ be a subset of V. We shall term S dense in V if any element $u \in \mathsf{V}$ can be approximated by a sequence, $u_n \in \mathsf{S}$, such that

$$\|u - u_n\| \to 0 \quad , \quad n \to \infty \quad .$$

Hence, any element in V can be approximated arbitrarily well by elements in S as measured through the norm associated with V.

As a prominent example, we have that $C^0[\mathsf{D}]$ is dense in $L^2_w[\mathsf{D}]$ since for any $u \in L^2_w[\mathsf{D}]$ there exists a sequence, $u_n \in C^0[\mathsf{D}]$, such that

$$\int_{\mathsf{D}} |u - u_n|^2 \, dx \leq \frac{1}{n^2} \quad ,$$

i.e. any function in $L^2_w[\mathsf{D}]$ can be represented arbitrarily well by a sequence of continuous functions.

A.3 Hilbert Spaces

A Banach space endowed with an inner product norm is known as a Hilbert space, e.g., $L^2_w[\mathsf{D}]$ constitutes a Hilbert space, H, with the weighted inner product

$$(u, v)_{L^2_w[\mathsf{D}]} = \int_{\mathsf{D}} u(\boldsymbol{x}) \, \overline{v}(\boldsymbol{x}) w(\boldsymbol{x}) \, d\boldsymbol{x} \quad ,$$

and the associated weighted norm

$$\|u\|_{L^2_w[\mathsf{D}]} = \left(\int_{\mathsf{D}} |u(\boldsymbol{x})|^2 w(\boldsymbol{x}) d\boldsymbol{x} \right)^{1/2} \quad .$$

Here $\overline{v}(\boldsymbol{x})$ refers to the complex conjugate $v(\boldsymbol{x})$, which is introduced to ensure symmetry of the inner product for functions defined on the complex field, C.

The Hilbert space plays a key role in modern numerical analysis due to its similarity with the more familiar notion of an Euclidean space. This enables the use of geometric intuition e.g., the validity of the triangle inequality can be appreciated in terms of the geometry of a triangle.

Moreover, the Cauchy-Schwarz inequality, which is a special case of the Hölder inequality, on the form

$$\left| \int_{\mathsf{D}} u(\boldsymbol{x}) v(\boldsymbol{x}) w(\boldsymbol{x}) \, d\boldsymbol{x} \right| \leq \|u\|_{L^2_w[\mathsf{D}]} \, \|v\|_{L^2_w[\mathsf{D}]} \quad ,$$

has a clear equivalence in the scalar product in an Euclidean geometry.

Taking this notion further, we shall call two functions $u, v \in \mathsf{H}$ orthogonal provided

$$(f, g)_{L^2_w[\mathsf{D}]} = 0 \Leftrightarrow f \perp g \quad .$$

For such functions, the Pythagorean theorem holds on a general Hilbert space as

$$u \perp v \Leftrightarrow \|u+v\|^2_{L^2_w[\mathsf{D}]} = \|u\|^2_{L^2_w[\mathsf{D}]} + \|v\|^2_{L^2_w[\mathsf{D}]} \ .$$

It is also this geometric interpretation of the Hilbert spaces that makes it natural to talk about projections onto finite dimensional spaces in that it has a simple analogy in the projection of vector components onto perpendicular basis vectors within an Euclidean geometry.

Expansions of functions, $u \in \mathsf{H}$, using orthogonal basis functions, $\phi_k \in \mathsf{H}$, play a special role in the development of spectral methods. Such orthogonal expansions,

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k \phi_k(x) \ \ , \ \ \hat{u}_k = \frac{1}{\gamma_k}(u,\phi_k)_{L^2_w[\mathsf{D}]} \ \ , \ \ \gamma_k = \|\phi_k\|^2_{L^2_w[\mathsf{D}]} \ \ ,$$

has the important property

$$\|u\|^2_{L^2_w[\mathsf{D}]} \leq \sum_{k=0}^{\infty} \gamma_k |\hat{u}_k|^2 \ \ ,$$

known as Bessel's inequality.

If, furthermore, the basis $\phi_k(x)$ is complete in H, i.e., if

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k \phi_k(x) \ \ ,$$

and

$$\lim_{N\to\infty} \|u - u_N\|_{L^2_w[\mathsf{D}]} = 0 \ \ ,$$

one recovers

$$\|u\|^2_{L^2_w[\mathsf{D}]} = \sum_{k=0}^{\infty} \gamma_k |\hat{u}_k|^2 \ \ ,$$

known as Parsevals identity.

A.4 Sobolev Spaces

Let us finally introduce the classical Sobolev spaces which provide a powerful framework in which to study linear operators.

We define the Sobolev space, $H_w^m[\mathsf{D}]$, of functions u as

$$H_w^m[\mathsf{D}] = \left\{ u \in L_w^2[\mathsf{D}] \middle| \, \forall n \in [0,m], u^{(n)} \in L_w^2[\mathsf{D}] \right\} \quad ,$$

where m is an integer.

The $H_w^m[\mathsf{D}]$ space is endowed with the inner product

$$(u,v)_{H_w^m[\mathsf{D}]} = \sum_{n=0}^{m} \int_{\mathsf{D}} u^{(n)}(\boldsymbol{x}) v^{(n)}(\boldsymbol{x}) w(\boldsymbol{x}) \, d\boldsymbol{x} \quad ,$$

and the norm

$$\|u\|_{H_w^m[\mathsf{D}]} = \left(\sum_{n=0}^{m} \|u^{(n)}\|_{L_w^2[\mathsf{D}]}^2 \right)^{1/2} \quad .$$

The Sobolev spaces form a hierarchy of Hilbert spaces in the sense that $H_w^{m+1}[\mathsf{D}] \subset H_w^m[\mathsf{D}] \subset \ldots \subset H_w^0[\mathsf{D}] = L_w^2[\mathsf{D}]$. Moreover, it is clear that $f \in C^m[\mathsf{D}] \Rightarrow f \in H^m[\mathsf{D}]$, i.e., $C^m[\mathsf{D}]$ is a dense subspace of $H^m[\mathsf{D}]$ as any function $u \in H^m[\mathsf{D}]$ can be approximated arbitrarily well by an element in $C^m[\mathsf{D}]$ as discussed in Sec. A.2.3. Conversely, if $u \in C^{m-1}[\mathsf{D}]$ then it is also true that $u \in H^m[\mathsf{D}]$, i.e., $H^m[\mathsf{D}] \subset C^{m-1}[\mathsf{D}]$.

As for the $L_w^p[\mathsf{D}]$ norms, the Sobolev norms are all equivalent, i.e.,

$$\forall 1 \leq p,q \leq \infty \; \exists C_1, C_2 \quad : \quad C_1 \|u\|_{H_w^p[\mathsf{D}]} \leq \|u\|_{H_w^q[\mathsf{D}]} \leq C_2 \|u\|_{H_w^p[\mathsf{D}]} \quad ,$$

provided only that u are in $H_w^p[\mathsf{D}]$ as well as $H_w^q[\mathsf{D}]$.

The definition of the Sobolev spaces given above is for functions of one variable only. However, the developments as well as the definitions generalize straightforwardly to problems in multiple dimensions.

Let us finally state a couple of inequalities that shall become useful in the subsequent analysis. Assume that $\mathsf{D} = [a,b]$ signifies a bounded interval on R and that we consider a function, $u \in H^1[\mathsf{D}]$. We have the Sobolev inequality

$$\|u\|_{L^\infty[\mathsf{D}]} \leq \left(2 + \frac{1}{b-a} \right)^{1/2} \|u\|_{L^2[\mathsf{D}]}^{1/2} \|u\|_{H^1[\mathsf{D}]}^{1/2} \quad .$$

Here, and in what remains, we shall by $H^m[\mathsf{D}]$ refer to the Sobolev space, $H_w^m[\mathsf{D}]$, with $w(x) = 1$ for brevity. This Sobolev inequality is important as it relates the $L^p[\mathsf{D}]$ spaces with the $H^1[\mathsf{D}]$ Sobolev space.

Also, we shall find it useful to recall the Poincaré inequality for a

function $u \in H^1[\mathsf{D}]$ where again $\mathsf{D} = [a, b]$ is assumed bounded. Then there exists a constant C such that

$$\|u\|_{L^2_w[\mathsf{D}]} \leq C\|u^{(1)}\|_{L^2_w[\mathsf{D}]} \ ,$$

connecting the function with its derivative.

A.5 Notation for Periodic Functions

In working with periodic functions, it is natural to slightly modify the meaning of the spaces and norms discussed above.

We shall define a function, $u(x)$, as being periodic in the interval $[0, 2\pi]$ if $u(0)$ and $u(2\pi)$ both exist and are equal, i.e., $u(0) = u(2\pi)$.

Based on this, we define the space, $\mathsf{C}^m_p[0, 2\pi]$, as the space of functions for which $u^{(n)}, n \leq m$ is continuous and periodic.

In a similar fashion, we define the Sobolev space, $\mathsf{H}^m_p[0, 2\pi]$, endowed with the inner product

$$(u, v)_{H^m_p[0,2\pi]} = \left(\sum_{n=0}^{m} \|u^{(n)}\|^2_{L^2[0,2\pi]} \right)^{1/2} \ ,$$

and the norm

$$\|u\|^2_{H^m_p[0,2\pi]} = (u, u)_{H^m_p[0,2\pi]} \ ,$$

as the space of periodic functions for which $\|u\|_{H^m_p[0,2\pi]}$ is bounded, i.e., it coincides with the space of functions, $u \in \mathsf{C}^{m-1}_p[0, 2\pi]$ for which $u^{(m)} \in L^2[0, 2\pi]$.

A.6 Linear Operators and Operator Norms

Let us consider the operator, $\mathcal{L} : \mathsf{V} \to \mathsf{V}$, where V is a Banach space. The operator is termed linear if

$$\mathcal{L}(au + bv) = a\mathcal{L}u + b\mathcal{L}v \ ,$$

where $u, v \in \mathsf{V}$ and a, b are scalars.

For linear operators, we define the subordinate norm

$$\|\mathcal{L}\| = \sup_{\|u\| \neq 0} \frac{\|\mathcal{L}u\|}{\|u\|} = \sup_{\|u\|=1} \|\mathcal{L}u\| \ , \quad u \in \mathsf{V} \ .$$

The operator is bounded if there exists a constant, C, such that

$$\|\mathcal{L}u\| \leq \|\mathcal{L}\|\|u\| \leq C\|u\| \quad ,$$

for all $u \in \boldsymbol{V}$. Conversely, is no such constant exists, the operator is termed unbounded.

Restricting the attention to V being a Hilbert space, i.e., we can define an inner product $(\cdot, \cdot)$ associated with the norm, we have that

$$(\mathcal{L}u, v) = (u, \mathcal{L}^* v) \quad ,$$

where $\mathcal{L}^*$ is termed the adjoint operator.

If there exists a constant, C, such that

$$\mathcal{L} + \mathcal{L}^* \leq C\mathcal{I} \quad ,$$

in the sense that

$$(u, (\mathcal{L} + \mathcal{L}^*)u) \leq C\|u\|^2 \quad ,$$

we shall call $\mathcal{L}$ a semi-bounded operator.

B

The Gamma Function

Let us briefly review some properties of the Euler-Gamma function, $\Gamma(x)$, defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t)\, dt \ ,$$

which is analytic for $x > 0$ and with a pole at $x = 0$. However, integration by parts immediately yields the recursive formula

$$\Gamma(x+1) = x\Gamma(x) \ ,$$

i.e., the pole at $x = 0$ is simple.

For x being an integer, this simple recurrence yields the important identity

$$\Gamma(n+1) = n! \ .$$

Useful formulas, of which many more can found in [?], include

$$\Gamma(2x) = \frac{1}{\sqrt{\pi}} 2^{2x-1} \Gamma(x)\Gamma(x+\tfrac{1}{2}) \ ,$$

known as the duplication formula, and the relation with binomial coefficients

$$\binom{x}{k} = \frac{x!}{k!(x-k)!} = \frac{\Gamma(x+1)}{\Gamma(k+1)\Gamma(x-k+1)} \ , \quad x > k-1.$$

Special values of $\Gamma(x)$ include

$$\Gamma(1) = 1 \ , \ \ \Gamma(\tfrac{1}{2}) = \sqrt{\pi} \ ,$$

while a useful asymptotic expression, known as Stirlings formula, reads

$$\Gamma(x) \simeq \sqrt{2\pi}\exp(-x)\frac{x^x}{\sqrt{x}}\left[1 + \frac{1}{12x} + \frac{1}{288x^2} - \frac{139}{51840x^3} - ...\right] \ , \ \ |x| \to \infty \ .$$

Recalling the close connection to the Beta function,

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \ ,$$

we recover

$$\int_{-1}^{1}(1-t)^{\alpha}(1+t)^{\beta}\,dx = 2^{\alpha+\beta+1}\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} \ .$$

C

A Zoo of Polynomials

C.1 Legendre Polynomials

The Legendre polynomials, $P_n(x)$, are defined as the solution to the Sturm-Liouville problem with $p(x) = 1 - x^2$, $q(x) = 0$ and $w(x) = 1$ as

$$\frac{d}{dx}(1-x^2)\frac{dP_n(x)}{dx} + n(n+1)P_n(x) = 0 \ ,$$

where $P_n(x)$ is assumed bounded for $x \in [-1, 1]$.

The Legendre polynomials are given as $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, $P_3(x) = \frac{1}{2}(5x^3 - 3x)$ and are orthogonal in $L^2_w[-1, 1]$ with as $w(x) = 1$ as

$$\int_{-1}^{1} P_n(x)P_m(x)\,dx = \frac{2}{2n+1}\delta_{mn} \ .$$

C.1.1 The Legendre Expansion

The continuous expansion is given as

$$u(x) = \sum_{n=0}^{N} \hat{u}_n P_n(x) \ \ , \ \ \hat{u}_n = \frac{2n+1}{2}\int_{-1}^{1} u(x)P_n(x)\,dx \ .$$

The discrete expansion coefficients depend on what family of Gauss points are chosen.

Legendre Gauss Quadrature

$$z_j = \{z|P_{N+1}(z) = 0\} \ , \ u_j = \frac{2}{(1-z_j^2)[P'_{N+1}(z_j)]^2} \ , \ j \in [0, \ldots, N] \ .$$

The normalization constant is given as

$$\tilde{\gamma}_n = \frac{2}{2n+1} \ ,$$

resulting in the expansion coefficients as

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(z_j) P_n(z_j) u_j \ .$$

Legendre Gauss-Radau Quadrature

$$y_j = \{y|P_N(y) + P_{N+1}(y) = 0\} \ ,$$

$$v_j = \begin{cases} \frac{2}{(N+1)^2} & j = 0 \\ \frac{1}{(N+1)^2} \frac{1-y_j}{[P_N(y_j)]^2} & j \in [1, \ldots, N] \end{cases} \ .$$

The normalization constant is given as

$$\tilde{\gamma}_n = \frac{2}{2n+1} \ ,$$

yielding the discrete expansion coefficients as

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(y_j) P_n(y_j) v_j \ .$$

Legendre Gauss-Lobatto Quadrature

$$x_j = \{x|(1-x^2)P'_N(x) = 0\} \ , \ w_j = \frac{2}{N(N+1)} \frac{1}{[P_N(x_j)]^2} \ .$$

The normalization constant is given as

$$\tilde{\gamma}_n = \begin{cases} \frac{2}{2n+1} & j \in [0, N-1] \\ \frac{2}{N} & j = N \end{cases} \ ,$$

from which the discrete expansion coefficients become

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) P_n(x_j) w_j \quad .$$

C.1.2 Recurrence and other Relations.

Here we give a number of useful recurrence relations.

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \quad .$$

$$P_n(x) = \frac{1}{2n+1}P'_{n+1}(x) - \frac{1}{2n+1}P'_{n-1}(x) \quad , \quad P_0(x) = P'_1(x) \quad .$$

We also have

$$\int P_n(x)\,dx = \begin{cases} P_1(x) & n = 0 \\ \frac{1}{6}(2P_2(x)+1) & n = 1 \\ \frac{1}{2n+1}P_{n+1}(x) - \frac{1}{2n+1}P_{n-1}(x) & n \geq 2 \end{cases} \quad .$$

$$P_n(-x) = (-1)^n P_n(x) \quad .$$

C.1.3 Special Values.

The Legendre polynomials have the following special values.

$$|P_n(x)| \leq 1 \quad , \quad |P'_n(x)| \leq \frac{1}{2}n(n+1) \quad .$$

$$P_n(\pm 1) = (\pm 1)^n \quad , \quad P'_n(\pm 1) = \frac{(\pm 1)^{n+1}}{2}n(n+1) \quad .$$

The values of P_n at the center $x = 0$ behaves as

$$P_{2n}(0) = (-1)^n \frac{(n-1)!}{(\prod_{i=1}^{n/2} 2i)^2} \quad , \quad P_{2n+1}(0) = 0 \quad .$$

Finally, we obtain the results for integration as

$$\int_{-1}^{1} P_n(x)\,dx = 2\delta_{0n} \quad .$$

C.1.4 Operators.

In the following we will consider the following question for Legendre expansions. Given a polynomial approximation as

$$f(x) = \sum_{n=0}^{\infty} \hat{a}_n P_n(x) \quad , \quad \mathcal{L}f(x) = \sum_{n=0}^{\infty} \hat{b}_n P_n(x) \quad ,$$

where $\mathcal{L}$ is a given operator, what is the relation between $\hat{a}_n$ and $\hat{b}_n$. We will give the result for the most commonly used operators, $\mathcal{L}$.

$$\mathcal{L} = \frac{d}{dx} \quad : \quad \hat{b}_n = (2n+1) \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{\infty} \hat{a}_p \quad .$$

$$\mathcal{L} = \frac{d^2}{dx^2} \quad : \quad \hat{b}_n = \frac{2n+1}{2} \sum_{\substack{p=n+2 \\ p+n \text{ even}}}^{\infty} (p(p+1) - n(n+1))\hat{a}_p \quad .$$

$$\mathcal{L} = x \quad : \quad \hat{b}_n = \frac{n}{2n-1}\hat{a}_{n-1} + \frac{n+1}{2n+3}\hat{a}_{n+1} \quad .$$

Finally, if we have

$$\frac{d^q}{dx^q}u(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} P_n(x) \quad ,$$

then

$$\frac{1}{2n-1}\hat{u}_{n-1}^{(q)} - \frac{1}{2n+3}\hat{u}_{n+1}^{(q)} = \hat{u}_n^{(q-1)} \quad .$$

C.2 Chebyshev Polynomials

The Chebyshev polynomials of the first kind, $T_n(x)$, appear as a solution to the singular Sturm-Liouville problem with $p(x) = \sqrt{1-x^2}$, $q(x) = 0$ and $w(x) = (\sqrt{1-x^2})^{-1}$ as

$$\frac{d}{dx}\left(\sqrt{1-x^2}\frac{dT_n(x)}{dx}\right) + \frac{n^2}{\sqrt{1-x^2}}T_n(x) = 0 \quad ,$$

where $T_n(x)$ is assumed bounded for $x \in [-1, 1]$.

The Chebyshev polynomials may be given on explicit form as

$$T_n(x) = \cos(n \arccos x) \quad .$$

Thus, $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$ etc.

The Chebyshev polynomials are orthogonal in $L^2_w[-1, 1]$

$$\int_{-1}^{1} T_n(x) T_m(x) \frac{1}{\sqrt{1-x^2}} \, dx = \frac{\pi}{2} c_n \delta_{mn} \quad ,$$

where

$$c_n = \begin{cases} 2 & n = 0 \\ 1 & \text{otherwise} \end{cases} \quad .$$

C.2.1 The Chebyshev Expansion.

The continuous expansion is given as

$$u(x) = \sum_{n=0}^{N} \hat{u}_n T_n(x) \quad , \quad \hat{u}_n = \frac{2}{\pi c_n} \int_{-1}^{1} u(x) T_n(x) \frac{1}{\sqrt{1-x^2}} \quad ,$$

where as the details of the discrete expansion depends on what family of Gauss points are chosen.

Chebyshev Gauss Quadrature

$$z_j = -\cos\left(\frac{(2j+1)\pi}{2N+2}\right) \quad , \quad u_j = \frac{\pi}{N+1} \quad , \quad j \in [0, \ldots, N] \quad .$$

The normalization constant is given as

$$\tilde{\gamma}_n = \begin{cases} \pi & n = 0 \\ \frac{\pi}{2} & n \in [1, \ldots, N] \end{cases} \quad ,$$

with the discrete expansion coefficients being

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(z_j) T_n(z_j) u_j \quad .$$

Chebyshev Gauss-Radau Quadrature

$$y_j = -\cos\left(\frac{2j\pi}{2N+1}\right) \quad , \quad v_j = \begin{cases} \frac{\pi}{2N+1} & j = 0 \\ \frac{2\pi}{2N+2} & j \in [1,\ldots,N] \end{cases} \quad .$$

The normalization constant is given as

$$\tilde{\gamma}_n = \begin{cases} \pi & n = 0 \\ \frac{\pi}{2} & n \in [1,\ldots,N] \end{cases} \quad ,$$

yielding the discrete expansion coefficients as

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(y_j) T_n(y_j) v_j \quad .$$

Chebyshev Gauss-Lobatto Quadrature

$$x_j = -\cos\left(\frac{\pi j}{N}\right) \quad , \quad w_j = \begin{cases} \frac{\pi}{2N} & j = 0, N \\ \frac{\pi}{N} & j \in [1,\ldots,N-1] \end{cases} \quad .$$

The normalization constant is given as

$$\tilde{\gamma}_n = \begin{cases} \frac{\pi}{2} & j \in [1, N-1] \\ \pi & j = 0, N \end{cases} \quad ,$$

resulting in the discrete expansion coefficients being

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n} \sum_{j=0}^{N} u(x_j) T_n(x_j) w_j \quad .$$

C.2.2 Recurrence and other Relations.

The number of recurrence relations is large and we will only give the most useful ones.

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad .$$

$$T_n = \frac{1}{2(n+1)} T'_{n+1}(x) - \frac{1}{2(n-1)} T'_{n-1}(x) \quad , \quad T_0(x) = T'_1(x) \quad .$$

Other useful relations are

$$2T_n^2(x) = 1 + T_{2n}(x) \quad .$$

$$2T_n(x)T_m(x) = T_{|n+m|}(x) + T_{|n-m|}(x) \quad .$$

$$\int T_n(x)\,dx = \begin{cases} T_1(x) & n = 0 \\ \frac{1}{4}(T_2(x)+1) & n = 1 \\ \frac{1}{2(n+1)}T_{n+1}(x) - \frac{1}{2(n-1)}T_{n-1}(x) & n \geq 2 \end{cases} \quad .$$

$$T_n(-x) = (-1)^n T_n(x) \quad .$$

C.2.3 Special Values.

From the definition of the Chebyshev polynomials we may make the following observations.

$$|T_n(x)| \leq 1 \quad , \quad |T_n'(x)| \leq n^2 \quad .$$

$$\frac{d^q}{dx^q}T_n(\pm 1) = (\pm 1)^{n+q} \prod_{k=0}^{q-1} \frac{n^2-k^2}{2k+1} \quad ,$$

with the special cases

$$T_n(\pm 1) = (\pm 1)^n \quad , \quad T_n'(\pm 1) = (\pm 1)^{n+1} n^2 \quad .$$

The values of T_n at the center $x = 0$ behaves as

$$T_{2n}(0) = (-1)^n \quad , \quad T_{2n+1}(0) = 0 \quad .$$

$$T_{2n}'(0) = 0 \quad , \quad T_{2n+1}'(0) = (-1)^n n \quad .$$

Finally, we obtain the results for integration as

$$\int_{-1}^{1} T_n(x)\,dx = \begin{cases} -\frac{2}{n^2-1} & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \quad .$$

C.2.4 Operators

In the following we consider the following question for Chebyshev expansions. Given a polynomial approximation as

$$f(x) = \sum_{n=0}^{\infty} \hat{a}_n T_n(x) \quad , \quad \mathcal{L}f(x) = \sum_{n=0}^{\infty} \hat{b}_n T_n(x) \quad ,$$

where $\mathcal{L}$ is a given operator, what is the relation between $\hat{a}_n$ and $\hat{b}_n$. We give the result for the most commonly used operators, $\mathcal{L}$.

$$\mathcal{L} = \frac{d}{dx} \quad : \quad \hat{b}_n = \frac{2}{c_n} \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{\infty} p\hat{a}_p \quad .$$

$$\mathcal{L} = \frac{d^2}{dx^2} \quad : \quad \hat{b}_n = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ p+n \text{ even}}}^{\infty} p(p^2 - n^2)\hat{a}_p \quad .$$

$$\mathcal{L} = \frac{d^3}{dx^3} \quad : \quad \hat{b}_n = \frac{1}{4c_n} \sum_{\substack{p=n+3 \\ p+n \text{ odd}}}^{\infty} p(p^2(p^2 - 2) - 2p^2n^2 + (n^2 - 1)^2)\hat{a}_p \quad .$$

$$\mathcal{L} = \frac{d^4}{dx^4} \quad : \quad \hat{b}_n = \frac{1}{24c_n} \sum_{\substack{p=n+4 \\ p+n \text{ even}}}^{\infty} p(p^2(p^2-4)^2 - 3p^4n^2 + 3p^2n^4 - n^2(n^2-4)^2)\hat{a}_p \quad .$$

$$\mathcal{L} = x \quad : \quad \hat{b}_n = \frac{1}{2}\left(c_{n-1}\hat{a}_{n-1} + \hat{a}_{n+1}\right) \quad .$$

$$\mathcal{L} = x^2 \quad : \quad \hat{b}_n = \frac{1}{4}\left(c_{n-2}\hat{a}_{n-2} + (c_n + c_{n-1})\hat{a}_n + \hat{a}_{n+2}\right) \quad .$$

Finally, if we have

$$\frac{d^q}{dx^q}u(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} T_n(x) \quad ,$$

then

$$c_{n-1}\hat{u}_{n-1}^{(q)} - \hat{u}_{n+1}^{(q)} = 2n\hat{u}_n^{(q-1)} \quad .$$

C.3 Laguerre Polynomials

The Laguerre polynomial, $L_n(x)$, is defined as the solution to the Sturm-Liouville problem with $p(x) = x\exp(-x)$, $q(x) = 0$ and $w(x) = \exp(-x)$ as

$$\frac{d}{dx}x\exp(-x)\frac{dL_n(x)}{dx} + n\exp(-x)L_n(x) = 0 \ ,$$

where $L_n(x)$ is defined for $x \in [0, \infty[$.

The Laguerre polynomials are given as $L_0(x) = 1$, $L_1(x) = 1 - x$, $L_2(x) = \frac{1}{2}x^2 - 2x + 1$, $L_3(x) = -\frac{1}{6}x^3 + \frac{3}{2}x^2 - 3x + 1$ and the Laguerre polynomials are orthogonal in $L^2_w[0, \infty]$ with as $w(x) = \exp(-x)$ as

$$\int_0^\infty L_n(x)L_m(x)\exp(-x)\,dx = \delta_{mn} \ .$$

C.3.1 The Laguerre Expansion

The continuous expansion is given as

$$u(x) = \sum_{n=0}^{N} \hat{u}_n L_n(x) \ \ , \ \ \hat{u}_n = \int_0^\infty u(x)L_n(x)\exp(-x)\,dx \ .$$

The discrete expansion depends on what family of Gauss points are chosen. Here we only consider the Gauss and the Gauss-Radau points as they are the most commonly used, as no quadrature point is situated at infinity.

Laguerre Gauss Quadrature

$$z_j = \{z | L_{N+1}(z) = 0\} \ ,$$

$$u_j = -\frac{1}{N+1}\left[L_N(z_j)L'_{N+1}(z_j)\right]^{-1} \ ,$$

with the discrete expansion coefficients being

$$\tilde{u}_n = \sum_{j=0}^{N} u(z_j)L_n(z_j)u_j \ .$$

Laguerre Gauss-Radau Quadrature

$$y_j = \{y | y_0 = 0 \ , \ L'_{N+1}(y) = 0\} \ ,$$

$$v_j = \begin{cases} \frac{1}{N+1} & j = 0 \\ \frac{1}{N+1} \left[L_{N+1}(y_j) L'_N(y_j) \right]^{-1} & j \in [1, N] \end{cases} \ ,$$

with the discrete expansion coefficients being

$$\tilde{u}_n = \sum_{j=0}^{N} u(y_j) L_n(y_j) v_j \ .$$

C.3.2 Recurrence and other Relations.

Here we give a number of useful recurrence relations.

$$(n+1) L_{n+1}(x) = (2n + 1 - x) L_n(x) - n L_{n-1}(x) \ .$$

$$L_n(x) = L'_n(x) - L'_{n+1}(x) \ , \ \ L_0(x) = -L'_1(x) \ .$$

C.3.3 Special Values.

The Laguerre polynomials have the following special values.

$$|L_n(x) \exp(-x)| \leq 1 \ .$$

$$L_n(0) = 1 \ .$$

C.3.4 Operators.

Consider

$$\frac{d^q}{dx^q} u(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} L_n(x) \ ,$$

then

$$\hat{u}_n^{(q)} = \hat{u}_{n+1}^{(q)} - \hat{u}_{n+1}^{(q-1)} \ .$$

C.4 Hermite Polynomials

The Hermite polynomial, $H_n(x)$, is defined as the solution to the Sturm-Liouville problem with $p(x) = \exp(-x^2)$, $q(x) = 0$ and $w(x) = \exp(-x^2)$ as

$$\frac{d}{dx}\exp(-x^2)\frac{dH_n(x)}{dx} + 2n\exp(-x^2)H_n(x) = 0 \ ,$$

where $H_n(x)$ is defined for $x \in]-\infty, \infty[$.

The Hermite polynomials are given as $H_0(x) = 1$, $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2$, $H_3(x) = 8x^3 - 12x$, with the polynomials being orthogonal in $L^2_w[-\infty, \infty]$ with $w(x) = \exp(-x^2)$ as

$$\int_{-\infty}^{\infty} H_n(x)H_m(x)\exp(-x^2)\,dx = 2^n n!\sqrt{\pi}\delta_{mn} \ .$$

C.4.1 The Hermite Expansion

The continuous expansion is given as

$$u(x) = \sum_{n=0}^{N} \hat{u}_n H_n(x) \ \ , \ \ \hat{u}_n = \frac{1}{2^n n!\sqrt{\pi}} \int_{-\infty}^{\infty} u(x)H_n(x)\exp(-x^2)\,dx \ .$$

Only for the Gauss points is it convenient to introduce the discrete expansion coefficients, since this set of quadrature points does not include the endpoints.

Hermite Gauss Quadrature

$$z_j = \{z | H_{N+1}(z) = 0\} \ ,$$

$$u_j = 4\sqrt{\pi}2^N(N+1)!\left[H'_{N+1}(z_j)\right]^{-2} \ \ , \ \ j \in [0, \ldots, N] \ .$$

The normalization constant is given as

$$\tilde{\gamma}_n = 2^n n!\sqrt{\pi} \ ,$$

leading to the discrete expansion coefficients given as

$$\tilde{u}_n = \sum_{j=0}^{N} u(z_j)H_n(z_j)u_j \ .$$

C.4.2 Recurrence and other Relations.

Here we give a number of useful recurrence relations.

$$H_{n+1}(x) = 2xH_n(x) - nH_{n-1}(x) \quad .$$

$$H_n(x) = \frac{1}{2(n+1)} H'_{n+1}(x) \quad , \quad H_0(x) = \frac{1}{2} H'_1(x) \quad .$$

$$H_n(-x) = (-1)^n H_n(x) \quad .$$

C.4.3 Special Values

The Hermite polynomials have the following special values.

$$|H_n(x) \exp(-x^2)| \leq 1 \quad .$$

$$H_{2n+1}(0) = 0 \quad , H_{2n}(0) = (-1)^n \frac{(2n)!}{n!} \quad .$$

C.4.4 Operators

Consider

$$\frac{d^q}{dx^q} u(x) = \sum_{n=0}^{\infty} \hat{u}_n^{(q)} H_n(x) \quad ,$$

then

$$\hat{u}_{n-1}^{(q)} = 2n \hat{u}_n^{(q-1)} \quad .$$

Bibliography

Bibliography

S. Abarbanel, D. Gottlieb, and E. Tadmor, *Spectral Methods for Discontinuous Problems.* In Numerical Methods for Fluid Dynamics II, K.W. Morton and M.J. Baines (Eds.), Clarendon Press, Oxford, 1986. pp. 129-153.

Ø. Andreassen and I. Lie, *Simulation of Acoustical and Elastic Waves and Interaction*, J. Acous. Soc. Am. **95**(1994), pp. 171-186.

N. S. Banerjee and J. Geer, *Exponential Approximations Using Fourier Series Partial Sums.* ICASE Report No. 97-56, NASA Langley Research Center, VA. 1997.

G. Ben-Yu, *Spectral Methods and Their Applications.* World Scientific, Singapore, 1998.

C. Bernardi and Y. Maday, *Polynomial Interpolation Results in Sobolev Spaces*, J. Comput. Appl. Math. **43**(1992), pp. 53-80.

C. Bernardi and Y. Maday, *Spectral Methods.* In Handbook of Numerical Analysis V by P. G. Ciarlet and J. L. Lions (Eds). Elsevier Sciences, North-Holland, The Netherlands, 1999.

J.P. Boyd, *Two Comments on Filtering (Artificial Viscosity) for Chebyshev and Legendre Spectral and Spectral Element Methods: Preserving Boundary Conditions and Interpretation of the Filter as a Diffusion*, J. Comput. Phys. **143**(1998), pp. 283-288.

W. Cai, D. Gottlieb and C.W. Shu, *Essentially Nonoscillatory Spectral Fourier Methods for Shock Wave Calculations*, Math. Comp. **52**(1989), pp. 389-410.

W. Cai, D. Gottlieb, and A. Harten, *Cell Averaging Chebyshev Methods for Hyperbolic Problems.* In Computers and Mathematics with Applications. Academic Press, New York, 1990.

C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics.* Springer Series in Computational Physics. Springer-Verlag. New York, 1988.

C. Canuto and A. Quarteroni, *Error Estimates for Spectral and Pseudospectral Approximations of Hyperbolic Equations*, SIAM J. Numer. Anal. **19**(1982), pp. 629-642.

C. Canuto and A. Quarteroni, *Approximation Results for Orthogonal Polynomials in Sobolev Spaces*, Math. Comp. **38**(1982), pp. 67-86.

M. H. Carpenter and D. Gottlieb, *Spectral Methods on Arbitrary Grids*, J. Comput. Phys. **129**(1996), pp. 74-86.

B. Cockburn and C.W. Shu, *Discontinuous Galerkin Methods for Convection-Dominated Problems*, SIAM Review, 2000 – submitted.

L. Dettori and B. Yang, *On the Chebyshev Penalty Method for Parabolic and Hyperbolic Equations*, M^2AN **30**(1996), pp. 907-920.

W. S. Don, *Numerical Study of Pseudospectral Methods in Shock Wave Applications*, J. Comput. Phys. **110**(1994), pp. 103-111.

W. S. Don and D. Gottlieb, *The Chebyshev-Legendre Method: Implementing Legendre Methods on Chebyshev Points*, SIAM J. Numer. Anal. **31**(1994), pp. 1519-1534.

W. S. Don and D. Gottlieb, *Spectral Simulation of Supersonic Reactive Flows*, SIAM J. Numer. Anal. **35**(1998), pp. 2370-2384.

W. S. Don and C. B. Quillen, *Numerical Simulation of Reactive Flow. Part I: Resolution*, J. Comput. Phys. **122**(1995), pp. 244-265.

K. S. Eckhoff, *On Discontinuous Solutions of Hyperbolic Equations*, Comput. Methods Appl. Mech. Engrg. **116**(1994), pp. 103-112.

K. S. Eckhoff, *Accurate Reconstructions of Functions of Finite Regularity from Truncated Series Expansions*, Math. Comp. **64**(1995), pp. 671-690.

P. Fischer and D. Gottlieb, *On the Optimal Number of Subdomains for Hyperbolic Problems on Parallel Computers*, Int. J. Supercomput. Appl. High Perform. Comput. **11**(1997), pp. 65-76.

B. Fornberg, *On A Fourier Method for the Integration of Hyperbolic Problems*, SIAM J. Numer. Anal. **12**(1975), pp. 509-528.

B. Fornberg, *A Practical Guide to Pseudospectral Methods*. Cambridge University Press, Cambridge, UK. 1996.

L. Fox, *Chebyshev Methods for Ordinary Differential Equations*, Computer J. **4**(1962), pp. 318-331.

D. Funaro, *Polynomial Approximation of Differential Equations*. Lecture Notes in Physics, **m 8**. Springer Verlag, Berlin. 1992.

D. Funaro and D. Gottlieb, *A New Method of Imposing Boundary Conditions in Pseudospectral Approximations of Hyperbolic Equations*, Math. Comp. **51**(1988), pp. 599-613.

D. Funaro and D. Gottlieb, *Convergence Results for Pseudospectral Approximations of Hyperbolic Systems by a Penalty-Type Boundary Treatment*, Math. Comp. **57**(1991), pp. 585-596.

A. Gelb and D. Gottlieb, *The Resolution of the Gibbs Phenomenon for Spliced Functions in One and Two Dimensions*, Computers Math. Applic. **33**(1997), pp. 35-58.

A. Gelb and E. Tadmor, *Enhanced Spectral Viscosity Approximations for Conservation Laws*, Appl. Numer. Math. 2000 – to appear.

J. G. Giannakouros and G. E. Karniadakis, *Spectral Element-FCT Method for the Compressible Euler Equations*, J. Comput. Phys. **115**(1994), pp. 65-85.

J. G. Giannakouros, D. Sidilkover, and G. E. Karniadakis, *Spectral Element-FCT Method for the One- and Two-Dimensional Compressible Euler Equations*, Comput. Methods Appl. Mech. Engrg. **116**(1994), pp. 113-121.

J. Goodman, T. Hou, and E. Tadmor, *On the Stability of the Unsmoothed*

Fourier Method for Hyperbolic Equations, Numer. Math. **67**(1994), pp. 93-129.

D. Gottlieb, *The Stability of Pseudospectral Chebyshev Methods*, Math. Comp. **36**(1981), pp. 107-118.

D. Gottlieb, L. Lustman, and S. A. Orszag, *Spectral Calculations of One-Dimensional Inviscid Compressible Flows*, SIAM J. Sci. Comp. **2**(1981), pp. 296-310.

D. Gottlieb, L. Lustman, and E. Tadmor, *Stability Analysis of Spectral Methods for Hyperbolic Initial-Boundary Value Systems*, SIAM J. Numer. Anal. **24**(1987), pp. 241-256.

D. Gottlieb, L. Lustman, and E. Tadmor, *Convergence of Spectral Methods for Hyperbolic Initial-Boundary Value Systems*, SIAM J. Numer. Anal. **24**(1987), pp. 532-537.

D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*. CBMS-NSF **26**. SIAM, Philadelphia, 1977.

D. Gottlieb, S. A. Orszag, and E. Turkel, *Stability of Pseudospectral and Finite-Difference Methods for Variable Coefficient Problems*, Math. Comp. **37**(1981), pp. 293-305.

D. Gottlieb and C.W. Shu, *On the Gibbs Phenomenon V: Recovering Exponential Accuracy from Collocation Point Values of a Piecewise Analytic Function*, Numer. Math. **71**(1995), pp. 511-526.

D. Gottlieb and C.W. Shu, *On the Gibbs Phenomenon and its Resolution*, SIAM Review **39**(1997), pp. 644-668.

D. Gottlieb and C.W. Shu, *A General Theory for the Resolution of the Gibbs Phenomenon*. In Tricomi's Ideas and Contemporary Applied Mathematics, National Italian Academy of Science, 1997.

D. Gottlieb and E. Tadmor, *The CFL Condition for Spectral Approximations to Hyperbolic Initial-Value Problems*, Math. Comp. **56**(1991), pp. 565-588.

D. Gottlieb and C. E. Wasberg, *Optimal Strategy in Domain Decomposition Spectral Methods for Wave-Like Phenomena*, SIAM J. Sci. Comput. **??**(1999), pp. ??-??.

J. S. Hesthaven, *A Stable Penalty Method for the Compressible Navier-Stokes Equations: II. One-Dimensional Domain Decomposition Schemes*, SIAM J. Sci. Comput. **18**(1997), pp. 658-685.

J. S. Hesthaven, *A Stable Penalty Method for the Compressible Navier-Stokes Equations: III. Multidimensional Domain Decomposition Schemes*, SIAM J. Sci. Comput. **20**(1999), pp. 62-93.

J. S. Hesthaven, *Spectral Penalty Methods*, Appl. Numer. Math. 2000 - to appear.

J. S. Hesthaven, P. G. Dinesen, and J. P. Lynov, *Spectral Collocation Time-Domain Modeling of Diffractive Optical Elements*, J. Comput. Phys. **155**(1999), pp. 287-306.

J. S. Hesthaven and D. Gottlieb, *A Stable Penalty Method for the Compressible Navier-Stokes Equations. I. Open Boundary Conditions*, SIAM J. Sci. Comp. **17**(1996), 579-612.

J. S. Hesthaven and D. Gottlieb, *Stable Spectral Methods for Conservation Laws on Triangles with Unstructured Grids*, Comput. Methods Appl. Mech. Engin. **175**(1999), pp. 361-381.

J. S. Hesthaven, J. Juul Rasmussen, L. Bergé and J. Wyller, *Numerical studies of localized wave fields govenered by the Raman-extended*

derivative nonlinear Schrödinger equation, J. Phys. A: Math. Gen. **30**(1997), pp. 8207-8224.

J. S. Hesthaven and C. H. Teng, *Stable Spectral Methods on Tetrahedral Elements*, SIAM J. Sci. Comput. 2000 – to appear.

E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*. Dover Publishing Inc, New York. 1966.

D. Jackson, *The Theory of Approximation*. American Mathematical Society, Colloquim Publication **11**. Providence. 1930.

D. A. Kopriva, *A Spectral Multidomain Method for the Solution of Hyperbolic Systems*, Appl. Numer. Math. **2**(1986), pp. 221-241.

D. A. Kopriva, *Computation of Hyperbolic Equations on Complicated Domains with Patched and Overset Chebyshev Grids*, SIAM J. Sci. Stat. Comput. **10**(1989), pp. 120-132.

D. A. Kopriva, *Multidomain Spectral Solution of the Euler Gas-Dynamics Equations*, J. Comput. Phys. **96**(1991), pp. 428-450.

D. A. Kopriva, *A Conservative Staggered-Grid Chebyshev Multidomain Method for Compressible Flows. II. A Semi-Structured Method*, J. Comput. Phys. **128**(1996), pp. 475-488.

D. A. Kopriva and J. H. Kolias, *A Conservative Staggered-Grid Chebyshev Multidomain Method for Compressible Flows. II. A Semi-Structured Method*, J. Comput. Phys. **125**(1996), pp. 244-261.

D. A. Kopriva, S.'L. Woodruff, and M. Y. Hussaini, *Discontinuous Spectral Element Approximation of Maxwell's Equations*. In Proc. of First International Symposium on Discontinuous Galerkin Methods. B. Cockburn, G. E. Karniadakis, and C.W. Shu (Eds). Newport, RI, 1999.

H. O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*. Series in Pure and Applied Mathematics, Academic Press, San Diego. 1989.

H. O. Kreiss and J. Oliger, *Comparison of Accurate Methods for the Integration of Hyperbolic Problems*, Tellus **24**(1972), pp. 199-215.

H.O. Kreiss and J. Oliger, *Stability of the Fourier Method*, SIAM J. Numer. Anal. **16**(1979), pp. 421-433.

C. Lanczos, *Applied Analysis*. Pitman, London, 1957.

P.D. Lax, *Accuracy and Resolution in the Computation of Solutions of Linear and Nonlinear Equations*. In Proc. of Recent Advances in Numerical Analysis, Univ. Wisconsin, Academic Press, 1978. pp. 107-117.

D. Levy and E. Tadmor, *From Semi-Discrete to Fully-Discrete: Stability of Runge-Kutta Schemes by the Energy Method*, SIAM Review **40**(1998), pp. 40-73.

I. Lomtev, C. B. Quillen and G. E. Karniadakis, *Spectral/hp Methods for Viscous Compressible Flows on Unstructured 2D Meshes*, J. Comput. Phys. **144**(1998), pp. 325-357.

A. Majda, J. McDonough, and S. Osher, *The Fourier Method for Nonsmooth Initial Data*, Math. Comp. **32**(1978), pp. 1041-1081.

S. A. Orszag, *Comparison of Pseudospectral and Spectral Approximation*, Stud. Appl. Math. **51**(1972), pp. 253-259.

J. E. Pasciak, *Spectral and Pseudospectral Methods for Advection Equations*, Math. Comp. **35**(1980), pp. 1081-1092.

A. Quarteroni, *Domain Decomposition Methods for Systems of Conservation Laws: Spectral Collocation Approximations*, SIAM J. Sci. Stat. Comput. **11**(1990), pp. 1029-1052.

S. C. Reddy and L. N. Trefethen, *Lax-Stability of Fully Discrete Spectral Methods via Stability Regions and Pseudo-Eigenvalues*, Comput. Methods Appl. Mech. Engin. **80**(1990), pp. 147-164.

S, C. Reddy and L. N. Trefethen, *Stability of the Method of Lines*, Numer. Math. **62**(1992), pp. 235-267.

D. Sidilkover and G. E. Karniadakis, *Non-Oscillatory Spectral Element Chebyshev Method for Shock Wave Calculations*, J. Comput. Phys. **107**(1993), pp. 10-22.

G. Szegö, *Orthogonal Polynomials*. 4th Ed. American Mathematical Society, Colloquim Publication **23**. Providence. 1975.

E. Tadmor, *Skew-Selfadjoint Form for Systems of Conservation Laws*, J. Math. Anal. App. **103**(1984), pp. 428-442.

E. Tadmor, *The Exponential Accuracy of Fourier and Chebyshev Differencing Methods*, SIAM Review **23**(1986), pp. 1-10.

E. Tadmor, *Stability Analysis of Finite-Difference, Pseudospectral, and Fourier-Galerkin Approximations for Time-Dependent Problems*, SIAM Review **29**(1987), pp. 525-555.

H. Tal-Ezer, Ph.D. Thesis, Tel Aviv University, 1983.

L. N. Trefethen and M. R. Trummer, *An Instability Phenomenon in Spectral Methods*, SIAM J. Numer. Anal. **24**(1987), pp. 1008-1023.

H. Vandeven, *Family of Spectral Filters for Discontinuous Problems*, J. Sci. Comput. **8**(1991), pp. 159-192.

L. Vozovoi, M. Israeli, and A. Averbuch, *Analysis and Application of Fourier-Gegenbauer Method to Stiff DIfferential Equations*, SIAM J. Numer. Anal. **33**(1996), pp. 1844-1863.

L. Vozovoi, A. Weill and M. Israeli, *Spectrally Accurate Solution of Nonperiodic Differential Equations by the Fourier-Gegenbauer Method*, SIAM J. Numer. Anal. **34**(1997), pp. 1451-1471.

T. Warburton and G. E. Karniadakis, *A Discontinuous Galerkin Method for the Viscous MHD Equations*, J. Comput. Phys, **152**(1999), pp. 608-641.

T. Warburton, *Application of the Discontinuous Galerkin Method to Maxwell's Equations Using Unstructured Polymorphic hp-finite Elements.* In Proceedings of the International Symposium on Discontinuous Galerkin Methods. Newport, RI, 1999.

T. Warburton, I. Lomtev, Y. Du, S. Sherwin, and G. E. Karniadakis *Galerkin and Discontinuous Galerkin Spectral/hp Methods*, Comput. Methods Appl. Mech. Engrg. **175**(1999), pp. 343-359.

B. Yang, D. Gottlieb, and J. S. Hesthaven, *Spectral Simulations of Electromagnetic Wave Scattering*, J. Comput. Phys. **134**(1997), pp. 216-230.

B. Yang and J. S. Hesthaven, *A Pseudospectral Method for Time-Domain Computation of Electromagnetic Scattering by Bodies of Revolution*, IEEE Trans. Antennas Propaga. **47**(1999), pp. 132-141.

D. Funaro, *Computational Aspects of Pseudospectral Laguerre Approximations*, Appl. Numer. Math. **6**(1990), pp. 447-457.

Y. Maday, B. Pernaud-Thomas, and H. Vandeven, *Reappraisal of Laguerre Type Spectral Methods*, Resh. Aerosp. **6**(1985), pp. 13-35.

P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*. Computer Science and Applied Mathematics. Academic Press, New York. 1975.

C. Mavriplis, *Laguerre Polynomials for Infinite Domain Spectral Methods*,

J. Comput. Phys. **80**(1989), pp. 480-488.

J. P. Boyd, *The Rate of Convergence of Hermite Function Series*, Math. Comp. **35**(1980), pp. 1309-1316.

J. P. Boyd, *Chebyshev and Fourier Spectral Methods, 2nd Edition*. Dover Publishers, New York. 2000.