Notes from *Spectral Methods: Algorithms, Analysis, and Applications*

by Jie Shen, Tao Tang, Li-Lian Wang

taken by Samuel T. Wallace

Publisher's Description

Along with finite differences and finite elements, spectral methods are one of the three main methodologies for solving partial differential equations on computers. This book provides a detailed presentation of basic spectral algorithms, as well as a systematical presentation of basic convergence theory and error analysis for spectral methods. Readers of this book will be exposed to a unified framework for designing and analyzing spectral algorithms for a variety of problems, including in particular high-order differential equations and problems in unbounded domains. The book contains a large number of figures which are designed to illustrate various concepts stressed in the book. A set of basic matlab codes has been made available online to help the readers to develop their own spectral codes for their specific applications.

A Note From the Transcriber

These notes were taken over summer 2020 as part of self-study preparing for PhD in applied math and numerical analysis. I am reading this without much interpolation theory knowledge, and some prior exposure to spectral methods through Trefethen's book *Spectral Methods in MATLAB* (sorry, no notes for that book). This book comes from a suggestion by a professor as my grad school, and it looked temptingly challenging.

Contents

0.1 Introduction

0.1.1 Weighted Residual Methods

Consider the following general problem:

$$\partial_t u(x,t) - \mathcal{L}u(x,t) = \mathcal{N}(u)(x,t), \quad t > 0, x \in \Omega \tag{1}$$

Where $\mathcal{L}$ is a leading spatial derivative operator, and $\mathcal{N}$ is a lower-order linear or non-linear operator involving only spatial derivatives. Here, Ω denotes a bounded domain of $\mathbb{R}^d$, $d = 1, 2,$ or 3. This equation 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, semi-implicit schemes or linearly-implicit schemes, in which the principal linear operators are treated *implicitly* to reduce the associated stability constraint, while the non-linear equations are treated explicitly to avoid the expensive process of solving nonlinear equations at each time step, are most frequently used int eh context of spectral methods.

Let τ be the 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 the equation:

$$\frac{u^{n+1} - u^{n-1}}{2\tau} - \mathcal{L}\left(\frac{u^{n+1} + u^{n-1}}{2}\right) = \mathcal{N}(u^n) \;\; n \geq 1 \tag{2}$$

We can rewrite this as

$$\mathbf{L}u(x) := \alpha u(x) - \mathcal{L}u(x) = f(x), \;\; x \in \Omega \tag{3}$$

where $u = \frac{u^{n+1}+u^{n+1}}{2}, \alpha = \tau^{-1}$ and $f = \alpha^{n-1} + \mathcal{N}(u^n)$. Hence, at each time step, we need to solve a steady-state problem of the form of (3).

At this point, it is important to emphasize that the construction of efficient numerical solvers for some important equations in the form of (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 for equations of the form (3) where $\mathcal{L}$ is a *linear elliptic* operator.

The starting point of the WRM is to approximate the solution u is to approximate (3) by a finite sum

$$u(x) \approx u_N(x) = \sum_{k=0}^{N} a_k \phi_k(x) \tag{4}$$

where $\{\phi_k\}$ are the *trial (or basis) functions*, and the expansion coefficients are to be determined. Substituting u_N for u in (3) leads to the *residual*

$$\mathbf{R}_N(x) = \mathbf{L}u_N(x) - f(x) \neq 0 \ x \in \Omega \tag{5}$$

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

$$(\mathbf{R}_N, \psi_j)_\omega = \int_\Omega \mathbf{R}_N(x)\psi_j(x)\omega(x)dx = 0, \ 0 \leq j \leq N \tag{6}$$

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

$$\langle \mathbf{R}_N, \psi_j \rangle_{N,\omega} := \sum_{k=0}^{N} \mathbf{R}_N(x_k)\psi_j(x_k)\omega_k = 0, \ 0 \leq j \leq N \tag{7}$$

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-elements 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 of 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)
- $\phi_k = L_k(x)$ (Legendre spectral method)
- $\phi_k = \mathcal{L}_k(x)$ (Laguerre spectral method)
- $\phi_k(x) = H_k(x)$ (Hermite spectral method)

Here, $T_k, L_k, \mathcal{L}_k,$ 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 (6) or (7)), assuming the boundary conditions are periodic of homogeneous.
- *Petrov-Galerkin.* The test functions are different from the trial ones.
- *Collocation.* The test functions $\{\psi_k\}$ in (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 each x_j , i.e. $\mathbf{R}_N(x_j) = 0$.

Remark 0.1.1. *In the literature, the term of pseudo-spectral methods is often used to describe any spectral method where some operations involve a collocation approach or a numerical quadrature which produces aliasing errors. In this sense, almost all practical spectral methods are pseudo-spectral. In this book, we shall not classify a method as pseudo-spectral of spectral. Instead, it will be classified as Galerkin type or collocation type.*

Remark 0.1.2. *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 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 formulas in a general manner. We shall revisit these illustrative examples in a more rigorous fashion in the main body of the book.

0.1.2 Spectral-Collocation Method

To fix the idea, let us consider the following linear problem:

$$\mathbf{L}u(x) = -u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \ \ x \in (-1, 1) \tag{8}$$

$$B_{\pm}u(\pm 1) = g_{\pm} \tag{9}$$

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

As mentioned above, 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 (8) amounts to finding $u_N \in P_N$ such that

1. the residual $\mathbf{R}_N(x_k) = \mathbf{L}u_N(x_k) - f(x_k) = 0, \ 1 \leq K \leq N - 1$

2. u_N satisfies exactly the boundary conditions, i.e.,

$$B_- u_N(x_0) = g_-, \;\; B_+ u_N(x_N) = g_+ \tag{10}$$

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

$$u_N(x) = \sum_{j=0}^{N} u_N(x_j) h_j(x) \tag{11}$$

where $\{h_j\}$ are the Lagrange basis polynomials (also referred to as *nodal* basis functions), i.e., $h_j \in P_N$ and $h_j(x_k) = \delta_{jk}$. Hence, in inserting (11) into (9)-(10) leads to the linear system

$$\sum_{j=0}^{N} \left[\mathbf{L} h_j(x_k)\right] u_N(x_j) = f(x_k), \tag{12}$$

$$\sum_{j=0}^{N} \left[\mathbf{L} h_j(x_k)\right] u_N(x_j) = f(x_k), \;\; 1 \le k \le N-1 \tag{13}$$

$$\sum_{j=0}^{N} \left[B_- j_j(x_0)\right] u_N(x_j) = g_-, \sum_{j=0}^{N} \left[B_+ j_j(x_N)\right] u_N(x_j) = g_+ \tag{14}$$

The above system contains $N+1$ unknowns, so we can rewrite it in a matrix form. To fix the idea, we consider (8) with Dirchlet boudnary 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 (12) reduces to

$$\sum_{j=1}^{N-1} \left[\mathbf{L} h_j(x_k)\right] u_N(x_j) = f(x_k) - \left\{\left[\mathbf{L} h_0(x_k)\right] g_- + \left[\mathbf{L} h_N(x_k)\right] g_+\right\} \tag{15}$$

for $1 \le K \le N-1$. Differentiating (11) m times leads to

$$u_N^{(m)}(x_k) = \sum_{j=0}^{N} d_{kj}^{(m)} u_N(x_j) \tag{16}$$

where

$$d_{kj}^{(m)} = h_j^{(m)}(x_k) \tag{17}$$

The matrix $D^{(m)} \left(d_{kj}^{(m)}\right)_{k,j=0\ldots N}$ is called the differentiation matrix of order m relative to the $\{s_j\}_{j=0}^{N}$. 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 (14) that

$$\mathbf{u}^{(m)} = D^{(m)} \mathbf{u}^{(0)}, \;\; m \ge 1 \tag{18}$$

Hence, we have

$$\mathbf{L}h_j\left(x_k\right) = -d_{kj}^{(2)} + p(x_k)d_{kj}^{(1)} + q\left(x_k\right)\delta_{kj} \tag{19}$$

Denote by $\mathbf{f}$ the vector with $N-1$ components given by the right-handside of (13). Setting

$$\tilde{D}_m = \left(d_{kj}^{(m)}\right)_{kj=1,\ldots,N}, \ m = 1, 2 \tag{20}$$

$$P = \operatorname{diag}\left(p(x_1), \ldots, p(x_{N-1})\right), \ Q = \operatorname{diag}\left(q(x_1), \ldots, q(x_{N-1})\right)$$

the system (13) reduces to

$$\left(-\tilde{D}_2 + P\tilde{D}_1 + Q\right)\mathbf{u}^{(0)} = \mathbf{f} \tag{21}$$

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 valuated by (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 $O\left(N^{2m}\right)$ (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 boudnary 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.

0.1.3 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 hte basic principles of the spectral-Galerkin method, spectral-Petrov-Galerkin method, and spectral-Galerkin method with numerical integration.

Galerkin Method

Without loss of generality, we consider (8) with $g_{\pm} = 0$. The non homogeneous boundary conditions can be easily handled by considering $\nu = u - \tilde{u}$, where $\tilde{u}$ is a "simple" function satisfying the non homogeneous boundary conditions (cf. Chap. 4).

Define the finite-dimensional approximation space:

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

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

$$u_N(x) = \sum_{k=0}^{N-2} \widehat{u}_k \phi_k(x) \in X_N \tag{22}$$

Then, the expansion coefficients $\{\widehat{u}_k\}_{k=0}^{N-2}$ can be determined by by the residual equation (6) with $\{\psi_{\mathbf{j}} = \phi_{\mathbf{j}}\}$:

$$\int_{-1}^{1} (\mathbf{L}u_N(x) - f(x))\,\phi_j(x)\omega(x)dx = 0,\ 0 \le j \le N-2 \tag{23}$$

which is equivalent to finding $u_N \in X_N$ such that

$$(\mathbf{L}u_N, \nu_N)_{\omega} = (f, \nu_N)_{\omega},\ \forall \nu_N \in X_N \tag{24}$$

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

The linear system of the above scheme is obtained by substituting by substituting (19) into (20). More precisely, setting

$$\mathbf{u} = (\widehat{u}_0, \widehat{u}_1, \ldots, \widehat{u}_{N-2})^T$$

$$f_j = (f, \phi_j)_{\omega}$$

$$\mathbf{f} = (f_0, f_1, \ldots, f_{N-2})^T$$

$$s_{jk} = (\mathbf{L}\phi_k, \phi_j)_{\omega}$$

$$S = (s_{jk})_{j,k=0,\ldots,N-2}$$

the system (20) reduces to

$$S\mathbf{u} = \mathbf{f}$$

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 (22) can be solved efficiently.

The key idea is to used *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 (20)-(22)). Let $L_k(x)$ be the Legendre polynomial of degree k , and set

$$\phi_k(x) = L_k(x) + \alpha_k L_{k+1}(x) + \beta_k L_{k+2}(x), \ k \geq 0 \tag{25}$$

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

$$X_N = \text{span}\{\phi_0, \phi_1, \ldots, \phi_{N-2}\} \tag{26}$$

Using the properties of the 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 (22) can be solved efficiently. However, for more general $p(x)$ and $q(x)$, the coefficient matrix S is full an done needs to resort to an iterative method (cf. Sect. 4.4).

In the case above, we just considered the Legendre case. In fact, th 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

$$(I_N f)(x) = \sum_{k=0}^{N} \tilde{f}_k \phi_k(x) \tag{27}$$

Where $\{\phi_k\}$ is an orthonormal polynomial of P_N (orthogonal with respect to ω , i.e. $(\phi_k, \phi_j)_\omega = \delta_{jk}$). Thanks to orthogonality, the *discrete transforms* between the physical values $\{f(x_j)\}_{j=0}^N$ and the expansion coefficients $\{f(x_j)\}_{j=0}^N$ can be computed efficiently. In particular, the computational complexity of using 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 Sec. 3.1.5.

It is important to point out that insolving time-dependent nonlinear problems, f usually contains nonlinear terms involving derivatives of hte numerical solution u_N at previous time steps (cf. (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.

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:

$$\mathbf{L}u(x) := u'''(x) + u(x) = f(x) \;\; x \in (-1,1) \tag{28}$$

$$u(\pm 1) = u'(1) = 0$$

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

$$X_N = \{\phi \in P_N | \phi(\pm 1) = \phi'(1) = 0\} \Rightarrow \dim(X_N) = N-2$$

Assuming that $\{\phi_k\}_{k=0}^{N-3}$ is a basis of X_N are dtermined by the residual equation (6) (with $\omega = 1$):

$$\int_{-1}^{1} (\mathbf{L}u_N(x) - f(x))\, \psi_j(x) dx = 0, \;\; 0 \le j \le N-3 \tag{29}$$

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

$$X_N^* = \{\psi \in P_N | \psi(\pm 1) = \psi'(-1) = 0\} \Rightarrow \dim(X_N^*) = N-2$$

Assume that $\{\psi_k\}_{k=0}^{N-3}$ is a basis of X_N^* . Then, (27) is equivalent to the variational formulation:

Find $u_N \in X_N$ such that

$$(\mathbf{L}u_N, \nu_N) = (f, \nu_N) \;\; \forall \nu_N \in X_N^*$$

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

$$\mathbf{u} = (\widehat{u}_0, \widehat{u}_1, \ldots, \widehat{u}_{N-3})^T$$

$$f_j = (f, \psi_j)$$

$$\mathbf{f}\, (f_0, f_1, \ldots, f_{N-3})^T ;$$

$$s_{jk} = \left(\phi_k', \psi_j''\right)$$

$$S = (s_{jk})_{j,k=0,\ldots,N-3}$$

$$m_{jk} = (\phi_k, \psi_j)$$

$$M = (m_{jk})_{j,k=0,\ldots,N-3}$$

The linear system becomes

$$(S + M)\mathbf{u} = \mathbf{f} \tag{30}$$

As described in the previous section, we wish to construct basis functions for X_N and X_N^* so that the linear system (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 $0 \le k \le N-3$,

$$\phi_k = L_k - \frac{2k+3}{2k+5}L_{k+1} - L_{k+2} + \frac{2k+3}{2k+5}L_{k+3} \in X_N$$

$$\psi_k = L_k + \frac{2k+3}{2k+5}L_{k+1} - L_{k+2} - \frac{2k+3}{2k+5}L_{k+3} \in X_N^*$$

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.

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 (8) with $g_\pm = 0$. The spectral method with numerical integration is

$$\text{Find } u_N \in X_N := \{\phi \in P_N | B_\pm \phi(\pm 1) = 0\} \text{ such that}$$

$$a_N(u_N, \nu_N) := \langle Lu_N, \nu_N \rangle_N = \langle f, \nu_N \rangle_N, \ \forall \nu_N \in X_N \tag{31}$$

where the discrete inner product is defined by

$$\langle u, v \rangle_N = \sum_{j=0}^{N} u(x_j) v_N(x_j) \omega_j \tag{32}$$

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 methods is easier to implement, thanks to the discrete inner product, than the spectral-Galerkin method (21). It is also more convenient for error analysis, thanks to the weak formulation, than the spectral-collocation method (12).

We note than in the particular case of homogeneous Dirichlet boundary conditions, i.e. $B_\pm u(\pm 1) = u(\pm 1) = 0$, by taking $\nu_N = h_j, 1 \leq j \leq N-1$ in (31) and using the exactness of Legendre-Gauss-Lobatto quadrature, i.e.,

$$\langle u, v \rangle_N = (u, v)\,, \ \forall u \cdot v \in P_{2N-1} \tag{33}$$

we find that the formulation (31) is equivalent to the collocation formulation (12). However, this is not true for general boundary conditions (see Chap. 4).

0.1.4 Fundamental Tools for Error Analysis