Chapter 1
Introduction

Numerical methods for partial differential equations can be classified into the *local* and *global* categories. The finite-difference and finite-element methods are based on local arguments, whereas the spectral method is global in character. In practice, finite-element methods are particularly well suited to problems in complex geometries, whereas spectral methods can provide superior accuracy, at the expense of domain flexibility. We emphasize that there are many numerical approaches, such as *hp* finite-elements and spectral-elements, which combine advantages of both the global and local methods. However in this book, we shall restrict our attentions to the *global* spectral methods.

Spectral methods, in the context of numerical schemes for differential equations, belong to the family of weighted residual methods (WRMs), which are traditionally regarded as the foundation of many numerical methods such as finite element, spectral, finite volume, boundary element (cf. Finlayson (1972)). WRMs represent a particular group of approximation techniques, in which the residuals (or errors) are minimized in a certain way and thereby leading to specific methods including Galerkin, Petrov-Galerkin, collocation and tau formulations.

The objective of this introductory chapter is to formulate spectral methods in a general way by using the notion of residual. Several important tools, such as *discrete transform* and *spectral differentiation*, will be introduced. These are basic ingredients for developing efficient spectral algorithms.

1.1 Weighted Residual Methods

Prior to introducing spectral methods, we first give a brief introduction to the WRM. Consider the general problem:

$$\partial_t u(x,t) - \mathscr{L}u(x,t) = \mathscr{N}(u)(x,t), \quad t > 0,\ x \in \Omega, \tag{1.1}$$

J. Shen et al., *Spectral Methods: Algorithms, Analysis and Applications*,
Springer Series in Computational Mathematics 41, DOI 10.1007/978-3-540-71041-7_1,

where $\mathscr{L}$ is a leading spatial derivative operator, and $\mathscr{N}$ is a lower-order linear or nonlinear operator involving only spatial derivatives. Here, Ω denotes a bounded domain of $\mathbb{R}^d$, $d = 1, 2$ or 3. Equation (1.1) is to be supplemented with an initial condition and suitable boundary conditions.

We shall only consider the WRM for the spatial discretization, and assume that the time derivative is discretized with a suitable time-stepping scheme. Among various time-stepping methods (cf. Appendix D), semi-implicit schemes or linearly-implicit schemes, in which the principal linear operators are treated *implicitly* to reduce the associated stability constraint, while the nonlinear terms are treated explicitly to avoid the expensive process of solving nonlinear equations at each time step, are most frequently used in the context of spectral methods.

Let τ be the time step size, and $u^k(\cdot)$ be an approximation of $u(\cdot, k\tau)$. As an example, we consider the Crank-Nicolson leap-frog scheme for (1.1):

$$\frac{u^{n+1} - u^{n-1}}{2\tau} - \mathscr{L}\Big(\frac{u^{n+1} + u^{n-1}}{2}\Big) = \mathscr{N}(u^n), \quad n \geq 1. \tag{1.2}$$

We can rewrite (1.2) as

$$\mathbf{L}u(x) := \alpha u(x) - \mathscr{L}u(x) = f(x), \quad x \in \Omega, \tag{1.3}$$

where, with a slight abuse of notation, $u = \frac{u^{n+1}+u^{n-1}}{2}$, $\alpha = \tau^{-1}$ and $f = \alpha u^{n-1} + \mathscr{N}(u^n)$. Hence, at each time step, we need to solve a steady-state problem of the form (1.3).

At this point, it is important to emphasize that the construction of efficient numerical solvers for some important equations in the form of (1.3), such as Poisson-type equations and advection-diffusion equations, is an essential step in solving general nonlinear PDEs. With this in mind, a particular emphasis of this book is to design and analyze efficient spectral algorithms for equations of the form (1.3) where $\mathscr{L}$ is a *linear elliptic* operator.

The starting point of the WRM is to approximate the solution u of (1.3) by a finite sum

$$u(x) \approx u_N(x) = \sum_{k=0}^{N} a_k \phi_k(x), \tag{1.4}$$

where $\{\phi_k\}$ are the *trial* (*or basis*) *functions*, and the expansion coefficients $\{a_k\}$ are to be determined. Substituting u_N for u in (1.3) leads to the *residual*:

$$\mathbf{R}_N(x) = \mathbf{L}u_N(x) - f(x) \neq 0, \quad x \in \Omega. \tag{1.5}$$

The notion of the WRM is to force the residual to zero by requiring

$$(\mathbf{R}_N, \psi_j)_\omega := \int_\Omega \mathbf{R}_N(x)\psi_j(x)\omega(x)dx = 0, \quad 0 \leq j \leq N, \tag{1.6}$$

where $\{\psi_j\}$ are the *test functions*, and ω is a positive weight function; or

$$\langle \mathbf{R}_N, \psi_j \rangle_{N,\omega} := \sum_{k=0}^{N} \mathbf{R}_N(x_k)\psi_j(x_k)\omega_k = 0, \quad 0 \le j \le N, \tag{1.7}$$

where $\{x_k\}_{k=0}^N$ are a set of preselected collocation points, and $\{\omega_k\}_{k=0}^N$ are the weights of a numerical quadrature formula.

The choice of trial/test functions is one of the main features that distinguishes spectral methods from finite-element and finite-difference methods. In the latter two methods, the trial/test functions are local in character with finite regularities. In contrast, spectral methods employ globally smooth functions as trial/test functions. The most commonly used trial/test functions are trigonometric functions or orthogonal polynomials (typically, the eigenfunctions of singular Sturm-Liouville problems), which include

- $\phi_k(x) = e^{ikx}$ (Fourier spectral method)
- $\phi_k(x) = T_k(x)$ (Chebyshev spectral method)
- $\phi_k(x) = L_k(x)$ (Legendre spectral method)
- $\phi_k(x) = \mathscr{L}_k(x)$ (Laguerre spectral method)
- $\phi_k(x) = H_k(x)$ (Hermite spectral method)

Here, $T_k, L_k, \mathscr{L}_k$ and H_k are the Chebyshev, Legendre, Laguerre and Hermite polynomials of degree k, respectively.

The choice of test functions distinguishes the following formulations:

- *Galerkin.* The test functions are the same as the trial ones (i.e., $\phi_k = \psi_k$ in (1.6) or (1.7)), assuming the boundary conditions are periodic or homogeneous.
- *Petrov-Galerkin.* The test functions are different from the trial ones.
- *Collocation.* The test functions $\{\psi_k\}$ in (1.7) are the Lagrange basis polynomials such that $\psi_k(x_j) = \delta_{jk}$, where $\{x_j\}$ are preassigned collocation points. Hence, the residual is forced to zero at $\{x_j\}$, i.e., $\mathbf{R}_N(x_j) = 0$.

Remark 1.1. *In the literature, the term of pseudo-spectral method is often used to describe any spectral method where some operations involve a collocation approach or a numerical quadrature which produces aliasing errors* (cf. Gottlieb and Orszag (1977)). *In this sense, almost all practical spectral methods are pseudo-spectral. In this book, we shall not classify a method as pseudo-spectral or spectral. Instead, it will be classified as Galerkin type or collocation type.*

Remark 1.2. *The so-called* tau *method is a particular class of Petrov-Galerkin method. While the* tau *method offers some advantages in certain situations, for most problems, it is usually better to use a well-designed Galerkin or Petrov-Galerkin method. So in this book, we shall not touch on this topic, and refer to* El-Daou and Ortiz (1998), Canuto et al. (2006) *and the references therein for a thorough discussion of this approach.*

In the forthcoming sections, we shall demonstrate how to construct spectral methods for solving differential equations by examining several spectral schemes based on Galerkin, Petrov-Galerkin and collocation formulations in a general manner. We shall revisit these illustrative examples in a more rigorous fashion in the main body of the book.

1.2 Spectral-Collocation Method

To fix the idea, we consider the following linear problem:

$$\begin{aligned} &\mathbf{L}u(x) = -u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \quad x \in (-1,1), \\ &B_{\pm}u(\pm 1) = g_{\pm}, \end{aligned} \tag{1.8}$$

where $B_{\pm}$ are linear operators corresponding to Dirichlet, Neumann or Robin boundary conditions (see Sect. 4.1), and the data p, q, f and $g_{\pm}$ are given such that the above problem is well-posed.

As mentioned earlier, the collocation method forces the residual to vanish pointwisely at a set of preassigned points. More precisely, let $\{x_j\}_{j=0}^{N}$ (with $x_0 = -1$ and $x_N = 1$) be a set of Gauss-Lobatto points (see Chap. 3), and let P_N be the set of all real algebraic polynomials of degree $\leq N$. The spectral-collocation method for (1.8) amounts to finding $u_N \in P_N$ such that (a) the residual $\mathbf{R}_N(x) = \mathbf{L}u_N(x) - f(x)$ equals to zero at the interior collocation points, namely,

$$\mathbf{R}_N(x_k) = \mathbf{L}u_N(x_k) - f(x_k) = 0, \quad 1 \leq k \leq N-1, \tag{1.9}$$

(b) u_N satisfies exactly the boundary conditions, i.e.,

$$B_{-}u_N(x_0) = g_{-}, \quad B_{+}u_N(x_N) = g_{+}. \tag{1.10}$$

The spectral-collocation method is usually implemented in the physical space by seeking approximate solution in the form

$$u_N(x) = \sum_{j=0}^{N} u_N(x_j) h_j(x), \tag{1.11}$$

where $\{h_j\}$ are the Lagrange basis polynomials (also referred to as *nodal* basis functions), i.e., $h_j \in P_N$ and $h_j(x_k) = \delta_{kj}$. Hence, inserting (1.11) into (1.9)-(1.10) leads to the linear system

$$\begin{aligned} &\sum_{j=0}^{N} \big[\mathbf{L}h_j(x_k)\big] u_N(x_j) = f(x_k), \quad 1 \leq k \leq N-1, \\ &\sum_{j=0}^{N} \big[B_{-}h_j(x_0)\big] u_N(x_j) = g_{-}, \quad \sum_{j=0}^{N} \big[B_{+}h_j(x_N)\big] u_N(x_j) = g_{+}. \end{aligned} \tag{1.12}$$

The above system contains $N+1$ equations and $N+1$ unknowns, so we can rewrite it in a matrix form. To fix the idea, we consider (1.8) with Dirichlet boundary conditions: $u(\pm 1) = g_{\pm}$. In this case, setting $u_N(x_0) = g_{-}$ and $u_N(x_N) = g_{+}$ in the first equation of (1.12), we find that the system (1.12) reduces to

$$\sum_{j=1}^{N-1} \big[\mathbf{L}h_j(x_k)\big] u_N(x_j) = f(x_k) - \big\{ \big[\mathbf{L}h_0(x_k)\big] g_- + \big[\mathbf{L}h_N(x_k)\big] g_+ \big\}, \tag{1.13}$$

for $1 \le k \le N-1$. Differentiating (1.11) m times leads to

$$u_N^{(m)}(x_k) = \sum_{j=0}^{N} d_{kj}^{(m)} u_N(x_j) \ \text{ where } \ d_{kj}^{(m)} = h_j^{(m)}(x_k). \tag{1.14}$$

The matrix $D^{(m)} = \big(d_{kj}^{(m)}\big)_{k,j=0,\ldots,N}$ is called the differentiation matrix of order m relative to $\{x_j\}_{j=0}^N$. If we denote by $\mathbf{u}^{(m)}$ the vector whose components are the values of $u_N^{(m)}$ at the collocation points, it follows from (1.14) that

$$\mathbf{u}^{(m)} = D^{(m)} \mathbf{u}^{(0)}, \quad m \ge 1. \tag{1.15}$$

Hence, we have

$$\mathbf{L}h_j(x_k) = -d_{kj}^{(2)} + p(x_k) d_{kj}^{(1)} + q(x_k)\delta_{kj}. \tag{1.16}$$

Denote by $\mathbf{f}$ the vector with $N-1$ components given by the right-hand side of (1.13). Setting

$$\begin{aligned} &\widetilde{D}_m = \big(d_{kj}^{(m)}\big)_{k,j=1,\ldots,N-1}, \quad m = 1, 2, \\ &P = \operatorname{diag}\big(p(x_1), \ldots, p(x_{N-1})\big), \quad Q = \operatorname{diag}\big(q(x_1), \ldots, q(x_{N-1})\big), \end{aligned} \tag{1.17}$$

the system (1.13) reduces to

$$\big(-\widetilde{D}_2 + P\widetilde{D}_1 + Q\big)\mathbf{u}^{(0)} = \mathbf{f}. \tag{1.18}$$

Observe that the collocation method is easy to implement, once the differentiation matrices are precomputed. Moreover, it is very convenient for solving problems with variable coefficients and/or nonlinear problems, since we work in the physical space and derivatives can be evaluated by (1.14) directly. As a result, the collocation method has been extensively used in practice. However, three important issues should be considered in the implementation and analysis of a collocation method:

- The coefficient matrix of the collocation system is always full with a condition number behaving like $O(N^{2m})$ (m is the order of the differential equation).
- The choice of collocation points is crucial in terms of stability, accuracy and ease of dealing with boundary conditions. In general, they are chosen as nodes (typically, zeros of orthogonal polynomials) of Gauss-type quadrature formulas.
- The aforementioned collocation scheme is formulated in a *strong* form. In terms of error analysis, it is more convenient to reformulate it as a (but not always equivalent) *weak* form, see Sect. 1.3.3 and Chap. 4.

1.3 Spectral Methods of Galerkin Type

The collocation method described in the previous section is implemented in the physical space. In this section, we shall describe Galerkin-type spectral methods in the frequency space, and present the basic principles of the spectral-Galerkin method, spectral-Petrov-Galerkin method, and spectral-Galerkin method with numerical integration.

1.3.1 Galerkin Method

Without loss of generality, we consider (1.8) with $g_{\pm} = 0$. The non-homogeneous boundary conditions can be easily handled by considering $v = u - \tilde{u}$, where $\tilde{u}$ is a "simple" function satisfying the non-homogeneous boundary conditions (cf. Chap. 4).

Define the finite-dimensional approximation space:

$$X_N = \left\{ \phi \in P_N \ : \ B_{\pm}\phi(\pm 1) = 0 \right\} \ \Rightarrow \ \dim(X_N) = N - 1.$$

Let $\{\phi_k\}_{k=0}^{N-2}$ be a set of basis functions of X_N. We expand the approximate solution as

$$u_N(x) = \sum_{k=0}^{N-2} \hat{u}_k \phi_k(x) \in X_N. \tag{1.19}$$

Then, the expansion coefficients $\{\hat{u}_k\}_{k=0}^{N-2}$ can be determined by the residual equation (1.6) with $\{\psi_j = \phi_j\}$:

$$\int_{-1}^{1} \left(\mathbf{L}u_N(x) - f(x) \right) \phi_j(x) \omega(x) dx = 0, \quad 0 \le j \le N-2, \tag{1.20}$$

which is equivalent to

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ \left(\mathbf{L}u_N, v_N \right)_{\omega} = \left(f, v_N \right)_{\omega}, \quad \forall v_N \in X_N. \end{cases} \tag{1.21}$$

Here, $(\cdot,\cdot)_{\omega}$ is the inner product of $L^2_{\omega}(-1,1)$ (cf. Appendix B).

The linear system of the above scheme is obtained by substituting (1.19) into (1.20). More precisely, setting

$$\begin{aligned} &\mathbf{u} = \left(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-2} \right)^T; \quad f_j = (f, \phi_j)_{\omega}, \quad \mathbf{f} = \left(f_0, f_1, \ldots, f_{N-2} \right)^T; \\ &s_{jk} = \left(\mathbf{L}\phi_k, \phi_j \right)_{\omega}, \quad S = (s_{jk})_{j,k=0,\ldots,N-2}, \end{aligned}$$

the system (1.20) reduces to

$$S\mathbf{u} = \mathbf{f}. \tag{1.22}$$

Therefore, it is crucial to choose basis functions $\{\phi_j\}$ such that:

- The right-hand side $(f, \phi_j)_\omega$ can be computed efficiently.
- The linear system (1.22) can be solved efficiently.

The key idea is to use *compact combinations* of orthogonal polynomials or orthogonal functions to construct basis functions. To demonstrate the basic principle, we consider the Legendre spectral approximation (i.e., $\omega \equiv 1$ in (1.20)-(1.22)). Let $L_k(x)$ be the Legendre polynomial of degree k, and set

$$\phi_k(x) = L_k(x) + \alpha_k L_{k+1}(x) + \beta_k L_{k+2}(x), \quad k \geq 0, \tag{1.23}$$

where the constants α_k and β_k are uniquely determined by the boundary conditions: $B_\pm \phi_k(\pm 1) = 0$ (cf. Sect. 4.1). We shall refer to such basis functions as *modal* basis functions. Therefore, we have

$$X_N = \text{span}\{\phi_0, \phi_1, \ldots, \phi_{N-2}\}. \tag{1.24}$$

Using the properties of Legendre polynomials (cf. Sect. 3.3), one verifies easily that, *if $p(x)$ and $q(x)$ are constants*, the coefficient matrix S is *sparse* so the linear system (1.22) can be solved efficiently. However, for more general $p(x)$ and $q(x)$, the coefficient matrix S is full and one needs to resort to an iterative method (cf. Sect. 4.4).

In the above, we just considered the Legendre case. In fact, the construction of such a basis is also feasible for the Chebyshev, Laguerre and Hermite cases (see Chaps. 4–7). The notion of using compact combinations of orthogonal polynomials/functions to develop efficient spectral solvers will be repeatedly emphasized in this book.

We now consider the evaluation of $(f, \phi_j)_\omega$. In general, this term can not be computed exactly and is usually approximated by $(I_N f, \phi_j)_\omega$, where I_N is an interpolation operator upon P_N relative to the Gauss-Lobatto points. Thus, we can write

$$(I_N f)(x) = \sum_{k=0}^{N} \tilde{f}_k \varphi_k(x), \tag{1.25}$$

where $\{\varphi_k\}$ is an orthonormal polynomial basis of P_N (orthogonal with respect to ω, i.e., $(\varphi_k, \varphi_j)_\omega = \delta_{jk}$). Thanks to the orthogonality, the *discrete transforms* between the physical values $\{f(x_j)\}_{j=0}^N$ and the expansion coefficients $\{\tilde{f}_k\}_{k=0}^N$ can be computed efficiently. In particular, the computational complexity of the Fourier and Chebyshev discrete transforms can be reduced to $O(N \log_2 N)$ by using the fast Fourier transform (FFT). An approach for implementing discrete transforms relative to general orthogonal polynomials is given in Sect. 3.1.5.

It is important to point out that in solving time-dependent nonlinear problems, f usually contains nonlinear terms involving derivatives of the numerical solution u_N at previous time steps (cf. (1.3)). Hence, numerical differentiations in the frequency space and/or in the physical space are required. Differentiation techniques relative to general orthogonal polynomials are addressed in Sects. 3.1.6 and 3.1.7.

1.3.2 Petrov-Galerkin Method

As pointed out in Sect. 1.1, the use of different test and trial functions distinguishes the Petrov-Galerkin method from the Galerkin method. Thanks to this flexibility, the Petrov-Galerkin method can be very useful for some non-self-adjoint problems such as odd-order equations.

As an illustrative example, we consider the following third-order equation:

$$\begin{aligned} &\mathbf{L}u(x) := u'''(x) + u(x) = f(x), \quad x \in (-1,1), \\ &u(\pm 1) = u'(1) = 0. \end{aligned} \tag{1.26}$$

As with the Galerkin case, we enforce the boundary conditions on the approximate solution. So we set

$$X_N = \left\{ \phi \in P_N : \phi(\pm 1) = \phi'(1) = 0 \right\} \;\Rightarrow\; \dim(X_N) = N-2.$$

Assuming that $\{\phi_k\}_{k=0}^{N-3}$ is a basis of X_N, we expand the approximate solution as

$$u_N(x) = \sum_{k=0}^{N-3} \hat{u}_k \phi_k(x) \in X_N.$$

The expansion coefficients $\{\hat{u}_k\}_{k=0}^{N-3}$ are determined by the residual equation (1.6) (with $\omega = 1$):

$$\int_{-1}^{1} \big(\mathbf{L}u_N(x) - f(x)\big)\psi_j(x)dx = 0, \quad 0 \le j \le N-3. \tag{1.27}$$

Since the leading third-order operator is not self-adjoint, it is natural to use a Petrov-Galerkin method with the test function space:

$$X_N^* = \left\{ \psi \in P_N : \psi(\pm 1) = \psi'(-1) = 0 \right\} \;\Rightarrow\; \dim(X_N^*) = N-2.$$

Assume that $\{\psi_k\}_{k=0}^{N-3}$ is a basis of X_N^*. Then, (1.27) is equivalent to the variational formulation:

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ \big(\mathbf{L}u_N, v_N\big) = \big(f, v_N\big), \quad \forall v_N \in X_N^*, \end{cases} \tag{1.28}$$

where $(\cdot,\cdot)$ is the inner product of the usual L^2-space.

The theoretical aspects of the above scheme will be examined in Chap. 6. We now consider its implementation. Setting

$$\begin{aligned} &\mathbf{u} = \big(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-3}\big)^T; \quad f_j = (f, \psi_j), \quad \mathbf{f} = \big(f_0, f_1, \ldots, f_{N-3}\big)^T; \\ &s_{jk} = (\phi_k', \psi_j''), \quad S = \big(s_{jk}\big)_{j,k=0,\ldots,N-3}; \\ &m_{jk} = (\phi_k, \psi_j), \quad M = \big(m_{jk}\big)_{j,k=0,\ldots,N-3}, \end{aligned}$$

the linear system (1.28) becomes

$$\big(S+M\big)\mathbf{u}=\mathbf{f}. \tag{1.29}$$

As described in the previous section, we wish to construct basis functions for X_N and X_N^*, so that the linear system (1.29) can be inverted efficiently. Once again, this goal can be achieved by using compact combinations of orthogonal polynomials. It can be checked that for $0 \le k \le N-3$,

$$\begin{aligned}\phi_k &= L_k - \frac{2k+3}{2k+5}L_{k+1} - L_{k+2} + \frac{2k+3}{2k+5}L_{k+3} \in X_N;\\ \psi_k &= L_k + \frac{2k+3}{2k+5}L_{k+1} - L_{k+2} - \frac{2k+3}{2k+5}L_{k+3} \in X_N^*,\end{aligned} \tag{1.30}$$

where L_n is the Legendre polynomial of degree n (cf. Sect. 3.3). Hence, $\{\phi_k\}_{k=0}^{N-3}$ (resp. $\{\psi_j\}_{j=0}^{N-3}$) forms a basis of X_N (resp. X_N^*). Moreover, using the properties of the Legendre polynomials, one verifies easily that the matrix M is seven-diagonal, i.e., $m_{jk}=0$ for all $|j-k|>3$. More importantly, the matrix S is diagonal.

1.3.3 Galerkin Method with Numerical Integration

We considered previously Galerkin-type methods in the frequency space, which are well suited for linear problems with constant (or polynomial) coefficients. However, their implementations are not convenient for problems with general variable coefficients. On the other hand, the collocation method is easy to implement, but it can not always be reformulated as a suitable variational formulation (most convenient for error analysis). A combination of these two approaches leads to the so-called *Galerkin method with numerical integration*, or sometimes called the *collocation method in the weak form.*

The key idea of this approach is to *replace the continuous inner products in the Galerkin formulation by the discrete ones*. As an example, we consider again (1.8) with $g_\pm = 0$. The spectral-Galerkin method with numerical integration is

$$\begin{cases}\text{Find } u_N \in X_N := \{\phi \in P_N : B_\pm\phi(\pm 1) = 0\} \text{ such that}\\ a_N(u_N, v_N) := \langle Lu_N, v_N\rangle_N = \langle f, v_N\rangle_N, \quad \forall v_N \in X_N,\end{cases} \tag{1.31}$$

where the discrete inner product is defined by

$$\langle u, v\rangle_N = \sum_{j=0}^{N} u(x_j)v(x_j)\omega_j,$$

with $\{x_j, \omega_j\}_{j=0}^N$ being the set of Legendre-Gauss-Lobatto quadrature nodes and weights (cf. Theorem 3.29).

For problems with variable coefficients, the above method is easier to implement, thanks to the discrete inner product, than the spectral-Galerkin method (1.21). It is also more convenient for error analysis, thanks to the weak formulation, than the spectral-collocation method (1.12).

We note that in the particular case of homogeneous Dirichlet boundary conditions, i.e., $B_{\pm}u(\pm 1) = u(\pm 1) = 0$, by taking $v_N = h_j$, $1 \le j \le N-1$ in (1.31) and using the exactness of Legendre-Gauss-Lobatto quadrature, i.e.,

$$\langle u, v\rangle_N = (u, v), \quad \forall u \cdot v \in P_{2N-1}, \tag{1.32}$$

we find that the formulation (1.31) is equivalent to the collocation formulation (1.12). However, this is not true for general boundary conditions (see Chap. 4).

1.4 Fundamental Tools for Error Analysis

In the previous sections, we briefly described several families of spatial discretization schemes using the notion of weighted residual methods. In this section, we present some fundamental apparatuses for stability and convergence analysis of numerical schemes based on weak (or variational) formulations.

We consider the linear boundary value problem (1.3):

$$\mathbf{L}u = f, \quad \text{in } \Omega; \quad Bu = 0, \quad \text{on } \partial\Omega, \tag{1.33}$$

where $\mathbf{L}$ and B are linear operators, and f is a given function on Ω.

As shown before, the starting point is to reformulate (1.33) in a *weak formulation*:

$$\begin{cases} \text{Find } u \in X \text{ such that} \\ a(u, v) = F(v), \quad \forall v \in Y, \end{cases} \tag{1.34}$$

where X is the space of trial functions, Y is the space of test functions, and F is a linear functional on Y. The expression $a(u, v)$ defines a bilinear form on $X \times Y$. It is conventional to assume that X and Y are Hilbert spaces. We refer to Appendix B for basic functional analysis settings.

Now, we consider what conditions should be placed on (1.34) to guarantee its well-posedness in the sense that:

- *Existence-uniqueness:* There exists exactly one solution of the problem.
- *Stability:* The solution must be stable which means that it depends on the data continuously. In other words, a small change of the given data produces a small change of the solution correspondingly.

The first fundamental result concerning the existence-uniqueness and stability is known as the Lax-Milgram lemma (see Theorem B.1) related to the abstract problem (1.34) with $X = Y$, i.e.,

$$\begin{cases} \text{Find } u \in X \text{ such that} \\ a(u,v) = F(v), \quad \forall v \in X. \end{cases} \tag{1.35}$$

More precisely, if the bilinear form $a(\cdot,\cdot) : X \times X \to \mathbb{R}$ satisfies

- Continuity:
$$\exists C > 0 \quad \text{such that} \quad |a(u,v)| \le C\|u\|_X \|v\|_X, \tag{1.36}$$
- Coercivity:
$$\exists \alpha > 0 \quad \text{such that} \quad a(u,u) \ge \alpha \|u\|_X^2, \tag{1.37}$$

then for any $F \in X'$ (the dual space of X as defined in Appendix B), the problem (1.35) admits a unique solution $u \in X$, satisfying

$$\|u\|_X \le \frac{1}{\alpha}\|F\|_{X'}. \tag{1.38}$$

Remark 1.3. *The constant*

$$\alpha = \inf_{0 \ne u \in X} \frac{|a(u,u)|}{\|u\|_X^2} \tag{1.39}$$

is referred to as the ellipticity constant of (1.35)*.*

The above result can only be applied to the problem (1.34) with $Y = X$. We now present a generalization of the Lax-Milgram lemma for the case $X \ne Y$ (see, e.g., Babuška and Aziz (1972)).

Theorem 1.1. *Let X and Y be two real Hilbert spaces, equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. Assume that $a(\cdot,\cdot) : X \times Y \to \mathbb{R}$ is a bilinear form and $F(\cdot) : Y \to \mathbb{R}$ is a linear continuous functional, i.e., $F \in Y'$ (the dual space of Y) satisfying*

$$\|F\|_{Y'} = \sup_{0 \ne v \in Y} \frac{|F(v)|}{\|v\|_Y} < \infty. \tag{1.40}$$

Further, assume that $a(\cdot,\cdot)$ satisfies

- *Continuity*:
$$\exists C > 0 \quad \textit{such that} \quad |a(u,v)| \le C\|u\|_X \|v\|_Y, \tag{1.41}$$
- *Inf-sup condition*:
$$\exists \beta > 0 \quad \textit{such that} \quad \sup_{0 \ne v \in Y} \frac{|a(u,v)|}{\|u\|_X \|v\|_Y} \ge \beta, \quad \forall 0 \ne u \in X, \tag{1.42}$$
- *"Transposed" inf-sup condition*:
$$\sup_{0 \ne u \in X} |a(u,v)| > 0, \quad \forall 0 \ne v \in Y. \tag{1.43}$$

Then, for any $F \in Y'$, the problem (1.34) *admits a unique solution $u \in X$, which satisfies*

$$\|u\|_X \le \frac{1}{\beta}\|F\|_{Y'}. \tag{1.44}$$

Remark 1.4. *The condition* (1.42) *is also known as the Babuška-Brezzi inf-sup condition* (cf. Babuška (1973), Brezzi (1974))*, and the real number*

$$\beta = \inf_{0\neq u\in X}\sup_{0\neq v\in Y}\frac{|a(u,v)|}{\|u\|_X\|v\|_Y} \tag{1.45}$$

is called the inf-sup constant.

Remark 1.5. *Theorem* 1.1 *with $X = Y$ is not equivalent to the Lax-Milgram lemma. In fact, one can verify readily the relation between the ellipticity and inf-sup constants: $\alpha \le \beta$. Indeed, by* (1.37)*,*

$$\alpha\|u\|_X \le \frac{|a(u,u)|}{\|u\|_X} \le \sup_{0\neq v\in X}\frac{|a(u,v)|}{\|v\|_X}, \quad \forall 0 \neq u \in X,$$

which implies

$$\alpha \le \inf_{0\neq u\in X}\sup_{0\neq v\in X}\frac{|a(u,v)|}{\|u\|_X\|v\|_X} = \beta.$$

This means that one can have $\alpha = 0$ but $\beta > 0$. In other words, the bilinear form is not coercive, but satisfies the inf-sup condition.

We review below the fundamental theory on convergence analysis of numerical approximations to (1.34).

We first consider the case $X = Y$. Assume that $X_N \subseteq X$ and

$$\forall v \in X, \quad \inf_{v_N\in X_N}\|v - v_N\|_X \to 0 \quad \text{as} \quad N \to \infty. \tag{1.46}$$

The Galerkin approximation to (1.35) is

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a(u_N, v_N) = F(v_N), \quad \forall v_N \in X_N. \end{cases} \tag{1.47}$$

The stability and convergence of this scheme can be established by using the following lemma (cf. Céa (1964)):

Theorem 1.2. (Céa Lemma). *Under the assumptions of the Lax-Milgram lemma* (*see Theorem* B.1)*, the problem* (1.47) *admits a unique solution $u_N \in X_N$ such that*

$$\|u_N\|_X \le \frac{1}{\alpha}\|F\|_{X'}. \tag{1.48}$$

Moreover, if u is the solution of (1.35)*, we have*

$$\|u - u_N\|_X \le \frac{C}{\alpha}\inf_{v_N\in X_N}\|u - v_N\|_X. \tag{1.49}$$

Here, the constants C and α are given in (1.36) *and* (1.37)*, respectively.*

Proof. Since X_N is a subspace of X, applying the Lax-Milgram lemma to (1.47) leads to the existence-uniqueness of u_N and the stability result (1.48). Now, taking $v = v_N$ in (1.35), and subtracting (1.47) from the resulting equation, we obtain the error equation

$$a(u - u_N, v_N) = 0, \quad \forall v_N \in X_N, \tag{1.50}$$

which, together with (1.36)-(1.37), implies

$$\begin{aligned} \alpha \|u - u_N\|_X^2 &\le a(u - u_N, u - u_N) = a(u - u_N, u - v_N) \\ &\le C \|u - u_N\|_X \|u - v_N\|_X, \quad \forall v_N \in X_N, \end{aligned}$$

from which (1.49) follows. □

Remark 1.6. *If, in addition, the bilinear form is symmetric, i.e., $a(u,v) = a(v,u)$, the Galerkin method is referred to as the Ritz method. In this case, the constant in the upper bound of* (1.49) *can be improved to $\sqrt{C\alpha^{-1}}$.*

Remark 1.7. *In performing error analysis of spectral methods, we usually take v_N in* (1.49) *to be a suitable orthogonal projection of u upon X_N, denoted by $\pi_N u$, which leads to*

$$\|u - u_N\|_X \le \frac{C}{\alpha} \|u - \pi_N u\|_X. \tag{1.51}$$

Hence, the error estimate follows from the approximation result on $\|u - \pi_N u\|_X$, which takes a typical form:

$$\|u - \pi_N u\|_X \le c N^{-\sigma(m)} \|u\|_{H^m}, \tag{1.52}$$

where c is a generic positive constant independent of N and any function, $\sigma(m) > 0$ is the so-called order of convergence in terms of the regularity index m, and H^m is a suitable Sobolev space with a norm involving derivatives of u up to m-th order. The establishment of such approximation results for each family of orthogonal polynomials/functions will be another emphasis of this book.

Typically, if u is sufficiently smooth, the estimate (1.52) *is valid for every m. However, for a finite-element method, the order of convergence is restricted by the order of local basis functions. The explicit dependence of the estimates of* (1.52) *type on the regularity index m will also be explored in this book.*

Observe that the bilinear form and the functional F in the discrete problem (1.47) are the same as those in the continuous problem (1.35). However, it is often convenient to use suitable approximate bilinear forms and/or functionals (see, for example, (1.31)). Hence, it is necessary to consider the following approximation to (1.35):

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a_N(u_N, v_N) = F_N(v_N), \quad \forall v_N \in X_N, \end{cases} \tag{1.53}$$

where X_N still satisfies (1.46), and $a_N(\cdot,\cdot)$ and $F_N(\cdot)$ are suitable approximations to $a(\cdot,\cdot)$ and $F(\cdot)$, respectively. In general, although X_N is a subspace of X, the

properties of the discrete bilinear form can not carry over from those of the continuous one. Hence, they have to be derived separately.

The result below, known as the first *Strang lemma* (see, e.g., Strang and Fix (1973), Ciarlet (1978)), is a generalization of Theorem 1.2.

Theorem 1.3. (First Strang lemma). *Under the assumptions of the Lax-Milgram lemma, suppose further that the discrete forms $F_N(\cdot)$ and $a_N(\cdot,\cdot)$ satisfy the same properties in the subspace $X_N \subset X$, and $\exists\, \alpha_* > 0$, independent of N, such that*

$$a_N(v,v) \geq \alpha_* \|v\|_X^2, \quad \forall v \in X_N. \tag{1.54}$$

Then, the problem (1.53) *admits a unique solution $u_N \in X_N$, satisfying*

$$\|u_N\|_X \leq \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|F_N(v_N)|}{\|v_N\|_X}. \tag{1.55}$$

Moreover, if u is the solution of (1.35)*, we have*

$$\begin{aligned} \|u - u_N\|_X \leq \inf_{w_N \in X_N} \Bigg\{ & \left(1 + \frac{C}{\alpha_*}\right) \|u - w_N\|_X \\ & + \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|a(w_N, v_N) - a_N(w_N, v_N)|}{\|v_N\|_X} \Bigg\} \\ & + \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|F(v_N) - F_N(v_N)|}{\|v_N\|_X}. \end{aligned} \tag{1.56}$$

Here, the constant C is given in (1.36)*.*

Proof. The existence-uniqueness and stability of (1.55) follow from the Lax-Milgram lemma. The proof of (1.56) is slightly different from that of (1.49). For any $w_N \in X_N$, let $e_N = u_N - w_N$. Using (1.54), (1.35) and (1.53) leads to

$$\begin{aligned} \alpha^* \|e_N\|_X^2 \leq a_N(e_N, e_N) &= a(u - w_N, e_N) + a(w_N, e_N) \\ &\quad - a_N(w_N, e_N) + F_N(e_N) - F(e_N). \end{aligned}$$

Since the result is trivial for $e_N = 0$, we derive from (1.36) that for $e_N \neq 0$,

$$\begin{aligned} \alpha^* \|e_N\|_X &\leq C\|u - w_N\|_X + \frac{|a(w_N, e_N) - a_N(w_N, e_N)|}{\|e_N\|_X} \\ &\quad + \frac{|F(e_N) - F_N(e_N)|}{\|e_N\|_X} \\ &\leq C\|u - w_N\|_X + \sup_{0 \neq v_N \in X_N} \frac{|a(w_N, v_N) - a_N(w_N, v_N)|}{\|v_N\|_X} \\ &\quad + \sup_{0 \neq v_N \in X_N} \frac{|F(v_N) - F_N(v_N)|}{\|v_N\|_X}, \end{aligned}$$

which, together with the triangle inequality, yields

$$\|u-u_N\|_X \le \|u-w_N\|_X + \|e_N\|_X.$$

Finally, taking the infimum over $w_N \in X_N$ leads to the desired result. □

The previous discussions were restricted to approximations of the abstract problem (1.35) based on Galerkin-type formulations. Similar analysis can be done for the Petrov-Galerkin approximation of (1.34) by using Theorem 1.1. Indeed, let $X_N \subseteq X$ and $Y_N \subseteq Y$. Consider the approximation to (1.34):

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a(u_N, v_N) = F(v_N), \quad \forall v_N \in Y_N. \end{cases} \tag{1.57}$$

Unlike the coercivity property, the inf-sup property can not carry over from the whole space to the subspace. Indeed, the infimum in (1.39) will not decrease if it is taken on a subspace, whereas the supremum in the inf-sup constant (1.45), in general, becomes smaller on a subspace. Consequently, we have to prove

- Discrete inf-sup condition:

$$\exists \beta_* > 0 \quad \text{such that} \quad \sup_{0\neq v_N \in Y_N} \frac{|a(u_N, v_N)|}{\|u_N\|_X \|v_N\|_Y} \ge \beta_*, \quad \forall 0 \neq u_N \in X_N, \tag{1.58}$$

- Discrete "transposed" inf-sup condition:

$$\sup_{0\neq u_N \in X_N} |a(u_N, v_N)| > 0, \quad \forall 0 \neq v_N \in Y_N. \tag{1.59}$$

The following result, which is another generalization of Theorem 1.2, can be found in Babuška and Aziz (1972).

Theorem 1.4. *Under the assumptions of Theorem 1.1, assume further that (1.58) and (1.59) hold. Then the discrete problem (1.57) admits a unique solution $u_N \in X_N$, satisfying*

$$\|u_N\|_X \le \frac{1}{\beta_*} \|F\|_{Y'}. \tag{1.60}$$

Moreover, if u is the solution of (1.34), we have

$$\|u-u_N\|_X \le \left(1 + \frac{C}{\beta_*}\right) \inf_{v_N \in X_N} \|u - v_N\|_X, \tag{1.61}$$

where the constant C is given in (1.41).

Remark 1.8. *If we consider the following approximation to (1.34):*

$$\begin{cases} \textit{Find } u_N \in X_N \textit{ such that} \\ a_N(u_N, v_N) = F_N(v_N), \quad \forall v_N \in Y_N, \end{cases} \tag{1.62}$$

then a result similar to Theorem 1.3 can be derived, provided that (1.58) and (1.59) hold in the subspaces X_N and Y_N.

1.5 Comparative Numerical Examples

The aim of this section is to provide some illustrative numerical examples for a qualitative comparison of:

- Global versus local approximations
- Spectral-Galerkin versus spectral-collocation methods

in terms of accuracy, computational complexity and/or conditioning of the linear systems.

1.5.1 Finite-Difference Versus Spectral-Collocation

In order to illustrate the main differences between the finite-difference and spectral methods, we compare numerical differentiations of a periodic function u by using a fourth-order finite-difference method and a spectral-collocation method.

Given $h = \frac{2\pi}{N}$ and a uniform grid $\{x_0, x_1, \ldots, x_N\}$ with $x_j = jh$, and a set of physical values $\{u_0, u_1, \ldots, u_N\}$ with $u_j = u(x_j)$, a fourth-order centered finite-difference approximation to $u'(x_j)$ is

$$w_j := \frac{u_{j-2} - 8u_{j-1} + 8u_{j+1} - u_{j+2}}{12h}. \tag{1.63}$$

To account for periodicity of u, we set

$$u_{-2} = u_{N-1}, \quad u_{-1} = u_N, \quad u_0 = u_{N+1}, \quad u_1 = u_{N+2}.$$

Then, the differentiation process (1.63) can be expressed as

$$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ w_{N-1} \\ w_N \end{bmatrix} = \frac{1}{12h} \begin{bmatrix} & & & \ddots & & & 1 & -8 \\ & & & \ddots & -1 & & 1 & \\ & & & \ddots & 8 & \ddots & & \\ & & & \ddots & 0 & \ddots & & \\ & & & \ddots & -8 & \ddots & & \\ -1 & & & & 1 & \ddots & & \\ 8 & -1 & & & & \ddots & & \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}. \tag{1.64}$$

Note that the coefficient matrix is sparse, reflecting the local nature of the finite-difference method.

On the other hand, the Fourier-collocation approximation of the function u is

$$\phi(x) = \sum_{k=0}^{N-1} h_k(x) u_k, \tag{1.65}$$

where $h_k(x_j) = \delta_{jk}$ and (cf. Lemma 2.2)

$$h_k(x) = \frac{1}{N} \frac{\sin\big(N(x-x_k)/2\big)}{\sin\big((x-x_k)/2\big)} \cos\big((x-x_k)/2\big). \tag{1.66}$$

Then, we approximate $u'(x_j)$ by

$$w_j = \phi'(x_j) = \sum_{k=0}^{N-1} h_k'(x_j) u_k, \quad j = 0, 1, \ldots, N-1, \tag{1.67}$$

where we have the explicit formula (cf. (2.34)):

$$h_k'(x_j) = \begin{cases} \frac{(-1)^{k+j}}{2} \cot\left[\frac{(j-k)\pi}{N}\right], & \text{if } j \neq k, \\ 0, & \text{if } j = k. \end{cases} \tag{1.68}$$

Thus, the matrix form of (1.67) becomes

$$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ w_{N-2} \\ w_{N-1} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \ddots & \vdots & \\ \ddots & -\cot\frac{2h}{2} & \\ \ddots & \cot\frac{h}{2} & \\ & 0 & \ddots \\ & -\cot\frac{h}{2} & \ddots \\ & \cot\frac{2h}{2} & \ddots \\ & \vdots & \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{bmatrix}. \tag{1.69}$$

Note that the coefficient matrix is full, reflecting the global nature of the spectral-collocation method. More detailed discussions of the Fourier method will be conducted in Chap. 2.

Next, we take $u(x) = \ln(2 + \sin x)$, which is 2π-periodic, and compare the exact derivative $u'(x) = \cos x/(2 + \sin x)$ with the numerical derivative $\{w_j\}$ obtained by the finite difference (1.64), and Fourier-collocation method (1.69) at the same grid. In Fig. 1.1, we plot the error $\max_{0 \le j \le N-1} |u'(x_j) - w_j|$ against various N. We observe a fourth-order convergence $O(h^4)$ (or $O(N^{-4})$) of the finite difference (1.64). We also observe that the Fourier-collocation method converges much faster than the

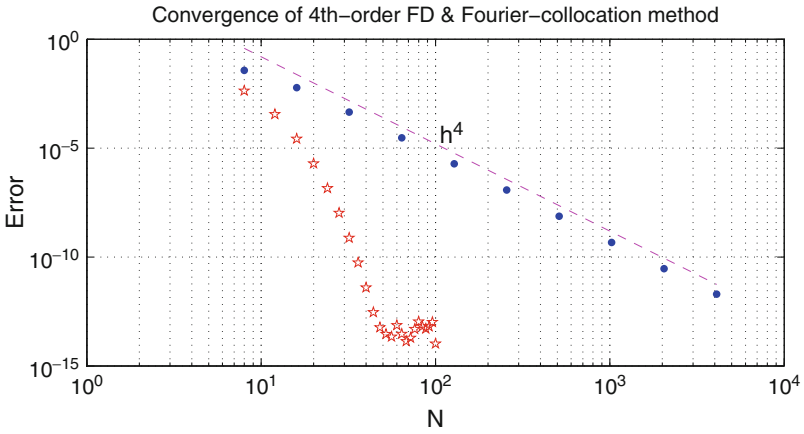

Fig. 1.1 Convergence of 4th-order finite difference (1.64) and Fourier-collocation (1.69) differentiation processes

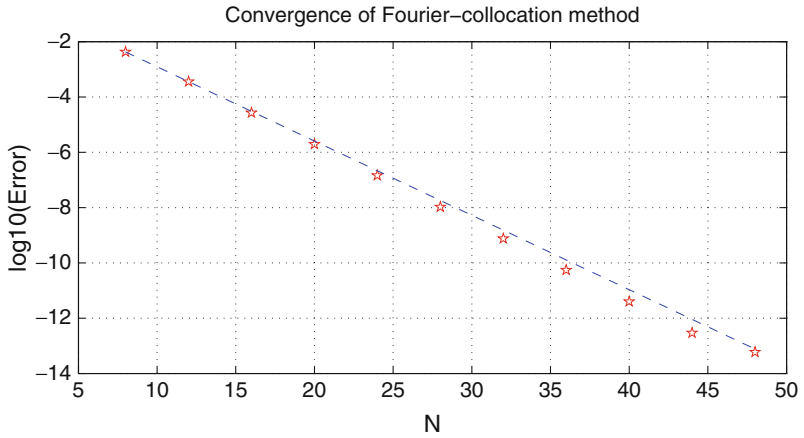

Fig. 1.2 Convergence of Fourier-collocation (1.69) differentiation process

finite difference method. To have a clearer picture of the convergence of the Fourier-collocation method (1.69), we plot in Fig. 1.2 the errors in the semi-log scale, which indicates an exponential convergence rate $O(e^{-cN})$ for some $c > 0$.

Remark 1.9. *The typical convergence behavior of a spectral method is* $O(N^{-m})$ *where* m *is a regularity index of the underlying function. In other words, its convergence rate is only limited by the regularity of the underlying function. A method exhibiting such a convergence behavior is often said to have spectral accuracy in the literature. On the other hand, the convergence rate of a finite element/finite difference method is limited by the order of the method, regardless of the regularity of the underlying function.*

A main advantage of a spectral method over a low-order finite element/finite difference method is that the former requires much fewer unknowns to resolve a given problem to a fixed accuracy, leading to potentially significant savings in storage and CPU time. For example, a rule of thumb (cf. Gottlieb and Orszag (1977)) *is that to achieve an engineering precision of* 1%, *a spectral method only needs* π *points per wave-length, as opposed to roughly ten points per wave-length required by a low-order method.*

Another important feature of spectral methods is that the derivatives of discrete functions are usually computed exactly (cf. (1.14)). *Therefore, spectral methods are usually free of phase errors, which can be very problematic for long-time integrations of partial differential equations.*

Remark 1.10. *If a function is analytic in a strip of width* 2β (*containing the underlying interval*) *in the complex plane, spectral approximations of such function can achieve an exponential convergence rate of* $O(e^{-\beta N})$. *We refer to* Davis (1975), Szegö (1975), *and* Gottlieb et al. (1992), Gottlieb and Shu (1997) *for such results on spectral projection errors, and to* Tadmor (1986) *for spectral differentiation errors* (*see* Reddy and Weideman (2005) *for a simpler analysis which also improved the estimates in* Tadmor (1986)). *Since the condition for an exponential convergence of order* $O(e^{-\beta N})$ *is quite generic, we shall not conduct analysis with exponential convergence in this book.*

1.5.2 Spectral-Galerkin Versus Spectral-Collocation

We compare in this section two versions of spectral methods: the Galerkin method in the frequency space and the collocation method in the physical space, in terms of the conditioning and round-off errors.

As an illustrative example, we consider the problem

$$u - u_{xx} = f, \quad u(\pm 1) = 0$$

with the exact solution $u(x) = \sin(10\pi x)$. In the comparison, the collocation solution is computed by (1.17)-(1.18) with $p = 0, q = \alpha$ and $g_{\pm} = 0$, while the Galerkin solution is obtained by solving (1.22) with the Legendre basis functions (cf. (1.23))

$$\phi_k(x) = \frac{1}{\sqrt{4k+6}}\big(L_k(x) - L_{k+2}(x)\big), \quad 0 \le k \le N-2.$$

Let us first examine the conditioning of the two linear systems. In Table 1.1, we list the condition numbers of the matrices resulted from the collocation method (COL) and the Galerkin method (GAL).

We see that for various α, the condition numbers of the GAL systems are all relatively small and independent of N, while those of the COL systems increase like $O\big(N^4\big)$.

Table 1.1 Condition numbers of COL and GAL

N	Method	$\alpha = 0$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1,000$
32	COL	1.04E+04	2.05E+03	2.50E+02	2.64E+01
32	GAL	1.00	5.07	41.6	396
64	COL	1.71E+05	3.37E+04	4.09E+03	4.18E+02
64	GAL	1.00	5.07	41.6	407
128	COL	2.77E+06	5.47E+05	6.63E+04	6.78E+03
128	GAL	1.00	5.07	41.7	408
256	COL	4.46E+07	8.81E+06	1.07E+06	1.09E+05
256	GAL	1.00	5.07	41.7	408

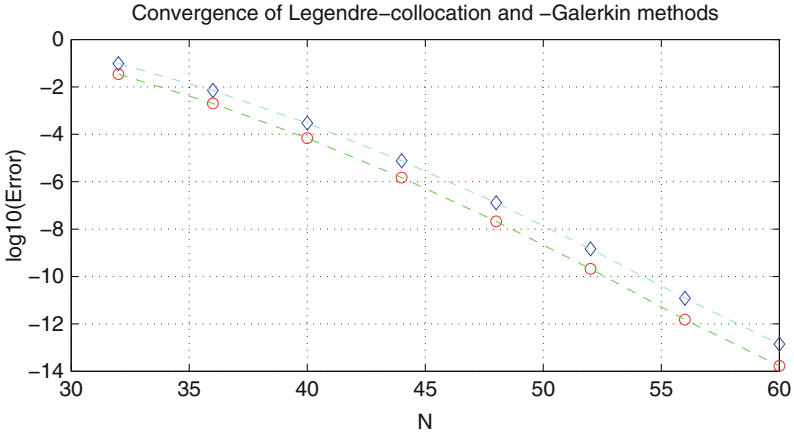

Fig. 1.3 Convergence: COL ("◇") vs. GAL ("○")

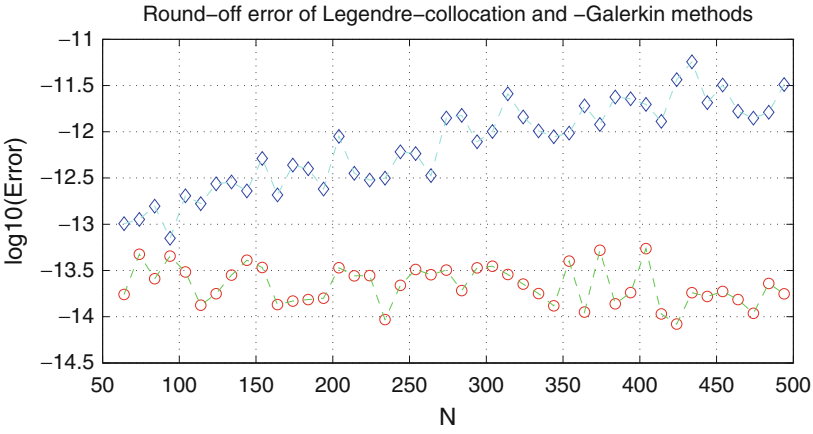

Fig. 1.4 Round-off errors: COL ("◇") vs. GAL ("○")

Next, we compare the effect of round-off errors. The maximum point-wise errors of two methods against various N are depicted in Figs. 1.3 and 1.4. We observe from Fig. 1.3 that, for relatively small N, both methods share essentially the same order of

convergence rate. However, Fig. 1.4 indicates that the effect of roundoff errors may become severer in a collocation method as N becomes large.

The above comparison is performed on a simple one-dimensional model problem. It should be pointed out that similar behaviors can be expected for multidimensional and/or higher-order problems. Finally, we want to emphasize that in a collocation method, the choice of the collocation points (the quadrature nodes) should be in agreement with underlying differential equations and boundary conditions. For instance, the Gauss-Lobatto points are not suitable for third-order equations (cf. Huang and Sloan (1992), Merryfield and Shizgal (1993)). However, in a spectral-Galerkin method, the use of quadrature rules is merely to evaluate the integrals, so the usual Gauss-Lobatto quadrature works for the third-order equation as well.

Problems

1.1. Consider the heat equation

$$u_t(x,t) = u_{xx}(x,t), \;\; t > 0; \quad u(x,0) = u_0(x), \tag{1.70}$$

where $u_0(x)$ is 2π-periodic. We expand the periodic solution u in terms of Fourier series (cf. Sect. 2.1.1)

$$u(x,t) = \sum_{|k|=0}^{\infty} a_k(t) e^{\mathrm{i}kx} \;\text{ with }\; a_k(t) = \frac{1}{2\pi}\int_0^{2\pi} u(x,t) e^{-\mathrm{i}kx} dx, \tag{1.71}$$

where $\mathrm{i} = \sqrt{-1}$ is the complex unit.

(a) Show that

$$a_k(t) = e^{-k^2 t} a_k(0), \quad \forall t \geq 0, \;\; k \in \mathbb{Z},$$

where

$$a_k(0) = \frac{1}{2\pi}\int_0^{2\pi} u_0(x) e^{-ikx} dx.$$

(b) Let

$$u_N(x,t) := \sum_{|k|=0}^{N-1} a_k(t) e^{ikx}.$$

Show that

$$\|(u - u_N)(\cdot,t)\|_\infty \leq c t^{-1/2} \|u_0\|_\infty \operatorname{erfc}(\sqrt{t}N),$$

where $\|v\|_\infty = \max_{x\in[0,2\pi]} |v(x)|$, and $\operatorname{erfc}(x)$ is the complementary error function defined by

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-y^2} dy.$$

(c) Use the property

$$\operatorname{erfc}(x) \cong \frac{e^{-x^2}}{\sqrt{\pi}x}, \quad x \gg 1,$$

to prove that

$$\|(u - u_N)(\cdot, t)\|_\infty \leq ct^{-3/2} N^{-1} e^{-tN^2}, \quad \forall\, t > 0.$$

Chapter 2
Fourier Spectral Methods for Periodic Problems

The spectral method was introduced in Orszag's pioneer work on using Fourier series for simulating incompressible flows about four decades ago (cf. Orszag (1971)). The word "spectral" was probably originated from the fact that the Fourier series are the eigenfunctions of the Laplace operator with periodic boundary conditions. This fact and the availability of the fast Fourier transform (FFT) are two major reasons for the extensive applications of Fourier methods to problems with periodic boundary conditions. In practice, a variety of physical problems exhibit periodicity. For instance, some problems are geometrically and physically periodic, such as crystal structures and homogeneous turbulence. On the other hand, many problems of scientific interest, such as the interaction of solitary waves and homogeneous turbulence, can be modeled by PDEs with periodic boundary conditions. Furthermore, even if an original problem is not periodic, the periodicity may be induced by using coordinate transforms, such as polar, spherical and cylindrical coordinates. Indeed, there are numerous circumstances where the problems are periodic in one or two directions, and non-periodic in other directions. In such cases, it is natural to use Fourier series in the periodic directions and other types of spectral expansions, such as Legendre or Chebyshev polynomials, in the non-periodic directions (cf. Chap. 7).

The objective of this chapter is to study some computational and theoretical aspects of Fourier spectral methods for periodic problems. In the first section, we introduce the continuous and discrete Fourier series, and examine the fundamental spectral techniques including discrete Fourier transforms, Fourier differentiation matrices and Fourier spectral differentiation based on FFT. The approximation properties of continuous and discrete Fourier series are surveyed in the second section. The applications of Fourier spectral methods to some linear and nonlinear problems are presented in the last section. For more detail and other aspects of Fourier approximations, we refer to Gottlieb and Orszag (1977), Gottlieb et al. (1984), Boyd (2001) and the references therein.

J. Shen et al., *Spectral Methods: Algorithms, Analysis and Applications*,
Springer Series in Computational Mathematics 41, DOI 10.1007/978-3-540-71041-7_2,

2.1 Continuous and Discrete Fourier Transforms

This section is devoted to a brief review of the properties of Fourier series and Fourier transforms. Our focus is put on the discrete Fourier transforms and Fourier differentiation techniques, which play an important role in the Fourier spectral methods.

2.1.1 Continuous Fourier Series

We denote the complex exponentials by

$$E_k(x) := e^{\mathrm{i}kx} = \cos kx + \mathrm{i}\sin kx = (\cos x + \mathrm{i}\sin x)^k, \quad k \in \mathbb{Z},\ x \in \mathbb{R},$$

where $\mathrm{i} = \sqrt{-1}$. The set $\{e^{\mathrm{i}kx} : k \in \mathbb{Z}\}$ forms a complete orthogonal system in the complex Hilbert space $L^2(0, 2\pi)$, equipped with the inner product and the norm

$$(u, v) = \frac{1}{2\pi}\int_0^{2\pi} u(x)\bar{v}(x)dx, \quad \|u\| = \sqrt{(u,u)}\,,$$

where $\bar{v}$ is the complex conjugate of v. The orthogonality of $\{E_k : k \in \mathbb{Z}\}$ reads

$$(E_k, E_m) = \frac{1}{2\pi}\int_0^{2\pi} e^{\mathrm{i}(k-m)x}dx = \delta_{km}, \tag{2.1}$$

where δ_{km} is the Kronecker Delta symbol.

For any complex-valued function $u \in L^2(0, 2\pi)$, its *Fourier series* is defined by

$$u(x) \sim \mathscr{F}(u)(x) := \sum_{k=-\infty}^{\infty} \hat{u}_k e^{\mathrm{i}kx}, \tag{2.2}$$

where the *Fourier coefficients* are given by

$$\hat{u}_k = (u, e^{\mathrm{i}kx}) = \frac{1}{2\pi}\int_0^{2\pi} u(x)e^{-\mathrm{i}kx}dx. \tag{2.3}$$

It is clear that if u is a real-valued function, its Fourier coefficients satisfy

$$\hat{u}_{-k} = \overline{\hat{u}}_k, \quad k \in \mathbb{Z}, \tag{2.4}$$

and $\hat{u}_0$ is obviously real.

In fact, the Fourier series can be defined for general absolutely integrable functions in $(0, 2\pi)$, and the convergence theory of Fourier expansions in different senses has been subjected to a rigorous and thorough investigation in Fourier analysis (see, e.g., Zygmund (2002), Stein and Shakarchi (2003)). It is well-known

that, for any $u \in L^2(0,2\pi)$, its truncated Fourier series $\mathscr{F}_N(u) := \sum_{|k|\le N} \hat{u}_k e^{\mathrm{i}kx}$ converges to u in the L^2-sense, and there holds the Parseval's identity:

$$\|u\|^2 = \sum_{k=-\infty}^{\infty} |\hat{u}_k|^2. \tag{2.5}$$

If u is continuous, periodic and of bounded variation on $[0,2\pi]$, then $\mathscr{F}_N(u)$ uniformly converges to u.

Notice that the truncated Fourier series can also be expressed in the convolution form, namely,

$$\mathscr{F}_N(u)(x) = \big(\mathscr{D}_N * u\big)(x) = \frac{1}{2\pi}\int_0^{2\pi} \mathscr{D}_N(x-t)u(t)dt, \tag{2.6}$$

where $\mathscr{D}_N$ is known as the *Dirichlet kernel* given by

$$\mathscr{D}_N(x) := \sum_{k=-N}^{N} e^{\mathrm{i}kx} = 1 + 2\sum_{k=1}^{N} \cos kx = \frac{\sin((N+1/2)x)}{\sin(x/2)}. \tag{2.7}$$

It is sometimes convenient to express the Fourier series in terms of the trigonometric polynomials:

$$u(x) \sim \mathscr{S}(u)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \big(a_k \cos kx + b_k \sin kx\big), \tag{2.8}$$

where the expansion coefficients are

$$a_k = \frac{1}{\pi}\int_0^{2\pi} u(x)\cos kx dx, \quad b_k = \frac{1}{\pi}\int_0^{2\pi} u(x)\sin kx dx.$$

The coefficients of the two different representations (2.2) and (2.8) are related by

$$\hat{u}_0 = \frac{a_0}{2}, \quad \hat{u}_k = \begin{cases} \dfrac{a_k - \mathrm{i}b_k}{2}, & \text{if } k \ge 1, \\ \dfrac{a_{-k} + \mathrm{i}b_{-k}}{2}, & \text{if } k \le -1. \end{cases} \tag{2.9}$$

In particular, if u is a real-valued function, then

$$a_0 = 2\hat{u}_0, \quad a_k = 2\mathrm{Re}(\hat{u}_k), \quad b_k = -2\mathrm{Im}(\hat{u}_k), \quad k \ge 1. \tag{2.10}$$

2.1.2 Discrete Fourier Series

Given a positive integer N, let

$$x_j = jh = j\frac{2\pi}{N}, \quad 0 \le j \le N-1, \tag{2.11}$$

be the N-equispaced grids in $[0, 2\pi)$, which are referred to as the Fourier collocation points. We define the discrete inner product by

$$\langle u, v\rangle_N = \frac{1}{N}\sum_{j=0}^{N-1} u(x_j)\bar{v}(x_j). \tag{2.12}$$

The following lemma is the discrete counterpart of (2.1).

Lemma 2.1. *Let $E_l(x) = e^{\mathrm{i}lx}$. For any integer $N \geq 1$, we have*

$$\big\langle E_k, E_m\big\rangle_N = \begin{cases} 1, & \text{if } k-m = lN,\ \forall l \in \mathbb{Z}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

Proof. Observe that if $k-m$ is not divisible by N, then

$$\begin{aligned} \big\langle E_k, E_m\big\rangle_N &= \frac{1}{N}\sum_{j=0}^{N-1} e^{\mathrm{i}(k-m)x_j} = \frac{1}{N}\sum_{j=0}^{N-1}\Big(e^{2\pi\mathrm{i}(k-m)/N}\Big)^j \\ &= \frac{1}{N}\frac{e^{2\pi\mathrm{i}(k-m)}-1}{e^{2\pi\mathrm{i}(k-m)/N}-1} = 0. \end{aligned}$$

If $k-m$ is divisible by N, we have $e^{2\pi\mathrm{i}(k-m)/N} = 1$, so the summation in the second line above equals to 1. □

In general, the Fourier coefficients $\{\hat{u}_k\}$ in (2.3) can not be evaluated exactly, so we have to resort to some quadrature formula. A simple and accurate quadrature formula for 2π-periodic functions is the rectangular rule

$$\frac{1}{2\pi}\int_0^{2\pi} v(x)dx \approx \frac{1}{N}\sum_{j=0}^{N-1} v(x_j), \quad \forall v \in C[0, 2\pi), \tag{2.14}$$

which is exact for all

$$v \in \operatorname{span}\big\{e^{\mathrm{i}kx} : 0 \leq |k| \leq N-1\big\}.$$

Moreover, one verifies readily that (2.14) is also exact for $v = \sin(\pm Nx)$ but not for $v = \cos(\pm Nx)$.

Applying (2.14) to (2.3) leads to the approximation

$$\hat{u}_k \approx \tilde{u}_k := \frac{1}{N}\sum_{j=0}^{N-1} u(x_j)e^{-\mathrm{i}kx_j}, \quad k = 0, \pm 1, \ldots. \tag{2.15}$$

Note that $\{\tilde{u}_k\}$ are N-periodic, that is,

$$\tilde{u}_{k\pm N} = \frac{1}{N}\sum_{j=0}^{N-1} u(x_j)e^{-\mathrm{i}(k\pm N)x_j} = \frac{1}{N}\sum_{j=0}^{N-1} u(x_j)e^{-\mathrm{i}kx_j}e^{\mp 2\pi\mathrm{i}j} = \tilde{u}_k,$$

which implies that for even N, we have

$$\tilde{u}_{-N/2} = \tilde{u}_{N/2}. \tag{2.16}$$

Hence, for even N, the grids $\{x_j\}_{j=0}^{N-1}$ can not distinguish the modes: $k = \pm N/2$, since

$$e^{\mathrm{i}Nx_j/2} = e^{\mathrm{i}j\pi} = (-1)^j = e^{-\mathrm{i}Nx_j/2}, \quad 0 \le j \le N-1. \tag{2.17}$$

In other words, the two modes $k = \pm N/2$ are *aliased*.

In order to have an effective implementation of the discrete Fourier transform (DFT), it is preferable to use an even N, and accordingly, a symmetric finite set of modes: $-N/2 \le k \le N/2$ in the discrete Fourier series (cf. (2.20) below). In view of (2.16)–(2.17), we redefine the approximation (2.15) by modifying the two modes $k = \pm N/2$:

$$\hat{u}_k \approx \tilde{u}_k = \frac{1}{Nc_k} \sum_{j=0}^{N-1} u(x_j) e^{-\mathrm{i}kx_j}, \quad k = -N/2, \ldots, N/2, \tag{2.18}$$

where $c_k = 1$ for $|k| < N/2$, and $c_k = 2$ for $k = \pm N/2$. The expression (2.18) is referred to as the (forward) *discrete Fourier transform* of $u(x)$ associated with the grid points in (2.11).

Due to (2.16), there are only N independent coefficients. Hence, we set

$$\mathscr{T}_N = \Big\{ u = \sum_{k=-N/2}^{N/2} \tilde{u}_k e^{\mathrm{i}kx} : \tilde{u}_{-N/2} = \tilde{u}_{N/2} \Big\}, \tag{2.19}$$

and define the mapping $I_N : C[0, 2\pi) \to \mathscr{T}_N$ by

$$(I_N u)(x) = \sum_{k=-N/2}^{N/2} \tilde{u}_k e^{\mathrm{i}kx}, \tag{2.20}$$

with $\{\tilde{u}_k\}$ given by (2.18). The following lemma shows that I_N is the interpolation operator from $C[0, 2\pi)$ to $\mathscr{T}_N$ such that

$$(I_N u)(x_j) = u(x_j), \quad x_j = \frac{2\pi j}{N}, \ 0 \le j \le N-1. \tag{2.21}$$

Lemma 2.2. *For any* $u \in C[0, 2\pi)$,

$$(I_N u)(x) = \sum_{j=0}^{N-1} u(x_j) h_j(x), \tag{2.22}$$

where

$$h_j(x) = \frac{1}{N} \sin\left[N\frac{x-x_j}{2}\right] \cot\left[\frac{x-x_j}{2}\right] \in \mathscr{T}_N \tag{2.23}$$

satisfying

$$h_j(x_k) = \delta_{jk}, \quad \forall j,k = 0,1,\ldots,N-1. \tag{2.24}$$

Proof. By (2.18) and (2.20),

$$\begin{aligned}(I_N u)(x) &= \sum_{k=-N/2}^{N/2} \left(\frac{1}{Nc_k} \sum_{j=0}^{N-1} u(x_j) e^{-\mathrm{i}kx_j} \right) e^{\mathrm{i}kx} \\ &= \sum_{j=0}^{N-1} \left(\frac{1}{N} \sum_{k=-N/2}^{N/2} \frac{1}{c_k} e^{\mathrm{i}k(x-x_j)} \right) u(x_j) \\ &=: \sum_{j=0}^{N-1} h_j(x) u(x_j).\end{aligned}$$

We derive from (2.7) and a direct calculation that

$$\begin{aligned}h_j(x) &= \frac{1}{N} \sum_{k=-N/2}^{N/2} \frac{1}{c_k} e^{\mathrm{i}k(x-x_j)} \\ &= \frac{1}{N} \left(D_{N/2-1}(x-x_j) + \cos\left[N \frac{x-x_j}{2} \right] \right) \\ &= \frac{1}{N} \left(\frac{\sin\left[(N-1)\frac{x-x_j}{2} \right]}{\sin \frac{x-x_j}{2}} + \cos\left[N \frac{x-x_j}{2} \right] \right) \\ &= \frac{1}{N} \sin\left[N \frac{x-x_j}{2} \right] \cot\left[\frac{x-x_j}{2} \right].\end{aligned} \tag{2.25}$$

Due to (2.17), we have $h_j(x) \in \mathscr{T}_N$, and it is clear that $h_j(x_i) = 0$ for $i \neq j$. Moreover, taking $x = x_j$ in the first identity of (2.25) yields $h_j(x_j) = 1$. □

Taking $x = x_j$ in (2.20) and using (2.21), leads to the *inverse* (*or backward*) *discrete transform*:

$$u(x_j) = \sum_{k=-N/2}^{N/2} \tilde{u}_k e^{\mathrm{i}kx_j}, \quad j = 0,1,\ldots,N-1. \tag{2.26}$$

It is obvious that the discrete Fourier transforms (2.18) and its inverse (2.26) can be carried out through matrix–vector multiplication with $O(N^2)$ operations. However, thanks to the fast Fourier transforms due to Cooley and Tukey (1965), such processes can be accomplished with $O(N \log_2 N)$ operations. Moreover, if u is a real valued function, then $\tilde{u}_{-k} = \overline{\tilde{u}}_k$, so only half of the coefficients in (2.26) need to be computed/stored.

The computational routines for FFT and IFFT are available in many software packages. Here, we restrict our attentions to their implementations in `MATLAB`. Given the data $\{\mathbf{v}(j) = u(x_{j-1})\}_{j=1}^N$ sampled at $\{x_k = 2\pi k/N\}_{k=0}^{N-1}$, the command "$\tilde{\mathbf{v}} = \texttt{fft}(\mathbf{v})$" returns the vector $\{\tilde{\mathbf{v}}(k)\}_{k=1}^N$, defined by

$$\tilde{\mathbf{v}}(k) = \sum_{j=1}^{N} \mathbf{v}(j) e^{-2\pi \mathrm{i}(j-1)(k-1)/N}, \quad 1 \le k \le N, \tag{2.27}$$

while the inverse FFT can be computed with the command "$\mathbf{v} = \texttt{ifft}(\tilde{\mathbf{v}})$" which returns the physical values $\{\mathbf{v}(j)\}_{j=1}^{N}$ via

$$\mathbf{v}(j) = \frac{1}{N} \sum_{k=1}^{N} \tilde{\mathbf{v}}(k) e^{2\pi \mathrm{i}(j-1)(k-1)/N}, \quad 1 \le j \le N. \tag{2.28}$$

Notice that some care has to be taken for the ordering of the modes. To illustrate this, we examine the one-to-one correspondence of the transforms in (2.18) and (2.26). More precisely, let

$$u(x_j) = \mathbf{v}(j+1), \quad x_j = \frac{2\pi j}{N}, \quad 0 \le j \le N-1. \tag{2.29}$$

We find that

$$\begin{aligned} \tilde{u}_k &= \frac{1}{N} \tilde{\mathbf{v}}(k+1), \qquad 0 \le k \le \frac{N}{2} - 1, \\ \tilde{u}_k &= \frac{1}{N} \tilde{\mathbf{v}}(k+N+1), \quad -\frac{N}{2} + 1 \le k \le -1, \\ \tilde{u}_{-N/2} &= \tilde{u}_{N/2} = \frac{1}{2N} \tilde{\mathbf{v}}(N/2+1). \end{aligned} \tag{2.30}$$

A tabulated view of the above relations is given in the following table.

Table 2.1 Correspondence of DFT and FFT & IFFT in MATLAB

j	1	2	…	$N/2-1$	$N/2$	$N/2+1$	$N/2+2$	…	$N-1$	N
$\mathbf{u} = \mathbf{v}$	u_0	u_1	…	…	…	…	…	…	u_{N-2}	u_{N-1}
$\tilde{\mathbf{u}} = \tilde{\mathbf{v}}/N$	$\tilde{u}_0$	$\tilde{u}_1$	…	$\tilde{u}_{N/2-2}$	$2\tilde{u}_{N/2-1}$	$\tilde{u}_{N/2}$	$\tilde{u}_{-N/2+1}$	…	$\tilde{u}_{-2}$	$\tilde{u}_{-1}$
$\mathbf{k}$	0	1	…	$N/2-2$	$N/2-1$	0	$-N/2+1$	…	-2	-1

In the table, we denote $\{u_j = \mathbf{u}(j)\}_{j=0}^{N-1}$ and $\{\tilde{u}_k = \tilde{\mathbf{u}}(k)\}_{k=-N/2}^{N/2}$. The last row gives the frequency vector $\mathbf{k}$ for the Fourier-spectral differentiation based on FFT, see Sect. 2.1.4 below. Note that the frequency $-N/2$ is aliased with the frequency $N/2$ in the discrete Fourier transform.

2.1.3 Differentiation in the Physical Space

In a Fourier spectral method, differentiation can be performed in the physical space as well as in the frequency space.

We start with differentiation in the physical space. Let $\{x_j\}$ and $\{h_j\}$ be defined in (2.11) and (2.23), respectively. Setting

$$u(x) = \sum_{j=0}^{N-1} u(x_j) h_j(x), \tag{2.31}$$

and taking the m-th derivative, we get

$$u^{(m)}(x) = \sum_{j=0}^{N-1} u(x_j) h_j^{(m)}(x). \tag{2.32}$$

This process can be formulated as a matrix–vector multiplication

$$\mathbf{u}^{(m)} = D^{(m)}\mathbf{u}, \quad m \geq 0, \tag{2.33}$$

where

$$\begin{aligned} D^{(m)} &= \left(d_{kj}^{(m)} := h_j^{(m)}(x_k)\right)_{k,j=0,\ldots,N-1}, \\ \mathbf{u} &= \left(u(x_0), u(x_1), \ldots, u(x_{N-1})\right)^T, \\ \mathbf{u}^{(m)} &= \left(u^{(m)}(x_0), u^{(m)}(x_1), \ldots, u^{(m)}(x_{N-1})\right)^T. \end{aligned}$$

In particular, we denote $D = D^{(1)}$. The compact form of the first-order differentiation matrix is given below.

Lemma 2.3. *The entries of the first-order Fourier differentiation matrix D are determined by*

$$d_{kj}^{(1)} = h_j'(x_k) = \begin{cases} \frac{(-1)^{k+j}}{2} \cot\left[\frac{(k-j)\pi}{N}\right], & \text{if } k \neq j, \\ 0, & \text{if } k = j. \end{cases} \tag{2.34}$$

Proof. Differentiating the Lagrange basis in (2.23) directly gives

$$\begin{aligned} h_j'(x) = &\frac{1}{2}\cos\left[N\frac{x-x_j}{2}\right]\cot\left[\frac{x-x_j}{2}\right] \\ &- \frac{1}{2N}\sin\left[N\frac{x-x_j}{2}\right]\csc^2\left[\frac{x-x_j}{2}\right]. \end{aligned}$$

It is clear that if $x = x_k \neq x_j$, then the second term is 0 and the first term can be simplified into the desired expression in (2.34).

We now consider the case $k = j$. For convenience, let $\theta = (x - x_j)/2$, and rewrite the above formula as

$$h_j'(x) = \frac{1}{2}\frac{\cos(N\theta)\cos\theta\sin\theta - N^{-1}\sin(N\theta)}{\sin^2\theta}. \tag{2.35}$$

Using the Taylor expansion, we find

$$\cos(N\theta)\cos\theta\sin\theta = \theta + O(\theta^3), \quad N^{-1}\sin(N\theta) = \theta + O(\theta^3), \quad |\theta| \ll 1.$$

Hence, we derive from (2.35) that $h_j'(x_j) = \lim_{x\to x_j} h_j'(x) = 0$, since $\theta \to 0$ as $x \to x_j$. □

Remark 2.1. *The first-order Fourier differentiation matrix has the following properties:*

- *D is a real and skew-symmetric matrix, since* $\cot(-x) = -\cot(x)$ *and* $d_{kk} = 0$.
- *D is a circulant Toeplitz matrix, since* $d_{kj} = d_{k+1,j+1}$.
- *The distinct eigenvalues of D are* $\{\mathrm{i}k : -N/2+1 \le k \le N/2-1\}$, *and the eigenvalue* 0 *has a multiplicity* 2.

The approximation of higher-order derivatives follows the same procedure. From the first relation in (2.25), we find

$$h_j^{(m)}(x_i) = \frac{1}{N}\sum_{k=-N/2}^{N/2} \frac{(\mathrm{i}k)^m}{c_k} e^{2\pi \mathrm{i}k(i-j)/N}. \tag{2.36}$$

In particular, the entries of the second-order differentiation matrix $D^{(2)}$ are given by

$$d_{kj}^{(2)} = h_j''(x_k) = \begin{cases} -\dfrac{(-1)^{k+j}}{2}\sin^{-2}\Big[\dfrac{(k-j)\pi}{N}\Big], & \text{if } k \ne j, \\ -\dfrac{N^2}{12} - \dfrac{1}{6}, & \text{if } k = j. \end{cases} \tag{2.37}$$

It is worthwhile to point out that $D^{(2)} \ne D^2$. Indeed, we consider $u = \cos(Nx/2)$ and denote $\mathbf{u}$ the vector that samples u at $\{x_j\}_{j=0}^{N-1}$. Since $u(x_j) = (-1)^j$, one verifies readily that $D\mathbf{u} = \mathbf{0}$ and $D^2\mathbf{u} = \mathbf{0}$, while $D^{(2)}\mathbf{u} = -N^2\mathbf{u}/4$.

It is clear that the differentiation procedure through (2.33) requires $O(N^2)$ operations. We shall demonstrate below how to perform the differentiation in the frequency space with $O(N\log_2 N)$ operations using FFT.

2.1.4 Differentiation in the Frequency Space

For a function given by (2.31), we can rewrite it as a finite Fourier series

$$u(x) = \sum_{k=-N/2}^{N/2} \tilde{u}_k e^{\mathrm{i}kx}, \tag{2.38}$$

where $\tilde{u}_{N/2} = \tilde{u}_{-N/2}$ as before. Thus, we have

$$u'(x_j) = \sum_{k=-N/2}^{N/2} \mathrm{i}k\tilde{u}_k e^{\mathrm{i}kx_j}, \tag{2.39}$$

where $\{x_j = 2\pi j/N\}_{j=0}^{N-1}$ are the grids given by (2.11). Given the physical values $\{u(x_j)\}_{j=0}^{N-1}$, the approximation of the derivative values $\{w_j = u'(x_j)\}_{j=0}^{N-1}$ can be computed as follows:

- Call $\tilde{\mathbf{v}} = \mathtt{fft}(\mathbf{v})$, where the components of the input vector $\mathbf{v}$ are $v(j) = u(x_{j-1})$, $j = 1,\ldots,N$, and which returns the frequency vector:

$$\tilde{\mathbf{v}} = \big(\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_N\big).$$

- Compute the coefficients of the expansion of the derivative:

$$\begin{aligned}\tilde{\mathbf{v}}^{(1)} &= \mathrm{i}\mathbf{k}.*\tilde{\mathbf{v}} \\ &= \mathrm{i}\Big(0, \tilde{v}_2, \ldots, (N/2-1)\tilde{v}_{N/2}, 0, (-N/2+1)\tilde{v}_{N/2+2}, \ldots, -\tilde{v}_N\Big),\end{aligned}$$

  where the multiplicative vector $\mathbf{k}$ is given in Table 2.1:

$$\mathbf{k} = \big(0, 1, \ldots, N/2-1, 0, -N/2+1, \ldots, -1\big). \tag{2.40}$$

- Call $\mathbf{w} = \mathtt{ifft}(\tilde{\mathbf{v}}^{(1)})$, which produces the desired derivative values $\{w_j\}_{j=0}^{N-1}$.

As a striking contrast to the differentiation process described in the previous section, the computational cost of the above procedure is $O\big(N\log_2 N\big)$. Moreover, higher-order derivatives can be computed using these three steps repeatedly. Multi-dimensional cases can be implemented similarly by using available routines such as `fft2.m` and `ifft2.m` in MATLAB.

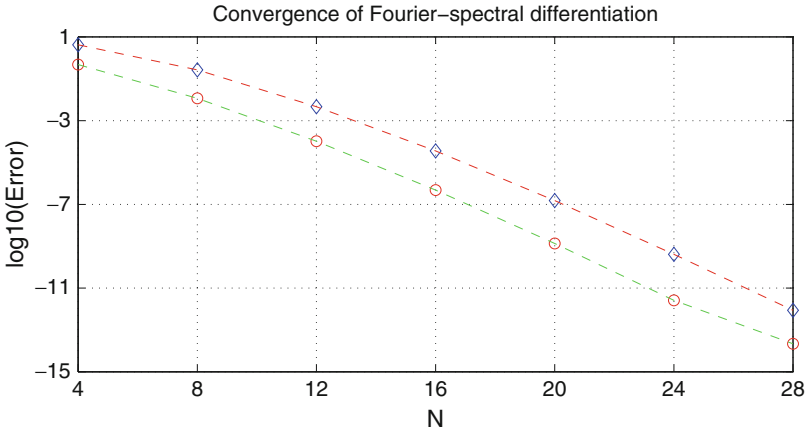

Fig. 2.1 Errors of Fourier-spectral differentiation of first-order ("○") and second-order ("◇")

As a numerical illustration, we consider the Fourier spectral differentiation of the $2\pi-$periodic function $u(x) = e^{1+\sin x}$. In Fig. 2.1, we plot, in semi-log scale, the maximum point-wise errors of the first- and second-order derivatives against various N. The plots in Fig. 2.1 clearly indicate the exponential convergence of the Fourier spectral differentiation process.

2.2 Fourier Approximation

In this section, we summarize some fundamental results on the approximation of periodic functions by the continuous and discrete Fourier series.

2.2.1 Inverse Inequalities

Since all norms of a finite dimensional space are equivalent, we can bound a strong norm by a weaker one with bounding constants depending on the dimension of the space. This type of inequality is called inverse inequality. Our aim in this section is to find the optimal constants in such inequalities.

For notational convenience, we use $A \lesssim B$ to mean that there exists a generic positive constant c, which is independent of N and any function, such that $A \leq cB$. We also use $\partial_x^m u$ or $u^{(m)}$ to denote the ordinary derivative $\frac{d^m u}{dx^m}$. Let $I := (0, 2\pi)$, and define the complex $(2N+1)$-dimensional space

$$X_N := \text{span}\{e^{ikx} : -N \leq k \leq N\}. \tag{2.41}$$

The Banach space $L^p(I)$ with $1 \leq p \leq \infty$ and its norm $\|\cdot\|_{L^p}$ are defined as in Appendix B.4.

We first recall the following *Nikolski's* inequality.

Lemma 2.4. *For any $u \in X_N$ and $1 \leq p \leq q \leq \infty$,*

$$\|u\|_{L^q} \leq \Big(\frac{Np_0 + 1}{2\pi}\Big)^{\frac{1}{p}-\frac{1}{q}} \|u\|_{L^p}, \tag{2.42}$$

where p_0 is the least even integer $\geq p$.

Another type of inverse inequality, i.e., the so-called *Bernstein* inequality, relates the L^p-norms of a function and its derivatives.

Lemma 2.5. *For any $u \in X_N$ and $1 \leq p \leq \infty$,*

$$\|\partial_x^m u\|_{L^p} \lesssim N^m \|u\|_{L^p}, \quad m \geq 1. \tag{2.43}$$

In particular, for $p = 2$,

$$\|\partial_x^m u\| \lesssim N^m \|u\|. \tag{2.44}$$

The proofs of these inverse inequalities can be found in Butzer and Nessel (1971) (also see Guo (1998b)). In particular, the derivation of (2.44) is straightforward by using $\partial_x^m(e^{ikx}) = (ik)^m e^{ikx}$, and the orthogonality of the Fourier series.

2.2.2 Orthogonal Projection

Let $P_N : L^2(I) \to X_N$ be the L^2-orthogonal projection, defined by

$$\big(P_N u - u, v\big) = 0, \quad \forall v \in X_N. \tag{2.45}$$

It is obvious that $P_N u$ is the truncated Fourier series, namely,

$$(P_N u)(x) = \sum_{k=-N}^{N} \hat{u}_k e^{ikx},$$

where $\{\hat{u}_k\}$ are given by (2.3).

We next measure the errors between $P_N u$ and u in Sobolev spaces. For this purpose, we denote by $H_p^m(I)$ the subspace of $H^m(I)$ (cf. Appendix B.4), which consists of functions with derivatives of order up to $m-1$ being 2π-periodic. In view of the relation $(e^{ikx})' = ike^{ikx}$, the norm and semi-norm of $H_p^m(I)$ can be characterized in the frequency space by

$$\|u\|_m = \left(\sum_{k=-\infty}^{\infty} (1+k^2)^m |\hat{u}_k|^2\right)^{1/2}, \quad |u|_m = \left(\sum_{k=-\infty}^{\infty} |k|^{2m} |\hat{u}_k|^2\right)^{1/2}. \tag{2.46}$$

We see that the space $H_p^m(I)$ with fractional m is also well-defined.

Formally, for any $u \in H_p^m(I)$, we can differentiate the Fourier series term-wisely, and obtain

$$\partial_x^l u(x) = \sum_{k=-\infty}^{\infty} (ik)^l \hat{u}_k e^{ikx}, \quad 0 \le l \le m,$$

which implies the commutability of the derivative and projection operators:

$$\partial_x^l (P_N u) = P_N(\partial_x^l u), \quad 0 \le l \le m. \tag{2.47}$$

The main approximation result is stated below (cf. Kreiss and Oliger (1979), Canuto and Quarteroni (1982)).

Theorem 2.1. *For any $u \in H_p^m(I)$ and $0 \le \mu \le m$,*

$$\|P_N u - u\|_\mu \lesssim N^{\mu - m} |u|_m. \tag{2.48}$$

Proof. By (2.46),

$$\begin{aligned}\|P_N u - u\|_\mu^2 &= \sum_{|k|>N} (1+k^2)^\mu |\hat{u}_k|^2 \\ &\lesssim N^{2\mu-2m} \sum_{|k|>N} |k|^{2m-2\mu}(1+k^2)^\mu |\hat{u}_k|^2 \\ &\lesssim N^{2\mu-2m} \sum_{|k|>N} |k|^{2m} |\hat{u}_k|^2 \\ &\lesssim N^{2\mu-2m} |u|_m^2.\end{aligned}$$

This completes the proof. □

This theorem indicates that the projection $P_N u$ is the best approximation of u in all Sobolev spaces $H_p^m(I)$ $(m \geq 0)$.

The L^∞-estimate of the projection errors is stated as follows.

Theorem 2.2. *For any $u \in H_p^m(I)$ with $m > 1/2$,*

$$\max_{x\in[0,2\pi]} \big|(P_N u - u)(x)\big| \leq \sqrt{\frac{1}{2m-1}} N^{1/2-m} |u|_m. \tag{2.49}$$

Proof. By the Cauchy–Schwarz inequality,

$$\begin{aligned}\big|(P_N u - u)(x)\big| \leq \sum_{|k|>N} |\hat{u}_k| &\leq \Big(\sum_{|k|>N} |k|^{-2m} \Big)^{1/2} \Big(\sum_{|k|>N} |k|^{2m} |\hat{u}_k|^2 \Big)^{1/2} \\ &\leq \sqrt{\frac{1}{2m-1}} N^{1/2-m} |u|_m.\end{aligned}$$

The last step is due to the fact that for $m > 1/2$,

$$\sum_{|k|>N} |k|^{-2m} \leq \int_N^\infty x^{-2m} dx \leq \frac{N^{1-2m}}{2m-1}.$$

This completes the proof. □

2.2.3 Interpolation

For the sake of consistency, we consider the Fourier interpolation on $2N$ collocation points $\{x_j = \pi j/N\}_{j=0}^{2N-1}$, but still denote the interpolation operator by I_N. That is,

$$(I_N u)(x) = \sum_{k=-N}^{N} \tilde{u}_k e^{\mathrm{i}kx} \tag{2.50}$$

with $\tilde{u}_N = \tilde{u}_{-N}$ and

$$\tilde{u}_k = \frac{1}{2Nc_k} \sum_{j=0}^{2N-1} u(x_j) e^{-\mathrm{i}kx_j}, \quad -N \le k \le N. \tag{2.51}$$

The interpolation error: $I_N u - u$ is characterized by the following theorem (see also Kreiss and Oliger (1979), Canuto and Quarteroni (1982)):

Theorem 2.3. *For any $u \in H_p^m(I)$ with $m > 1/2$,*

$$\|\partial_x^l (I_N u - u)\| \lesssim N^{l-m} |u|_m, \qquad 0 \le l \le m. \tag{2.52}$$

Proof. We first show that the expansion coefficients of the continuous (cf. (2.3)) and discrete Fourier series (cf. (2.51)) are connected by

$$c_k \tilde{u}_k = \hat{u}_k + \sum_{|p|>0}^{\infty} \hat{u}_{k+2pN}. \tag{2.53}$$

Indeed, plugging $u(x_j) = \sum_{|p|=0}^{\infty} \hat{u}_p e^{\mathrm{i}px_j}$ into (2.51) gives

$$c_k \tilde{u}_k = \frac{1}{2N} \sum_{j=0}^{2N-1} \Big(\sum_{|p|=0}^{\infty} \hat{u}_p e^{\mathrm{i}(p-k)x_j} \Big) = \frac{1}{2N} \sum_{|p|=0}^{\infty} \Big(\sum_{j=0}^{2N-1} e^{\mathrm{i}(p-k)x_j} \Big) \hat{u}_p,$$

where the constant $c_k = 1$ for $|k| < N$ and $c_k = 2$ for $k = \pm N$.

We deduce from Lemma 2.1 that $e^{\mathrm{i}(p-k)x_j} = 1$, if and only if $p - k = 2lN$ with $l \in \mathbb{Z}$, otherwise, it equals to zero. Hence, we have

$$c_k \tilde{u}_k = \sum_{|p|=0}^{\infty} \hat{u}_{k+2pN} = \hat{u}_k + \sum_{|p|>0}^{\infty} \hat{u}_{k+2pN},$$

which yields (2.53). Thus, a direct calculation leads to

$$\begin{aligned}
\|P_N u - I_N u\|^2 &= \sum_{|k|\le N} |\hat{u}_k - \tilde{u}_k|^2 \\
&= \sum_{|k|<N} |\hat{u}_k - \tilde{u}_k|^2 + \frac{1}{4} \sum_{k=\pm N} |2\hat{u}_k - 2\tilde{u}_k|^2 \\
&\le \sum_{|k|<N} |\hat{u}_k - \tilde{u}_k|^2 + \frac{1}{2} \sum_{k=\pm N} |\hat{u}_k - 2\tilde{u}_k|^2 + \frac{1}{2} \sum_{k=\pm N} |\hat{u}_k|^2 \\
&\le \sum_{|k|\le N} |\hat{u}_k - c_k \tilde{u}_k|^2 + \frac{1}{2} \sum_{k=\pm N} |\hat{u}_k|^2.
\end{aligned}$$

The last term is bounded by

$$|\hat{u}_N|^2 + |\hat{u}_{-N}|^2 \le N^{-2m} \sum_{|k|=N}^{\infty} |k|^{2m} |\hat{u}_k|^2 \le N^{-2m} |u|_m^2,$$

and the first term can be estimated by using the relation (2.53) and the Cauchy–Schwarz inequality:

$$\begin{aligned}\sum_{|k|\le N}|\hat{u}_k - c_k\tilde{u}_k|^2 &= \sum_{|k|\le N}\Big|\sum_{|p|>0}^{\infty}\hat{u}_{k+2pN}\Big|^2\\ &\le \sum_{|k|\le N}\Big\{\Big(\sum_{|p|>0}^{\infty}|k+2pN|^{-2m}\Big)\Big(\sum_{|p|>0}^{\infty}|k+2pN|^{2m}|\hat{u}_{k+2pN}|^2\Big)\Big\}\\ &\le \max_{|k|\le N}\Big\{\sum_{|p|>0}^{\infty}|k+2pN|^{-2m}\Big\}\Big(\sum_{|k|\le N}\sum_{|p|>0}^{\infty}|k+2pN|^{2m}|\hat{u}_{k+2pN}|^2\Big).\end{aligned}$$

It is clear that

$$\max_{|k|\le N}\Big\{\sum_{|p|>0}^{\infty}|k+2pN|^{-2m}\Big\} \le \frac{1}{N^{2m}}\sum_{|p|>0}^{\infty}\frac{1}{|2p-1|^{2m}} \lesssim N^{-2m},$$

and

$$\sum_{|k|\le N}\sum_{|p|>0}^{\infty}|k+2pN|^{2m}|\hat{u}_{k+2pN}|^2 \le 2|u|_m^2.$$

Hence, a combination of the above estimates leads to

$$\|P_Nu - I_Nu\| \lesssim N^{-m}|u|_m.$$

Moreover, by the inverse inequality (2.44),

$$\|\partial_x^l(P_Nu - I_Nu)\| \lesssim N^l\|P_Nu - I_Nu\| \lesssim N^{l-m}|u|_m.$$

Finally, using the triangle inequality and Theorem 2.1 yields

$$\|\partial_x^l(I_Nu - u)\| \le \|\partial_x^l(P_Nu - I_Nu)\| + \|\partial_x^l(P_Nu - u)\| \lesssim N^{l-m}|u|_m.$$

This ends the proof. □

We presented above some basic Fourier approximations in the Sobolev spaces. The interested readers are referred to the books on Fourier analysis (see, e.g., Körner (1988), Folland (1992)) for a thorough discussion on the Fourier approximations in different contexts.

2.3 Applications of Fourier Spectral Methods

In this section, we apply Fourier spectral methods to several nonlinear PDEs with periodic boundary conditions, including the Korteweg–de Vries (KdV) equation (cf. Korteweg and de Vries (1895)), the Kuramoto–Sivashinsky (KS) equation

(cf. Kuramoto and Tsuzuki (1976)) and the Allen–Cahn equation (cf. Allen and Cahn (1979)). The emphasis will be put on the treatment for nonlinear terms and time discretizations.

2.3.1 Korteweg–de Vries (KdV) Equation

The KdV equation is a celebrated mathematical model of waves on shallow water surfaces. A fascinating property of the KdV equation is that it admits soliton-type solutions (cf. Zabusky and Galvin (1971)). Consider the KdV equation in the whole space:

$$\begin{aligned} &\partial_t u + u\partial_y u + \partial_y^3 u = 0, \quad y \in (-\infty, \infty), \ t > 0, \\ &u(y,0) = u_0(y), \quad y \in (-\infty, \infty), \end{aligned} \tag{2.54}$$

which has the exact soliton solution

$$u(y,t) = 12\kappa^2 \operatorname{sech}^2\big(\kappa(y - y_0) - 4\kappa^3 t\big), \tag{2.55}$$

where y_0 is the center of the initial profile $u(y,0)$, and κ is a constant related to the traveling phase speed.

Since $u(y,t)$ decays exponentially to zero as $|y| \to \infty$, we can truncate the infinite interval to a finite one $(-\pi L, \pi L)$ with $L > 0$, and approximate the boundary conditions by the periodic boundary conditions on $(-\pi L, \pi L)$. It is expected that the initial-boundary valued problem (2.54) with periodic boundary conditions can provide a good approximation to the original initial-valued problem as long as the soliton does not reach the boundaries.

For convenience, we map the interval $[-\pi L, \pi L]$ to $[0, 2\pi]$ through the coordinate transform:

$$x = \frac{y}{L} + \pi, \quad y = L(x - \pi), \qquad x \in [0, 2\pi], \ y \in [-\pi L, \pi L],$$

and denote

$$v(x,t) = u(y,t), \quad v_0(x) = u_0(y). \tag{2.56}$$

The transformed KdV equation reads

$$\begin{aligned} &\partial_t v + \frac{1}{L} v\partial_x v + \frac{1}{L^3}\partial_x^3 v = 0, \quad x \in (0, 2\pi), \ t > 0, \\ &v(\cdot, t) \text{ periodic on } [0, 2\pi), \quad t \geq 0; \quad v(x,0) = v_0(x), \quad x \in [0, 2\pi]. \end{aligned} \tag{2.57}$$

Writing $v(x,t) = \sum_{|k|=0}^{\infty} \hat{v}_k(t) e^{\mathrm{i}kx}$, taking the inner product of the first equation with $e^{\mathrm{i}kx}$, and using (2.3) and the fact that $v\partial_x v = \frac{1}{2}\partial_x(v^2)$, we obtain that

$$\frac{d\hat{v}_k}{dt} - \frac{\mathrm{i}k^3}{L^3}\hat{v}_k + \frac{\mathrm{i}k}{2L}\widehat{(v^2)}_k = 0, \quad k = 0, \pm 1, \ldots, \tag{2.58}$$

with the initial condition

$$\hat{v}_k(0) = \frac{1}{2\pi}\int_0^{2\pi} v_0(x)e^{-\mathrm{i}kx}dx. \tag{2.59}$$

The ODE systems (2.58)–(2.59) can be solved by various numerical methods such as the Runge–Kutta methods or the semi-implicit/linearly implicit schemes in which the nonlinear terms are treated explicitly while the leading linear term is treated implicitly.

Here, we use a combination of the integrating factor and Runge–Kutta methods as suggested in Trefethen (2000). More precisely, multiplying (2.58) by the integrating factor $e^{-\mathrm{i}k^3t/L^3}$, we can rewrite the resulting equation as

$$\frac{d}{dt}\left[e^{-\mathrm{i}k^3t/L^3}\hat{v}_k\right] = -\frac{\mathrm{i}k}{2L}e^{-\mathrm{i}k^3t/L^3}\widehat{(v^2)}_k, \quad k = 0, \pm 1, \ldots. \tag{2.60}$$

Such a treatment makes the linear term disappear and can relax the stiffness of the system. The system (2.60) can then be solved by a standard ODE solver.

We now describe the Fourier approximation of (2.57) in MATLAB. Let $\mathbf{k}$ be the vector as in (2.40), and denote

$$\tilde{\mathbf{v}} = \left(\tilde{v}_0, \ldots, \tilde{v}_{N/2}, \tilde{v}_{-N/2+1}, \ldots, \tilde{v}_{-1}\right), \ \mathbf{g} = e^{-\mathrm{i}\mathbf{k}^3t/L^3}, \ \tilde{\mathbf{u}} = \mathbf{g}.*\tilde{\mathbf{v}},$$

where the operations on the vectors are component-wise. Then, the Fourier approximation scheme based on (2.60) is as follows:

$$\frac{d\tilde{\mathbf{u}}}{dt} = -\frac{\mathrm{i}\mathbf{k}}{2L}.*\mathbf{g}.*\mathtt{fft}\left(\left[\mathtt{ifft}\left(\mathbf{g}^{-1}.*\tilde{\mathbf{u}}\right)\right]^2\right), \quad t > 0. \tag{2.61}$$

Therefore, a Runge–Kutta method, such as the fourth-order MATLAB routine rk4.m, can be directly applied to (2.61).

We present below some numerical results obtained by Program 27 (with some minor modifications) in Trefethen (2000). We first take $\kappa = 0.3$, $y_0 = -20$ and $L = 15$. On the left of Fig. 2.2, we plot the time evolution of the approximate solution (with $N = 256$, time step size $\tau = 0.01$ and $t \in [0, 60]$), and on the right, we plot the maximum errors at $t = 1, 30, 60$ for various N with $\tau = 0.001$. Observe that the errors decay like $O(e^{-cN})$, which is typical for smooth solutions. The superior accuracy for this soliton solution indicates that the KdV equation on a finite interval can be used to effectively simulate the KdV equation on the whole space before the solitary wave reaches the boundaries.

In the next example, we consider the interaction of five solitary waves. More precisely, we consider the KdV equation (2.54) with the initial condition which consists of five solitary waves,

$$u_0(y) = \sum_{j=1}^{5} 12\kappa_j^2 \mathrm{sech}^2\left(\kappa_j(y - y_0)\right), \tag{2.62}$$

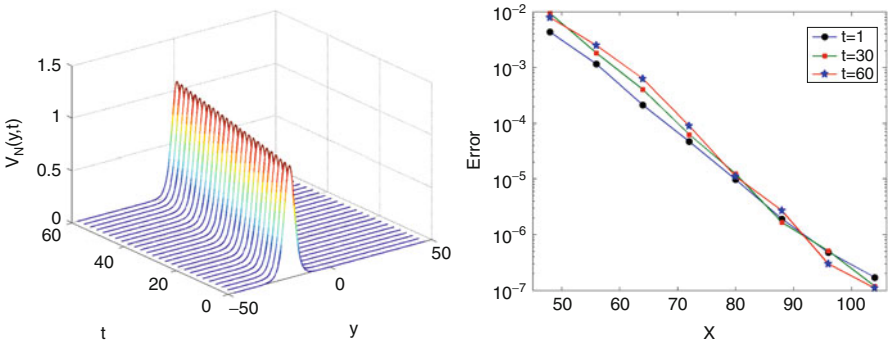

Fig. 2.2 *Left*: time evolution of the numerical solution; *right*: maximum errors vs. N

with

$$\begin{aligned} &\kappa_1 = 0.3,\ \ \kappa_2 = 0.25,\ \ \kappa_3 = 0.2,\ \ \kappa_4 = 0.15,\ \ \kappa_5 = 0.1, \\ &y_1 = -120,\ \ y_2 = -90,\ \ y_3 = -60,\ \ y_4 = -30,\ \ y_5 = 0. \end{aligned} \tag{2.63}$$

We fix $L = 50, N = 512$ and $\tau = 0.01$. In Fig. 2.3, we plot the time evolution of the approximate solution for $t \in [0, 600]$, and depict the initial profile and final profile at $t = 600$ in Fig. 2.4. We observe that the soliton with large amplitude travels with a faster speed, and the amplitudes of the five solitary waves are well preserved at the final time. This indicates the scheme has an excellent conservation property.

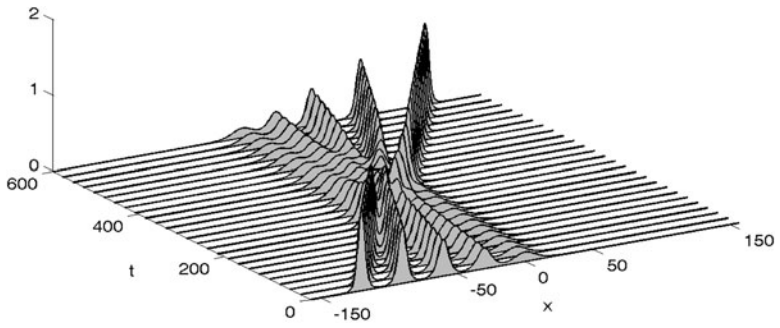

Fig. 2.3 Interaction of five solitary waves

2.3.2 Kuramoto–Sivashinsky (KS) Equation

The KS equation has been used in the study of a variety of reaction-diffusion systems (cf. Kuramoto and Tsuzuki (1976)), and is also an interesting dynamical PDE that can exhibit chaotic solutions (cf. Hyman and Nicolaenko (1986), Nicolaenko et al. (1985)).

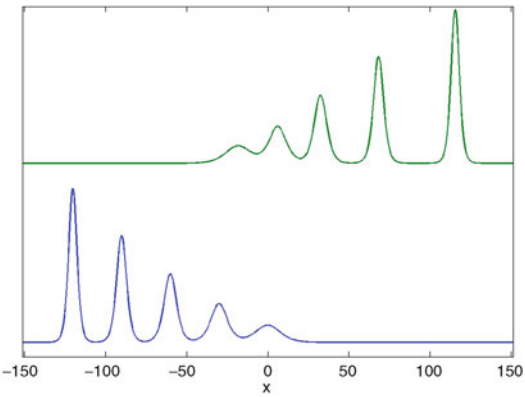

Fig. 2.4 Profiles at $t = 0, 600$

We consider the KS equation of the form

$$\begin{aligned} &\partial_t u + \partial_x^4 u + \partial_x^2 u + uu_x = 0, \quad x \in (-\infty, \infty), \ t > 0, \\ &u(x,t) = u(x + 2L\pi, t), \ \partial_x u(x,t) = \partial_x u(x + 2L\pi, t), \quad t \geq 0, \\ &u(x,0) = u_0(x), \quad x \in (-\infty, \infty), \end{aligned} \tag{2.64}$$

where the given function u_0 is $2L\pi$-periodic.

Thanks to the periodicity, it suffices to consider (2.64) in the reference interval $[0, 2L\pi]$. We discretize (2.64) in space by seeking the approximate solution

$$u_N(x,t) = \sum_{k=-N/2}^{N/2} \tilde{u}_k(t) e^{\mathrm{i}kx/L}, \quad t > 0, \tag{2.65}$$

where $\tilde{u}_{N/2}(t) = \tilde{u}_{-N/2}(t)$. The N independent frequencies are determined by the scheme

$$\partial_t u_N + \partial_x^4 u_N + \partial_x^2 u_N = -\frac{1}{2} \partial_x I_N \left(u_N^2 \right), \quad t > 0, \tag{2.66}$$

where I_N is the interpolation operator associated with the grids $\left\{ x_j = 2L\pi j/N \right\}_{j=0}^{N-1}$. Thus, for each frequency k, we have

$$\tilde{u}_k'(t) + \left(\frac{k^4}{L^4} - \frac{k^2}{L^2} \right) \tilde{u}_k(t) = -\frac{1}{2L} \mathrm{i}k \tilde{w}_k(t), \quad t > 0, \tag{2.67}$$

where $\tilde{w}_k(t)$ is the discrete Fourier coefficient of the nonlinear term, i.e.,

$$I_N \left(u_N^2 \right) = \sum_{k=-N/2}^{N/2} \tilde{w}_k(t) e^{\mathrm{i}kx/L}. \tag{2.68}$$

It is important to point out that, using the Fourier transform, linear operators with constant coefficients can always be diagonalized (i.e., the frequencies are separable) like (2.67). This leads to efficient time integrations for the resulting equation in the frequency space. We refer to Kassam and Trefethen (2005) for a review of various time-stepping schemes. Here, we use an exponential time-differencing (ETD) method as suggested in Kassam and Trefethen (2005).

Denote by $\tilde{\boldsymbol{u}}(t)$ the vector of the expansion coefficients as arranged in Table 2.1. Let $\boldsymbol{L}$ be the diagonal matrix with the diagonal $\boldsymbol{k}^2/L^2 - \boldsymbol{k}^4/L^4$, where $\boldsymbol{k}$ is the index vector given by Table 2.1, and $\boldsymbol{N}(t) := \boldsymbol{N}(\tilde{\boldsymbol{u}}, t)$ be the vector of the nonlinear term in (2.67)–(2.68). Then, we can rewrite (2.67)–(2.68) as a nonlinear ODE system

$$\tilde{\boldsymbol{u}}'(t) = \boldsymbol{L}\tilde{\boldsymbol{u}}(t) + \boldsymbol{N}(t), \quad t > 0. \tag{2.69}$$

Let τ be the time step size. It is clear that (2.69) is equivalent to

$$\tilde{\boldsymbol{u}}(t_n + \tau) = e^{\boldsymbol{L}\tau}\tilde{\boldsymbol{u}}(t_n) + e^{\boldsymbol{L}\tau}\int_0^\tau e^{-\boldsymbol{L}s}\boldsymbol{N}\big(\tilde{\boldsymbol{u}}(t_n + s), t_n + s\big)ds. \tag{2.70}$$

Based on how one approximates the integral, various ETD schemes may be constructed. For example, let $\tilde{\boldsymbol{u}}_n$ be the approximation of $\tilde{\boldsymbol{u}}(t_n)$. The following modified fourth-order ETD Runge–Kutta (ETDRK4) has been shown to be a very stable and accurate scheme for stiff equations (cf. Kassam and Trefethen (2005)):

$$\begin{aligned}
\boldsymbol{a}_n &= e^{\boldsymbol{L}\tau/2}\tilde{\boldsymbol{u}}_n + \boldsymbol{L}^{-1}\big(e^{\boldsymbol{L}\tau/2} - \boldsymbol{I}\big)\boldsymbol{N}(\tilde{\boldsymbol{u}}_n, t_n),\\
\boldsymbol{b}_n &= e^{\boldsymbol{L}\tau/2}\tilde{\boldsymbol{u}}_n + \boldsymbol{L}^{-1}\big(e^{\boldsymbol{L}\tau/2} - \boldsymbol{I}\big)\boldsymbol{N}(\boldsymbol{a}_n, t_n + \tau/2),\\
\boldsymbol{c}_n &= e^{\boldsymbol{L}\tau/2}\boldsymbol{a}_n + \boldsymbol{L}^{-1}\big(e^{\boldsymbol{L}\tau/2} - \boldsymbol{I}\big)\big[2\boldsymbol{N}(\boldsymbol{b}_n, t_n + \tau/2) - \boldsymbol{N}(\tilde{\boldsymbol{u}}_n, t_n)\big],\\
\tilde{\boldsymbol{u}}_{n+1} &= e^{\boldsymbol{L}\tau}\tilde{\boldsymbol{u}}_n + \Big\{\alpha\boldsymbol{N}(\tilde{\boldsymbol{u}}_n, t_n) + 2\beta\big[\boldsymbol{N}(\boldsymbol{a}_n, t_n + \tau/2)\\
&\qquad + \boldsymbol{N}(\boldsymbol{b}_n, t_n + \tau/2)\big] + \gamma\boldsymbol{N}(c_n, t_n + \tau)\Big\},
\end{aligned} \tag{2.71}$$

where the coefficients

$$\begin{aligned}
\alpha &= \tau^{-2}\boldsymbol{L}^{-3}\Big[-4 - \boldsymbol{L}\tau + e^{\boldsymbol{L}\tau}\big(4 - 3\boldsymbol{L}\tau + (\boldsymbol{L}\tau)^2\big)\Big],\\
\beta &= \tau^{-2}\boldsymbol{L}^{-3}\Big[2 + \boldsymbol{L}\tau + e^{\boldsymbol{L}\tau}\big(-2 + \boldsymbol{L}\tau\big)\Big],\\
\gamma &= \tau^{-2}\boldsymbol{L}^{-3}\Big[-4 - 3\boldsymbol{L}\tau + e^{\boldsymbol{L}\tau}\big(4 - \boldsymbol{L}\tau\big) - (\boldsymbol{L}\tau)^2\Big].
\end{aligned} \tag{2.72}$$

In the following computations, we take $L = 16$ and impose the initial condition:

$$u_0(x) = \cos(x/L)(1 + \sin(x/L))$$

as in Kassam and Trefethen (2005). In Fig. 2.5, we depict the time evolution of the KS equation (2.64) obtained by the above algorithm with $\tau = 10^{-4}$ and $N = 128$. We plot in Fig. 2.6 the profiles of the numerical solution at various time in the waterfall format. We can observe the same pattern of the deterministic chaos as illustrated in Kassam and Trefethen (2005).

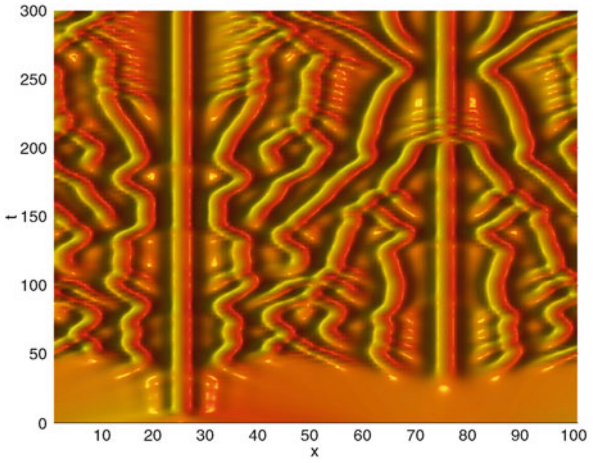

Fig. 2.5 Time evolution of the KS equation. Time runs from 0 at the bottom to 300 at the top

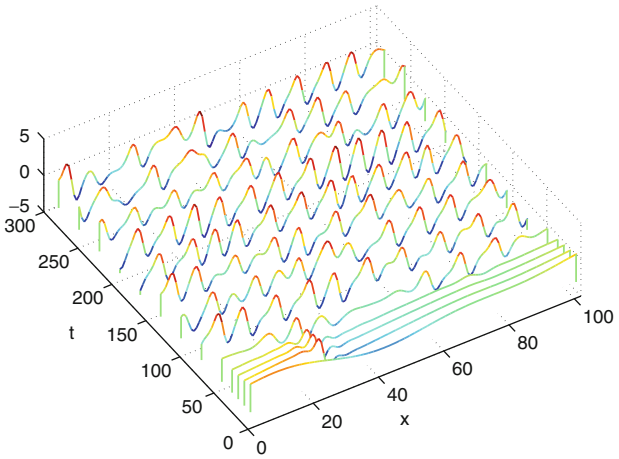

Fig. 2.6 Waterfall plot of the profiles of the numerical solution

2.3.3 Allen–Cahn Equation

The Allen–Cahn equation was originally introduced in Allen and Cahn (1979) to describe the motion of anti-phase boundaries in crystalline solids. It has been widely used in material science applications.

We consider the two-dimensional Allen–Cahn equation with periodic boundary conditions:

$$\begin{aligned} &\partial_t u - \varepsilon^2 \Delta u + u^3 - u = 0, \quad (x,y) \in \Omega = (-1,1)^2,\ t > 0, \\ &u(-1,y,t) = u(1,y,t), \quad u(x,-1,t) = u(x,1,t), \quad t \geq 0, \\ &u(x,y,0) = u_0(x,y), \quad (x,y) \in \bar{\Omega}, \end{aligned} \tag{2.73}$$

where ε is a small parameter which describes the inter-facial width. We refer to Sect. 9.3 for a more thorough discussion on the Allen–Cahn equation.

Let us write the Fourier approximation of the solution as

$$u_N(x,y,t) = \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{kl}(t) e^{\mathrm{i}(kx+ly)\pi}, \tag{2.74}$$

and denote by u_N^m the approximation of u_N at time $t_m = m\tau$ with τ being the time step size. Then a second-order stabilized semi-implicit scheme in time is (cf. Shen and Yang (2010)):

$$\begin{aligned} &\frac{3u_N^{m+1} - 4u_N^m + u_N^{m-1}}{2\tau} - \varepsilon^2 \Delta u_N^{m+1} + \big(2F_N(u_N^m) - F_N(u_N^{m-1})\big) \\ &\quad + s\big(u_N^{m+1} - 2u_N^m + u_N^{m-1}\big) = 0, \quad m = 1,2,\ldots, \end{aligned} \tag{2.75}$$

where $s > 0$ is an adjustable parameter, and $F_N(v) = I_N(v^3) - v$ with I_N being the two-dimensional tensorial interpolation operator on the computational grid. Notice that the extra dissipative term $s\big(u_N^{m+1} - 2u_N^m + u_N^{m-1}\big)$ (of order $s\tau^2$) is added to improve the stability while preserving the simplicity. At each time step, we only need to solve the linear problem

$$-2\tau\varepsilon^2 \Delta u_N^{m+1} + (3 + 2s\tau) u_N^{m+1} = \boldsymbol{N}\big(u_N^m, u_N^{m-1}\big), \tag{2.76}$$

where

$$\begin{aligned} \boldsymbol{N}\big(u_N^m, u_N^{m-1}\big) &= 4\big(1 + \tau s + \tau\big) u_N^m - \big(1 + 2\tau s + 2\tau\big) u_N^{m-1} \\ &\quad - 2\tau I_N\big[2(u_N^m)^3 - (u_N^{m-1})^3\big]. \end{aligned} \tag{2.77}$$

Applying the Fourier Galerkin method yields the equations in the frequency space:

$$\big(2\tau\varepsilon^2(k^2 + l^2)\pi^2 + 3 + 2s\tau\big)\tilde{u}_{kl}^{m+1} = \tilde{w}_{kl}, \tag{2.78}$$

where $\{\tilde{w}_{kl}\}$ are the discrete Fourier coefficients of the nonlinear term $\boldsymbol{N}\big(u_N^m, u_N^{m-1}\big)$. The above semi-implicit scheme leads to an efficient implementation with the main cost coming from the treatment for the nonlinear term, which can be manipulated by FFT through a pseudo-spectral approach.

To test the numerical scheme, we consider the motion of a circular interface and impose the initial condition:

$$u_0(x,y) = \begin{cases} 1, & (x-0.5)^2 + (y-0.5)^2 < 1, \\ -1, & \text{otherwise.} \end{cases} \tag{2.79}$$

The motion of the interface is driven by the mean curvature of the circle, so the circle will shrink and eventually disappear, see Fig. 2.7. Note that the rate at which the diameter of the circle shrinks can be determined analytically (cf. Chen and Shen (1998)).

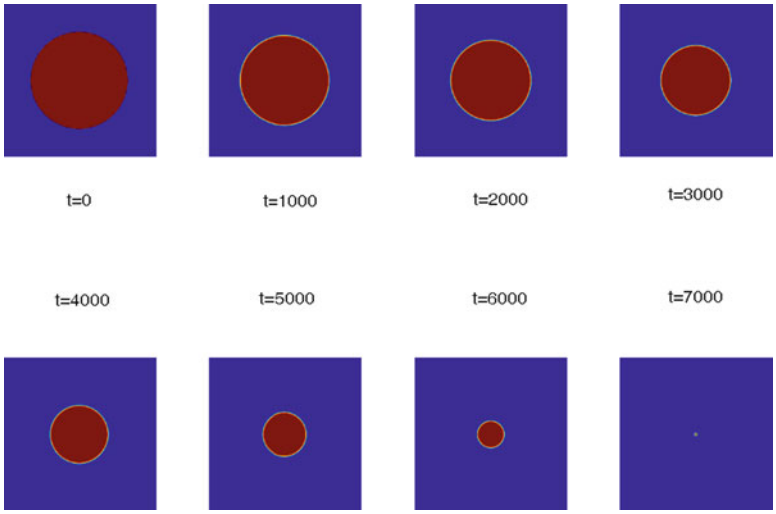

Fig. 2.7 Time evolution of a circular domain

Problems

2.1. Let $\mathscr{D}_N(x)$ be the Dirichlet kernel defined in (2.7).
(a) Show that $\mathscr{D}_N(x)$ is an even function, and it is symmetric about $x = 1/2$, namely,

$$\mathscr{D}_N(-x) = \mathscr{D}_N(x), \quad \mathscr{D}_N(1/2+x) = \mathscr{D}_N(1/2-x),$$

(b) Show that

$$\frac{1}{2\pi}\int_0^{2\pi} \mathscr{D}_N(x)dx = 1,$$

and

$$\frac{1}{2\pi}\int_0^{2\pi} \mathscr{D}_N^2(x)dx = 2N+1.$$

(c) Show that

$$\int_0^{2\pi} |\mathscr{D}_N(x)|dx \le c\ln N, \quad N \ge 2,$$

where c is a positive constant independent of N.
(d) Prove that for any $\phi \in X_N$ (defined in (2.41)),

$$\phi(x) = \frac{1}{2\pi}\int_0^{2\pi} \phi(t)\mathscr{D}_N(x-t)dt,$$

and

$$\|\phi\|_\infty \le \sqrt{2N+1}\|\phi\|.$$

2.2. The Fejér kernel is defined as the Nth arithmetic mean of the Dirichlet kernels:

$$F_N(x) = \frac{1}{N} \sum_{n=0}^{N-1} \mathscr{D}_n(x).$$

Show that

$$F_N(x) = \frac{\sin^2\left(Nx/2\right)}{N \sin^2\left(x/2\right)}.$$

2.3. Use the Sobolev inequality (B.33) and Theorem 2.1 to prove Theorem 2.2 with $m \geq 1$ in place of $m > 1/2$.

2.4. Determine $D^{(2)}$ and D^2 with $N = 4$ and confirm that $D^2 \neq D^{(2)}$.

2.5. Derive the formula (2.37) for the entries of the second-order differentiation matrix $D^{(2)}$.

2.6. Describe and implement a fourth-order Runge–Kutta and Fourier method for the Burger equation with periodic boundary conditions:

$$u_t = \varepsilon u_{xx} + u u_x, \quad x \in (-\pi, \pi); \quad u(x,0) = e^{-10\sin^2(x/2)},$$

with $\varepsilon = 0.03$ and the simulation running up to $t = 1$.