Chapter 4
Spectral Methods for Second-Order Two-Point Boundary Value Problems

We consider in this chapter spectral algorithms for solving the two-point boundary value problem:

$$-\varepsilon U'' + p(x)U' + q(x)U = F, \quad \text{in } I := (-1,1), \tag{4.1}$$

(where $\varepsilon > 0$) with the general boundary conditions

$$a_-U(-1) + b_-U'(-1) = c_-, \quad a_+U(1) + b_+U'(1) = c_+, \tag{4.2}$$

which include in particular the Dirichlet boundary conditions ($a_\pm = 1$ and $b_\pm = 0$), Neumann boundary conditions ($a_\pm = 0$ and $b_\pm = \pm 1$), and Robin (or mixed) boundary conditions ($a_- = b_+ = 0$ or $a_+ = b_- = 0$). Whenever possible, we shall give a uniform treatment for all these boundary conditions. Without loss of generality, we assume that:

$$\begin{aligned} &\text{(i)}\ \ a_\pm \geq 0;\\ &\text{(ii)}\ \ a_-^2 + b_-^2 \neq 0,\ a_-b_- \leq 0;\ \ a_+^2 + b_+^2 \neq 0,\ a_+b_+ \geq 0;\\ &\text{(iii)}\ \ q(x) - p'(x)/2 \geq 0,\ \ \forall x \in I;\\ &\text{(iv)}\ \ p(1) > 0 \text{ if } b_+ \neq 0;\ \ p(-1) < 0 \text{ if } b_- \neq 0. \end{aligned} \tag{4.3}$$

The above conditions are necessary for the well-posedness of (4.1)–(4.2).

Let us first reduce the problem (4.1)–(4.2) to a problem with homogeneous boundary conditions.

- *Case I.* $a_\pm = 0$ and $b_\pm \neq 0$
  We set $\tilde{u} = \beta x^2 + \gamma x$, where β and γ are uniquely determined by asking $\tilde{u}$ to satisfy (4.2), namely,

$$\begin{aligned} -2b_-\beta + b_-\gamma &= c_-,\\ 2b_+\beta + b_+\gamma &= c_+. \end{aligned}$$

J. Shen et al., *Spectral Methods: Algorithms, Analysis and Applications*,
Springer Series in Computational Mathematics 41, DOI 10.1007/978-3-540-71041-7_4,

- *Case II.* $a_-^2 + a_+^2 \neq 0$
  We set $\tilde{u} = \beta x + \gamma$, where β and γ again can be uniquely determined by requiring $\tilde{u}$ to satisfy (4.2). Indeed, we have

$$\begin{aligned}(-a_- + b_-)\beta + a_-\gamma &= c_-,\\ (a_+ + b_+)\beta + a_+\gamma &= c_+.\end{aligned}$$

  The determinant of the coefficient matrix is

$$\text{DET} = -2a_-a_+ + a_+b_- - a_-b_+.$$

  The assumption (4.3) implies that $b_- \leq 0$ and $b_+ \geq 0$, so we have DET < 0.

Now, we set

$$u = U - \tilde{u}, \quad f = F - (-\varepsilon\tilde{u}'' + p(x)\tilde{u}' + q(x)\tilde{u}).$$

Then u satisfies the following equation

$$-\varepsilon u'' + p(x)u' + q(x)u = f, \quad \text{in } I = (-1,1), \tag{4.4}$$

with the homogeneous boundary condition

$$a_-u(-1) + b_-u'(-1) = 0, \quad a_+u(1) + b_+u'(1) = 0. \tag{4.5}$$

Let us denote

$$H^1_\diamond(I) = \{u \in H^1(I) : u(\pm 1) = 0 \text{ if } b_\pm = 0\}, \tag{4.6}$$

and

$$h_- = \begin{cases} 0, & \text{if } a_-b_- = 0,\\ \frac{a_-}{b_-}, & \text{if } a_-b_- \neq 0,\end{cases} \quad h_+ = \begin{cases} 0, & \text{if } a_+b_+ = 0,\\ \frac{a_+}{b_+}, & \text{if } a_+b_+ \neq 0.\end{cases} \tag{4.7}$$

Then, a standard weak formulation for (4.4)-(4.5) is:

$$\begin{cases} \text{Find } u \in H^1_\diamond(I) \text{ such that}\\ \mathscr{B}(u,v) = (f,v), \quad \forall v \in H^1_\diamond(I),\end{cases} \tag{4.8}$$

where

$$\begin{aligned}\mathscr{B}(u,v) := {} & \varepsilon(u',v') + \varepsilon h_+u(1)v(1) - \varepsilon h_-u(-1)v(-1)\\ & + (p(x)u',v) + (q(x)u,v).\end{aligned} \tag{4.9}$$

It is easy to see that the bilinear form $\mathscr{B}(\cdot,\cdot)$ defined above is continuous and coercive in $H^1_\diamond(I) \times H^1_\diamond(I)$ under the conditions (4.3) (see Problem 4.1). One derives immediately from the Lax-Milgram lemma (see Appendix B) that the problem (4.8) admits a unique solution. Note that only the Dirichlet boundary condition(s) is enforced *exactly* in $H^1_\diamond(I)$, but all other boundary conditions are treated *naturally*.

The rest of this chapter is organized as follows. In the first section, we consider the problem (4.1)–(4.2) with constant coefficients and present several *Galerkin* schemes based on weak formulations using *continuous* inner products. In the second section, we consider the *Galerkin method with numerical integration* which is based on the weak formulation (4.8) using *discrete* inner products. In the third section, we present the *collocation methods* which look for approximate solutions to satisfy (4.2) and (4.1) *exactly* at a set of collocation points. In Sect. 4.4, we introduce some preconditioned iterative methods for solving the linear systems arising from spectral approximations of two-point boundary value problems. In Sect. 4.5, we provide error analysis for two model cases and the one-dimensional Helmholtz equation.

For a thorough discussion on other numerical methods for more general two-point boundary value problems, we refer to Ascher et al. (1995).

4.1 Galerkin Methods

To simplify the presentation, we shall restrict ourselves in this section to a special case of (4.4), namely,

$$\begin{aligned} &-u'' + \alpha u = f, \quad \text{in } I = (-1,1), \\ &a_- u(-1) + b_- u'(-1) = 0, \quad a_+ u(1) + b_+ u'(1) = 0, \end{aligned} \tag{4.10}$$

where $\alpha \geq 0$ is a given constant. The general case (4.1)-(4.2) will be treated in Sects. 4.2 and 4.3.

As a special case of (4.8), the standard weak formulation for (4.10) is

$$\begin{cases} \text{Find } u \in H^1_\diamond(I) \text{ such that} \\ (u',v') + h_+ u(1)v(1) - h_- u(-1)v(-1) \\ \qquad + \alpha(u,v) = (f,v), \quad \forall v \in H^1_\diamond(I). \end{cases} \tag{4.11}$$

4.1.1 Weighted Galerkin Formulation

We consider the approximation of (4.10) by using a weighted Galerkin method in the polynomial space

$$\tilde{X}_N = \{\phi \in P_N : \phi(\pm 1) = 0 \text{ if } b_\pm = 0\}. \tag{4.12}$$

A straightforward extension of (4.11) using the weighted inner product leads to the following formulation:

$$\begin{cases} \text{Find } u_N \in \tilde{X}_N \text{ such that} \\ (u_N', \omega^{-1}(v_N\omega)')_\omega + \omega(1)h_+u_N(1)v_N(1) \\ \qquad\qquad - \omega(-1)h_-u_N(-1)v_N(-1) + \alpha(u_N, v_N)_\omega \\ \qquad\qquad = (f, v_N)_\omega, \ \ \forall v_N \in \tilde{X}_N. \end{cases} \tag{4.13}$$

However, there are several problems associated with this formulation. First, the above formulation does not make sense if $\lim_{x\to\pm1}\omega(x)$ does not exist, except in the case of Dirichlet boundary conditions. Hence, it can not be used for the Jacobi weight function with $\alpha < 0$ or $\beta < 0$, including in particular the Chebyshev weight (cf. Canuto and Quarteroni (1994) and pp. 194–196 in Funaro (1992) for some special weighted weak formulations of (4.10)). Secondly, as it will become clear later in this section, even in the case $\omega(x) \equiv 1$, this formulation will not lead to a sparse or special linear system that can be inverted efficiently. The cure is to use a new weighted weak formulation in which the general boundary conditions in (4.10) are enforced *exactly* rather than *approximately* in (4.13).

Let us denote

$$X_N = \big\{v \in P_N : \ a_\pm v(\pm1) + b_\pm v'(\pm1) = 0\big\}. \tag{4.14}$$

The new weighted Galerkin method for (4.10) is

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ -(u_N'', v_N)_\omega + \alpha(u_N, v_N)_\omega = (f_N, v_N)_\omega, \quad \forall\, v_N \in X_N, \end{cases} \tag{4.15}$$

where f_N is an appropriate polynomial approximation of f, which is usually taken to be the interpolation of f associated with the Gauss-type quadrature points. The main difference with (4.13) is that the Robin boundary conditions are enforced *exactly* here. We shall see below that by choosing appropriate basis functions of X_N, we shall be able to reduce (4.15) to a linear system with a sparse or special coefficient matrix that can be solved efficiently.

Given a set of basis functions $\{\phi_j\}_{j=0}^{N-2}$ of X_N, we denote

$$\begin{aligned} &f_k = \int_I f_N\,\phi_k\omega dx, \quad \mathbf{f} = (f_0, f_1, \ldots, f_{N-2})^T; \\ &u_N = \sum_{j=0}^{N-2} \hat{u}_j\phi_j, \quad \mathbf{u} = (\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-2})^T; \\ &s_{kj} = -\int_I \phi_j''\,\phi_k\omega dx, \quad m_{kj} = \int_I \phi_j\,\phi_k\omega dx, \end{aligned} \tag{4.16}$$

and

$$S = \big(s_{kj}\big)_{0\le k,j\le N-2}, \quad M = \big(m_{kj}\big)_{0\le k,j\le N-2}.$$

Taking $v_N = \phi_k$, $0 \le k \le N-2$ in (4.15), we find that (4.15) is equivalent to the following linear system:

$$\big(S + \alpha M\big)\mathbf{u} = \mathbf{f}. \tag{4.17}$$

Below, we determine the entries of S and M for two special cases: $\omega = 1, (1-x^2)^{-1/2}$.

4.1.2 Legendre-Galerkin Method

We set $\omega(x) \equiv 1$ and $f_N = I_N f$ (the Legendre interpolation polynomial of f relative to the Legendre-Gauss-Lobatto points (cf. Sect. 3.3)). Then (4.15) becomes

$$-\int_I u_N'' v_N \, dx + \alpha \int_I u_N v_N \, dx = \int_I I_N f \, v_N \, dx, \quad \forall \, v_N \in X_N, \tag{4.18}$$

which is referred to as the Legendre-Galerkin method for (4.10).

The actual linear system for (4.18) depends on the choice of basis functions of X_N. Just as in the finite-element methods, where neighboring points are used to form basis functions so as to minimize their interactions in the physical space, neighboring orthogonal polynomials should be used to form basis functions in a spectral-Galerkin method so as to minimize their interactions in the frequency space. Therefore, we look for basis functions as a *compact combination of Legendre polynomials* (cf. Shen (1994)), namely,

$$\phi_k(x) = L_k(x) + a_k L_{k+1}(x) + b_k L_{k+2}(x), \tag{4.19}$$

where the parameters $\{a_k, b_k\}$ are chosen to satisfy the boundary conditions in (4.10). Such basis functions are referred to as *modal* basis functions.

Lemma 4.1. *For all $k \geq 0$, there exists a unique set of $\{a_k, b_k\}$ such that $\phi_k(x) = L_k(x) + a_k L_{k+1}(x) + b_k L_{k+2}(x)$ verifies the boundary conditions in (4.10).*

Proof. Since $L_k(\pm 1) = (\pm 1)^k$ and $L_k'(\pm 1) = \frac{1}{2}(\pm 1)^{k-1} k(k+1)$ (see Sect. 3.3), the boundary conditions in (4.10) lead to the following system for $\{a_k, b_k\}$:

$$\begin{aligned} \Big(a_+ + \frac{b_+}{2}(k+1)(k+2)\Big)a_k + \Big(a_+ + \frac{b_+}{2}(k+2)(k+3)\Big)b_k \\ = -a_+ - \frac{b_+}{2}k(k+1), \\ -\Big(a_- - \frac{b_-}{2}(k+1)(k+2)\Big)a_k + \Big(a_- - \frac{b_-}{2}(k+2)(k+3)\Big)b_k \\ = -a_- + \frac{b_-}{2}k(k+1). \end{aligned} \tag{4.20}$$

The determinant of the coefficient matrix is

$$\begin{aligned} \mathrm{DET}_k = 2a_+a_- + a_-b_+(k+2)^2 - a_+b_-(k+2)^2 \\ -b_-b_+(k+1)(k+2)^2(k+3)/2. \end{aligned}$$

We then derive from (4.3) that the four terms (including the signs before them) of DET_k are all nonnegative, and at least one is positive for any k. Hence, $\{a_k, b_k\}$ can be uniquely determined by solving (4.20), namely,

$$\begin{aligned} a_k =& (2k+3)(a_+b_- + a_-b+)/\mathrm{DET}_k, \\ b_k =& \Big\{ -2a_-a_+ + (k+1)^2\big(a_+b_- - a_-b_+\big) \\ &+ \frac{b_-b_+}{2}k(k+1)^2(k+2)\Big\}\Big/\mathrm{DET}_k. \end{aligned} \tag{4.21}$$

This completes the proof. □

Note that in particular:

- If $a_\pm = 1$ and $b_\pm = 0$ (Dirichlet boundary conditions), we have $a_k = 0$ and $b_k = -1$.
- If $a_\pm = 0, b_\pm = \pm 1$ (Neumann boundary conditions), we have $a_k = 0$ and $b_k = -k(k+1)/((k+2)(k+3))$.

It is obvious that $\{\phi_k\}$ are linearly independent. Therefore, by dimension argument, we have

$$X_N = \mathrm{span}\big\{\phi_k :\ k = 0, 1, \ldots, N-2\big\}.$$

Remark 4.1. *In the very special case*

$$-u_{xx} = f,\ \ x \in (-1,1); \quad u_x(\pm 1) = 0,$$

with the condition $\int_{-1}^{1} f dx = 0$*, since the solution is only determined up to a constant, we should use*

$$X_N = \mathrm{span}\big\{\phi_k :\ k = 1, 2, \ldots, N-2\big\}.$$

This remark also applies to the Chebyshev-Galerkin method presented below.

Lemma 4.2. *The stiffness matrix S is a diagonal matrix with*

$$s_{kk} = -(4k+6)b_k, \quad k = 0, 1, \ldots. \tag{4.22}$$

The mass matrix M is symmetric penta-diagonal whose nonzero elements are

$$m_{jk} = m_{kj} = \begin{cases} \dfrac{2}{2k+1} + a_k^2\dfrac{2}{2k+3} + b_k^2\dfrac{2}{2k+5}, & j = k, \\ a_k\dfrac{2}{2k+3} + a_{k+1}b_k\dfrac{2}{2k+5}, & j = k+1, \\ b_k\dfrac{2}{2k+5}, & j = k+2. \end{cases} \tag{4.23}$$

Proof. Integrating by parts and using the fact that $\{\phi_k\}$ satisfy the boundary conditions (4.5), we find that

$$\begin{aligned} s_{jk} &= -\int_I \phi_k''(x)\,\phi_j(x)dx \\ &= \int_I \phi_k'(x)\,\phi_j'(x)dx + h_+\phi_k(1)\phi_j(1) - h_-\phi_k(-1)\phi_j(-1) \\ &= -\int_I \phi_k(x)\,\phi_j''(x)dx = s_{kj}, \end{aligned} \tag{4.24}$$

where $h_\pm$ are defined in (4.7). It is then obvious from (4.24) and the definition of $\{\phi_k\}$ that S is a diagonal matrix. Thanks to (3.176c) and (3.174), we find

$$\begin{aligned} s_{kk} &= -b_k \int_I L_{k+2}''(x)L_k(x)dx \\ &= -b_k(k+1/2)(4k+6)\int_I L_k^2(x)dx = -b_k(4k+6). \end{aligned}$$

The nonzero entries for M in (4.23) can be easily obtained by using (3.174). □

Remark 4.2. *An immediate consequence is that* $\{\phi_k\}_{k=0}^{N-2}$ *forms an orthogonal basis of* X_N *with respect to the inner product* $-(u_N'', v_N)$. *Furthermore, an* orthonormal *basis of* X_N *with respect to this inner product is*

$$\tilde{\phi}_k(x) := \frac{1}{\sqrt{-b_k(4k+6)}}\phi_k(x).$$

Notice that under the assumption (4.3), $b_k < 0$ *for all* k.

We now provide a detailed implementation procedure. Given the values of f at the LGL points $\{x_j\}_{j=0}^N$, we determine the values of u_N (solution of (4.15)) at $\{x_j\}_{j=0}^N$ as follows:

1. (Pre-computation) Compute the LGL points, $\{a_k, b_k\}$ and nonzero elements of S and M.
2. Evaluate the Legendre coefficients of $I_N f$ from $\{f(x_j)\}_{j=0}^N$ (forward Legendre transform, see (3.193)) and evaluate $\mathbf{f}$.
3. Solve $\mathbf{u}$ from (4.17).
4. Evaluate $u_N(x_j) = \sum_{i=0}^{N-2} \hat{u}_i\phi_i(x_j)$, $j = 0, 1, \ldots, N$ (backward Legendre transform, see (3.194)).

Although the solution of the linear system (4.17) can be done in $O(N)$ flops, the two discrete Legendre transforms in the above procedure cost about $2N^2$ flops. To reduce the cost of the discrete transforms between the physical and frequency spaces, a natural choice is to use Chebyshev polynomials so that the discrete Chebyshev transforms can be accelerated by using FFT.

4.1.3 Chebyshev-Galerkin Method

We set $\omega = (1-x^2)^{-1/2}$ and $f_N = I_N^c f$ (the Chebyshev interpolation polynomial of f relative to the Chebyshev-Gauss-Lobatto points (see Sect. 3.4)). Then, (4.15) becomes

$$-\int_I u_N'' v_N \omega\, dx + \alpha \int_I u_N v_N \omega\, dx = \int_I I_N^c f v_N \omega\, dx, \quad \forall v_N \in X_N, \tag{4.25}$$

which is referred to as the Chebyshev-Galerkin method for (4.10).

As before, we would like to seek the basis functions of X_N in the form

$$\phi_k(x) = T_k(x) + a_k T_{k+1}(x) + b_k T_{k+2}(x). \tag{4.26}$$

Lemma 4.3. *For all $k \geq 0$, there exists a unique set of $\{a_k, b_k\}$ such that $\phi_k(x) = T_k(x) + a_k T_{k+1}(x) + b_k T_{k+2}(x)$ satisfies the boundary conditions in* (4.10).

Proof. Since $T_k(\pm 1) = (\pm 1)^k$ and $T_k'(\pm 1) = (\pm 1)^{k-1} k^2$, we find from (4.5) that $\{a_k, b_k\}$ must satisfy the system

$$\begin{aligned} &(a_+ + b_+(k+1)^2)a_k + (a_+ + b_+(k+2)^2)b_k = -a_+ - b_+ k^2, \\ &-(a_- - b_-(k+1)^2)a_k + (a_- - b_-(k+2)^2)b_k = -a_- + b_- k^2, \end{aligned} \tag{4.27}$$

whose determinant is

$$\begin{aligned} \mathrm{DET}_k = {} & 2a_+ a_- + \big\{(k+1)^2 + (k+2)^2\big\}\big(a_- b_+ - a_+ b_-\big) \\ & - 2b_- b_+ (k+1)^2 (k+2)^2. \end{aligned}$$

As in the Legendre case, the conditions in (4.3) imply that $\mathrm{DET}_k > 0$. Hence, $\{a_k, b_k\}$ are uniquely determined by

$$\begin{aligned} a_k = {} & 4(k+1)(a_+ b_- + a_- b_+)/\mathrm{DET}_k, \\ b_k = {} & \big\{(-2a_- a_+ + (k^2 + (k+1)^2)(a_+ b_- - a_- b_+) \\ & + 2b_- b_+ k^2 (k+1)^2\big\}/\mathrm{DET}_k. \end{aligned} \tag{4.28}$$

This ends the proof. □

Therefore, we have from the dimension argument that

$$X_N = \mathrm{span}\big\{\phi_k :\ k = 0, 1, \ldots, N-2\big\}.$$

One easily derives from (3.214) that the mass matrix M is a symmetric positive definite penta-diagonal matrix whose nonzero elements are

$$m_{jk} = m_{kj} = \begin{cases} \dfrac{\pi}{2}(c_k + a_k^2 + b_k^2), & j = k, \\ \dfrac{\pi}{2}(a_k + a_{k+1}b_k), & j = k+1, \\ \dfrac{\pi}{2}b_k, & j = k+2, \end{cases} \tag{4.29}$$

where $c_0 = 2$ and $c_k = 1$ for $k \geq 1$. However, the computation of s_{kj} is much more involved. Below, we shall derive the explicit expression of s_{kj} for two special cases.

Lemma 4.4. *For the case $a_\pm = 1$ and $b_\pm = 0$ (Dirichlet boundary conditions), we have $a_k = 0$, $b_k = -1$ and*

$$s_{kj} = \begin{cases} 2\pi(k+1)(k+2), & j = k, \\ 4\pi(k+1), & j = k+2, k+4, k+6, \ldots, \\ 0, & j < k \text{ or } j+k \text{ odd}. \end{cases} \tag{4.30}$$

For the case $a_\pm = 0$, $b_+ = 1$ and $b_- = -1$ (Neumann boundary conditions), we have $a_k = 0$, $b_k = -\frac{k^2}{(k+2)^2}$ and

$$s_{kj} = \begin{cases} 2\pi(k+1)k^2/(k+2), & j = k, \\ 4\pi j^2(k+1)/(k+2)^2, & j = k+2, k+4, k+6, \ldots, \\ 0, & j < k \text{ or } j+k \text{ odd}. \end{cases} \tag{4.31}$$

Proof. One observes immediately that

$$s_{kj} = -\int_I \phi_j'' \phi_k \omega \, dx = 0, \quad \text{for } j < k.$$

Hence, S is an upper triangular matrix. By the odd-even parity of the Chebyshev polynomials, we have also $s_{kj} = 0$ for $j + k$ odd.

Thanks to (3.216b), we have

$$\begin{aligned} T_{k+2}''(x) = &\frac{1}{c_k}(k+2)\big((k+2)^2 - k^2\big)T_k(x) \\ &+ \frac{1}{c_{k-2}}(k+2)\big((k+2)^2 - (k-2)^2\big)T_{k-2}(x) + \ldots. \end{aligned} \tag{4.32}$$

We first consider the case $a_\pm = 1$ and $b_\pm = 0$. From (4.21), we find $\phi_k(x) = T_k(x) - T_{k+2}(x)$. It follows immediately from (4.32) and (3.214) that

$$\begin{aligned} -(\phi_k'', \phi_k)_\omega = (T_{k+2}'', T_k)_\omega &= \frac{1}{c_k}(k+2)\big((k+2)^2 - k^2\big)(T_k, T_k)_\omega \\ &= 2\pi(k+1)(k+2). \end{aligned}$$

Setting $\phi_j''(x) = \sum_{n=0}^{j} d_n T_n(x)$, by a simple computation using (4.32), we derive

$$d_n = \begin{cases} -\dfrac{4}{c_j}(j+1)(j+2), & n = j, \\ -\dfrac{1}{c_n}\{(j+2)^3 - j^3 - 2n^2\}, & n < j. \end{cases}$$

Hence for $j = k+2, k+4, \ldots$, we find

$$-(\phi_j'', \phi_k)_\omega = -d_k(T_k, T_k)_\omega + d_{k+2}(T_{k+2}, T_{k+2})_\omega = 4\pi(k+1).$$

The case with $a_\pm = 0$ and $b_\pm = \pm 1$ can be treated similarly as above. □

Similar to the Legendre-Galerkin method, the implementation of the Chebyshev-Galerkin method for (4.10) involves the following steps:

1. (pre-computation) Compute $\{a_k, b_k\}$ and nonzero elements of S and M.
2. Evaluate the Chebyshev coefficients of $I_N^c f$ from $\{f(x_j)\}_{j=0}^N$ (forward Chebyshev transform, see (3.222)) and evaluate $\mathbf{f}$.
3. Solve $\mathbf{u}$ from (4.17).
4. Evaluate $u_N(x_j) = \sum_{i=0}^{N-2} \hat{u}_i \phi_i(x_j)$, $j = 0, 1, \ldots, N$ (backward Chebyshev transform, see (3.223)).

Remark 4.3. *Note that the forward and backward Chebyshev transforms can be performed by using FFT in $O(N \log_2 N)$ operations. However, the cost of Step* 3 *depends on the boundary conditions in* (4.5)*. For the special but important cases described in the above lemma, the special structures of S would allow us to solve the system* (4.17) *in $O(N)$ operations. More precisely, in* (4.30) *and* (4.31)*, the nonzero elements of S take the form $s_{kj} = a(j) * b(k)$. Hence, a special Gaussian elimination procedure for* (4.17) (cf. Shen (1995)) *would only require $O(N)$ flops instead of $O(N^3)$ flops for a general full matrix.*

Therefore, thanks to FFT, the computational complexity of Chebyshev-Galerkin method for the above cases is $O(N \log_2 N)$ which is quasi-optimal (*i.e., optimal up to a logarithmic term*)*.*

Remark 4.4. *In the case of Dirichlet boundary conditions, one can also use the basis functions $\psi_k(x) = (1-x^2)T_k(x)$* (cf. Heinrichs (1989))*, which lead to a banded stiffness matrix.*

4.1.4 Chebyshev-Legendre Galerkin Method

The main advantage of using Chebyshev polynomials is that the discrete Chebyshev transforms can be performed in $O(N \log_2 N)$ operations by using FFT. However, the Chebyshev-Galerkin method leads to non-symmetric formulations which may

cause difficulties in analysis and implementation. On the other hand, the Legendre-Galerkin method leads to symmetric formulation and sparse matrices for problems with constant coefficients, but the discrete Legendre transforms are expensive (with $O(N^2)$ operations). In order to take advantage of both the Legendre and Chebyshev methods (cf. Don and Gottlieb (1994)), one may use the so-called Chebyshev-Legendre Galerkin method (cf. Shen (1996)):

$$-\int_I u_N'' v_N\,dx+\alpha\int_I u_N v_N\,dx=\int_I I_N^c f v_N\,dx, \tag{4.33}$$

where I_N^c denotes the interpolation operator relative to the Chebyshev-Gauss-Lobatto points. So the only difference with (4.18) is that the Chebyshev interpolation operator I_N^c is used here to replace the Legendre interpolation operator in (4.18). Therefore, (4.33) leads to the linear system (4.17) with $\mathbf{u}$, S and M defined in (4.16) and (4.22)-(4.23), but with $\mathbf{f}$ defined by

$$f_k=\int_I I_N^c f\,\phi_k dx,\quad \mathbf{f}=(f_0,f_1,\ldots,f_{N-2})^T. \tag{4.34}$$

Hence, the solution procedure of (4.33) is essentially the same as that of (4.18) except that *Chebyshev-Legendre transforms* (between the value of a function at the *CGL* points and the coefficients of its *Legendre* expansion) are needed instead of the *Legendre transforms*. More precisely, given the values of f at the CGL points $\{x_i=\cos(\frac{i\pi}{N})\}_{0\le i\le N}$, we determine the values of u_N (solution of (4.33)) at the CGL points as follows:

1. (Pre-computation) Compute $\{a_k,b_k\}$ and nonzero elements of S and M.
2. Evaluate the Legendre coefficients of $I_N^c f$ from $\{f(x_i)\}_{i=0}^N$ (forward Chebyshev-Legendre transform).
3. Evaluate $\mathbf{f}$ from (4.34) and solve $\mathbf{u}$ from (4.17).
4. Evaluate $u_N(x_j)=\sum_{i=0}^{N-2}\hat{u}_i\phi_i(x_j)$, $j=0,1,\ldots,N$ ("modified" backward Chebyshev-Legendre transform).

The forward and ("modified") backward Chebyshev-Legendre transforms can be implemented efficiently. Indeed, each Chebyshev-Legendre transform can be split into two steps:

1. The transform between its physical values at Chebyshev-Gauss-Lobatto points and the coefficients of its Chebyshev expansion. This can be done by using FFT in $O(N\log_2 N)$ operations.
2. The transform between the coefficients of the Chebyshev expansion and of the Legendre expansion. Alpert and Rokhlin (1991) developed an $O(N)$ algorithm for this transform given a prescribed precision.

Therefore, the total computational cost for (4.33) is of order $O(N\log_2 N)$.

The algorithm in Alpert and Rokhlin (1991) is based on the fast multipole method (cf. Greengard and Rokhlin (1987)). Hence, it is most attractive for very large N. For small to moderate N, a simple algorithm described in Shen (1996) appears to be more competitive.

4.2 Galerkin Method with Numerical Integration

The *Galerkin* methods presented in the previous section lead to very efficient algorithms for problems with constant coefficients. However, they are not feasible for problems with general variable coefficients for which the exact integration is often not possible. Therefore, for problems with variable coefficients, we need to replace the continuous inner product by a suitable discrete inner product, leading to the so-called *Galerkin method with numerical integration*. More precisely, the Legendre-Galerkin method with numerical integration for (4.8) is

$$\begin{cases} \text{Find } u_N \in \tilde{X}_N = P_N \cap H^1_\diamond(I) \text{ such that} \\ \mathscr{B}_N(u_N, v_N) = \langle f, v_N \rangle_N, \quad \forall v_N \in \tilde{X}_N, \end{cases} \tag{4.35}$$

where

$$\begin{aligned} \mathscr{B}_N(u_N, v_N) := {} & \varepsilon \langle u_N', v_N' \rangle_N + \varepsilon h_+ u_N(1) v_N(1) - \varepsilon h_- u_N(-1) v_N(-1) \\ & + \langle p(x) u_N', v_N \rangle_N + \langle q(x) u_N, v_N \rangle_N, \end{aligned}$$

with $\langle \cdot, \cdot \rangle_N$ being the discrete inner product relative to the Legendre-Gauss-Lobatto quadrature.

Let $\{h_j\}$ be the Lagrange basis polynomials (also referred to as *nodal basis*) associated with $\{x_j\}_{j=0}^N$. To fix the idea, we assume $b_\pm \neq 0$, so $\tilde{X}_N = P_N$ and we can write

$$u_N(x) = \sum_{j=0}^{N} u_N(x_j) h_j(x). \tag{4.36}$$

Plugging the above expression into (4.35) and taking $v_N = h_k$, we find that (4.35) reduces to the linear system

$$B\mathbf{w} = W\mathbf{f}, \tag{4.37}$$

where

$$\begin{aligned} & \mathbf{w} = \big(u_N(x_0), u_N(x_1), \ldots, u_N(x_N)\big)^T; \\ & b_{kj} = \mathscr{B}_N(h_j, h_k), \quad B = (b_{kj})_{k,j=0,1,\ldots,N}; \\ & \mathbf{f} = \big(f(x_0), f(x_1), \ldots, f(x_N)\big)^T; \\ & W = \operatorname{diag}(\omega_0, \omega_1, \ldots, \omega_N), \end{aligned} \tag{4.38}$$

with $\{\omega_k\}_{k=0}^N$ being the weights of the Legendre-Gauss-Lobatto quadrature (see Theorem 3.29).

The entries b_{kj} can be determined as follows. Let $\{x_j\}_{j=0}^N$ be arranged in ascending order[1] with $x_0 = -1$ and $x_N = 1$. Using (3.59) and integration by parts, we have

$$\begin{aligned}\langle h_j', h_k' \rangle_N = (h_j', h_k') &= -(h_j'', h_k) + h_j' h_k \big|_{-1}^{1} \\ &= -\left(D^2\right)_{kj} \omega_k + d_{Nj}\delta_{Nk} - d_{0j}\delta_{0k}.\end{aligned} \tag{4.39}$$

Consequently,

$$\begin{aligned}b_{kj} = &\left[-\varepsilon \left(D^2\right)_{kj} + p(x_k) d_{kj} + q(x_k)\delta_{kj}\right]\omega_k \\ &+ \varepsilon\,(d_{Nj} + h_+\delta_{Nj})\delta_{Nk} - \varepsilon\,(d_{0j} + h_-\delta_{0j})\delta_{0k}.\end{aligned} \tag{4.40}$$

We can also reinterpret (4.35) as a collocation form. Observe that

$$\langle u_N', h_k' \rangle_N = -u_N''(x_k)\omega_k + u_N'(1)\delta_{Nk} - u_N'(-1)\delta_{0k}, \quad 0 \le k \le N.$$

Then, taking $v_N = h_j$ in (4.35) for $j = 0, 1, \ldots, N$, since $\omega_0 = \omega_N = \frac{2}{N(N+1)}$, we find

$$\begin{cases} -\varepsilon\, u_N''(x_j) + p(x_j) u_N'(x_j) + q(x_j) u_N(x_j) = f(x_j),\ 1 \le j \le N-1, \\ a_- u_N(-1) + b_- u_N'(-1) = -\dfrac{b_-}{\varepsilon}\dfrac{2}{N(N+1)} \\ \qquad \big[f(-1) - (-\varepsilon\, u_N''(-1) + p(-1)u_N'(-1) + q(-1)u_N(-1))\big], \\ a_+ u_N(1) + b_+ u_N'(1) = \dfrac{b_+}{\varepsilon}\dfrac{2}{N(N+1)} \\ \qquad \big[f(1) - (-\varepsilon\, u_N''(1) + p(1)u_N'(1) + q(1)u_N(1))\big]. \end{cases} \tag{4.41}$$

Remark 4.5. *Note that the solution of* (4.35) *satisfies* (4.4) *exactly at the interior collocation points* $\{x_j\}_{j=1}^{N-1}$*, but the boundary conditions* (4.5) *are only satisfied approximately with an error proportional to the residual of* (4.4)*, with u replaced by the approximate solution* u_N*, at the boundary. Thus,* (4.35) *does not correspond exactly to a collocation method, so it is sometimes referred to as a* collocation method in the weak form*. However, it is clear from* (4.41) *that in the Dirichlet case* (*i.e.,* $b_\pm = 0$)*,* (4.41) *becomes* a collocation method (*see the next section*)*. In other words, the Galerkin method with numerical integration* (4.35)*, in the case of Dirichlet boundary conditions, is equivalent to the collocation method.*

Remark 4.6. *The matrix B in the linear system* (4.37)*, even for the simplest differential equation, is full and ill-conditioned, so it is in general not advisable to solve* (4.37) *using a direct method for large N. Instead, an iterative method using an appropriate preconditioner should be used, see Sect.* 4.4*.*

[1] Historically (cf. Gottlieb and Orszag (1977)), the Chebyshev-collocation points were defined as $x_j = \cos\frac{j\pi}{N}$ which were in descending order. For the sake of consistency, we choose to arrange the collocation points in ascending order in this book.

4.3 Collocation Methods

The collocation method, or more specifically the *collocation method in the strong form*, is fundamentally different from the Galerkin method, in the sense that it is not based on a weak formulation. Instead, it looks for an approximate solution which enforces the boundary conditions in (4.5) and collocates (4.4) at a set of interior collocation points. On the other hand, the *collocation method in the weak form* presented in the last section is based on a weak formulation in which the general boundary conditions are treated *naturally* and are only satisfied *asymptotically*, and the approximate solution verifies (4.4) at a set of interior collocation points.

We describe below the collocation method for the two-point boundary value problem (4.1) with the general boundary conditions (4.2). Notice that the non-homogeneous boundary conditions can be treated directly in a collocation method so there is no need to "homogenize" the boundary conditions as we did previously for the Galerkin methods.

Given any set of distinct collocation points $\{x_j\}_{j=0}^N$ on $[-1,1]$ in ascending order with $x_0 = -1$ and $x_N = 1$, the collocation method for (4.1) with (4.2) is

$$\begin{cases} \text{Find } u_N \in P_N \text{ such that} \\ -\varepsilon u_N''(x_i) + p(x_i)u_N'(x_i) + q(x_i)u_N(x_i) = F(x_i), \ \ 1 \le i \le N-1, \\ a_- u_N(-1) + b_- u_N'(-1) = c_-, \quad a_+ u_N(1) + b_+ u_N'(1) = c_+. \end{cases} \tag{4.42}$$

Let $\{h_j\}$ be the Lagrange basis polynomials associated with $\{x_j\}_{j=0}^N$, and let $D = \big(d_{kj} := h_j'(x_k)\big)_{k,j=0,1,\ldots,N}$. Writing $w_j = u_N(x_j)$ and $u_N(x) = \sum_{j=0}^N w_j h_j(x)$, we have

$$\begin{aligned} u_N(x_k) &= \sum_{j=0}^N w_j h_j(x_k) = w_k, \\ u_N'(x_k) &= \sum_{j=0}^N w_j h_j'(x_k) = \sum_{j=0}^N d_{kj} w_j \\ &= \sum_{j=1}^{N-1} d_{kj} w_j + d_{k0} w_0 + d_{kN} w_N, \\ u_N''(x_k) &= \sum_{j=0}^N w_j h_j''(x_k) = \sum_{j=0}^N \big(D^2\big)_{kj} w_j \\ &= \sum_{j=1}^{N-1} \big(D^2\big)_{kj} w_j + \big(D^2\big)_{k0} w_0 + \big(D^2\big)_{kN} w_N. \end{aligned}$$

Substituting the above into (4.42) leads to

$$\begin{cases} \sum_{j=0}^{N} \left[-\varepsilon \left(D^2\right)_{ij} + p(x_i)d_{ij} + q(x_i)\delta_{ij} \right] w_j = F(x_i), \quad i = 1,2,\ldots,N-1, \\ a_- w_0 + b_- \sum_{j=0}^{N} d_{0j} w_j = c_-, \quad a_+ w_N + b_+ \sum_{j=0}^{N} d_{Nj} w_j = c_+. \end{cases} \tag{4.43}$$

Let us denote

$$\begin{aligned} &a_{ij} = -\varepsilon \left(D^2\right)_{ij} + p(x_i)d_{ij} + q(x_i)\delta_{ij}, \quad 1 \le i \le N-1,\ 0 \le j \le N, \\ &a_{0j} = a_-\delta_{0j} + b_- d_{0j}, \ \ a_{Nj} = a_+\delta_{Nj} + b_+ d_{Nj}, \quad 0 \le j \le N, \\ &\mathbf{b} = \left(c_-, F(x_1), F(x_2), \ldots, F(x_{N-1}), c_+\right)^T, \\ &\mathbf{w} = \left(w_0, w_1, \ldots, w_N\right)^T, \quad A = (a_{ij})_{0 \le i,j \le N}. \end{aligned} \tag{4.44}$$

Then, the linear system (4.43) reduces to

$$A\mathbf{w} = \mathbf{b}. \tag{4.45}$$

Remark 4.7. *Notice that the above formulation is valid for any set of collocation points. However, the choice of collocation points is essential for the stability, convergence and efficiency of the collocation method. For two-point boundary value problems, the Gauss-Lobatto points are commonly used. Due to the global nature of the Lagrange basis polynomials, the system matrix A in* (4.45) *is always full and ill-conditioned, even for problems with constant coefficients.*

Remark 4.8. *For the case of homogeneous Dirichlet boundary conditions, i.e.,* $u(\pm 1) = 0$*, the collocation method* (4.42) *with* $\{x_j\}$ *being the Legendre-Gauss-Lobatto points, as observed in Remark* 4.5*, is equivalent to the Galerkin method with numerical integration* (4.35)*.*

It is interesting to note that in the case of Dirichlet boundary conditions, after eliminating w_0 *and* w_N *from* (4.37) *and* (4.45)*, the reduced* $(N-1)\times(N-1)$ *matrices B and A are related by* $B = WA$*, where W is the diagonal matrix* $W = \mathrm{diag}(\omega_1, \omega_2, \ldots, \omega_{N-1})$*. Furthermore, the condition number of A behaves like* $O(N^4)$ (cf. Orszag (1980))*, while that of B behaves like* $O(N^3)$ (cf. Bernardi and Maday (1992a)).

Remark 4.9. *If the bilinear form* $\mathscr{B}(\cdot,\cdot)$ *in* (4.9) *is self-adjoint, then the matrix B in* (4.37) *from the Galerkin method with numerical integration is symmetric. However, the matrix A in* (4.45) *from the collocation method is always non-symmetric.*

4.3.1 Galerkin Reformulation

We show below that in the case of homogeneous Dirichlet boundary conditions, the collocation method (4.42) with $\{x_j\}$ being the Jacobi-Gauss-Lobatto points, can be reformulated as a Galerkin method with numerical integration.

Lemma 4.5. *Let $\omega = (1+x)^\alpha(1-x)^\beta$ be the Jacobi weight function with $\alpha,\beta > -1$, $\{x_j\}_{j=0}^N$ be the Jacobi-Gauss-Lobatto points, and $\langle\cdot,\cdot\rangle_{N,\omega}$ be the discrete inner product associated with the Jacobi-Gauss-Lobatto quadrature* (*cf. Theorem* 3.27). *Then* (4.42) *with $b_\pm = c_\pm = 0$ is equivalent to*

$$\begin{cases} \text{Find } u_N \in P_N^0 = P_N \cap H_0^1(I) \text{ such that} \\ \varepsilon\,\langle u_N', \omega^{-1}(v_N\omega)'\rangle_{N,\omega} + \langle p(x)u_N', v_N\rangle_{N,\omega} \\ \qquad + \langle q(x)u_N, v_N\rangle_{N,\omega} = \langle F, v_N\rangle_{N,\omega}, \quad \forall v_N \in P_N^0. \end{cases} \tag{4.46}$$

Proof. By a direct computation, we find that

$$\omega^{-1}(v_N\omega)' = \omega^{-1}(v_N'\omega + v_N\omega') = v_N' - \big(\alpha(1+x) - \beta(1-x)\big)\frac{v_N}{1-x^2}.$$

Since $v_N(\pm 1) = 0$ and $v_N \in P_N$, we derive that $\omega^{-1}(v_N\omega)' \in P_{N-1}$. Therefore, thanks to (3.59), we find that

$$\begin{aligned} \langle u_N', \omega^{-1}(v_N\omega)'\rangle_{N,\omega} &= (u_N', \omega^{-1}(v_N\omega)')_\omega \\ &= -(u_N'', v_N)_\omega = -\langle u_N'', v_N\rangle_{N,\omega}. \end{aligned} \tag{4.47}$$

Therefore, the formulation (4.46) is equivalent to

$$\langle -\varepsilon u_N'' + p(x)u_N' + q(x)u_N, v_N\rangle_{N,\omega} = \langle F, v_N\rangle_{N,\omega}, \quad \forall v_N \in P_N^0. \tag{4.48}$$

Notice that

$$P_N^0 = \operatorname{span}\big\{h_1(x), h_2(x), \ldots, h_{N-1}(x)\big\},$$

Taking $v_N = h_i$ for $1 \le i \le N-1$ in (4.48) leads to (4.42) with $b_\pm = c_\pm = 0$.

On the other hand, taking the discrete inner product of (4.42) with $h_k(x)$ for $1 \le k \le N-1$, we find that the solution u_N of (4.42) with $b_\pm = c_\pm = 0$ verifies (4.46). □

This lemma indicates that for (4.4) with Dirichlet boundary conditions, the Jacobi-collocation method, including the Legendre- and Chebyshev-collocation methods, can be reformulated as a *Galerkin method with numerical integration*. An obvious advantage of this reformulation is that error estimates for the Jacobi-collocation method can be carried out in the same way as the Jacobi-Galerkin method.

4.3.2 Petrov-Galerkin Reformulation

Except for the Dirichlet case, the collocation method (4.42) can not be reformulated as a Galerkin method with numerical integration. However, it can be reformulated as a Petrov-Galerkin method for which the trial functions and test functions are taken from different spaces.

Lemma 4.6. *Let* $\omega = (1+x)^{\alpha}(1-x)^{\beta}$ *be the Jacobi weight function with* $\alpha,\beta > -1$, $\{x_j\}_{j=0}^{N}$ *be the set of Jacobi-Gauss-Lobatto points, and* $\langle\cdot,\cdot\rangle_{N,\omega}$ *be the discrete inner product associated with the Jacobi-Gauss-Lobatto quadrature* (*cf. Theorem* 3.27). *Then,* (4.42) *with* $c_{\pm} = 0$ *is equivalent to the following Petrov-Galerkin method*:

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ \varepsilon\left\langle u_N', \omega^{-1}(v_N\omega)'\right\rangle_{N,\omega} + \left\langle p(x)u_N', v_N\right\rangle_{N,\omega} \\ \qquad\qquad + \left\langle q(x)u_N, v_N\right\rangle_{N,\omega} = \left\langle F, v_N\right\rangle_{N,\omega}, \quad \forall v_N \in P_N^0, \end{cases} \tag{4.49}$$

where X_N *is defined in* (4.14).

Proof. By definition, the solution u_N of (4.42) with $c_{\pm} = 0$ is in X_N. The property (4.47) still holds for $u_N \in X_N$ and $v_N \in P_N^0$, so does (4.48). Taking the discrete inner product of (4.42) with $h_k(x)$ for $k = 1,2,\ldots,N-1$, we find that the solution u_N of (4.42) with $c_{\pm} = 0$ verifies (4.49). Conversely, taking $v_N = h_i$ for $1 \le i \le N-1$ in (4.48) gives (4.42) with $c_{\pm} = 0$. □

This reformulation will allow us to obtain error estimates for the collocation method (4.42) by using the standard techniques developed for Petrov-Galerkin methods.

4.4 Preconditioned Iterative Methods

As noted in the previous two sections, there is no suitable direct spectral solver for equations with general variable coefficients. Hence, an appropriate iterative method should be used. Since the bilinear form associated with (4.4)–(4.5) is generally not symmetric nor necessarily positive definite, it is in general not advisable to apply an iterative method directly, unless the equation is diffusion dominant, i.e., ε is sufficiently large, when compared with $p(x)$. Instead, it is preferable to transform (4.4)-(4.5) into an equivalent equation whose bilinear form becomes positive definite. Indeed, multiplying (4.4) by the function

$$a(x) = \exp\left(-\frac{1}{\varepsilon}\int p(x)dx\right)$$

and using $-\varepsilon a'(x) = a(x)p(x)$, we find that (4.4) is equivalent to

$$-(a(x)u'(x))' + b(x)u(x) = g(x), \tag{4.50}$$

where $b(x) = a(x)q(x)/\varepsilon$ and $g(x) = a(x)f(x)/\varepsilon$. Hereafter, we assume that there are three constants c_1, c_2 and c_3 such that

$$0 < c_1 \le a(x) \le c_2, \quad 0 \le b(x) \le c_3, \quad \forall x \in [-1,1]. \tag{4.51}$$

We denote

$$\begin{aligned} \mathscr{B}(u,v) := &\int_{-1}^{1} a(x)u'v'dx + a(1)h_+u(1)v(1) - a(-1)h_-u(-1)v(-1) \\ &+ \int_{-1}^{1} b(x)uvdx, \ \ \forall u,v \in H^1_\diamond(I), \end{aligned} \tag{4.52}$$

where $H^1_\diamond(I)$ and $h_\pm$ are defined in (4.6) and (4.7), respectively. The weak formulation associated with (4.50) with general boundary conditions (4.5) is

$$\begin{cases} \text{Find } u \in H^1_\diamond(I) \text{ such that} \\ \mathscr{B}(u,v) = (g,v), \quad \forall v \in H^1_\diamond(I). \end{cases} \tag{4.53}$$

Hence, under the conditions (4.3) and (4.51), we find that $\mathscr{B}(u,v)$ is self-adjoint, continuous and coercive in $H^1_\diamond(I)$ so that the problem (4.53) admits a unique solution. Instead of dealing with the original equation (4.4)–(4.5), we shall consider below the equivalent problem (4.53) whose bilinear form is symmetric and positive definite.

4.4.1 Preconditioning in the Modal Basis

Let p_k be the Legendre or Chebyshev polynomial of degree k, X_N be defined in (4.14), and $\{\phi_k = p_k + a_k p_{k+1} + b_k p_{k+2}\}_{k=0}^{N-2}$ be the basis functions of X_N constructed in Sect. 4.1. Let I_N be the interpolation operator based on the Legendre or Chebyshev Gauss-Lobatto points $\{x_j\}_{j=0}^N$, and $\langle\cdot,\cdot\rangle_{N,\omega}$ (with $\omega = 1, (1-x^2)^{-1/2}$) be the associated discrete inner product. We consider the following *Galerkin method with numerical integration* for (4.53):

$$\begin{cases} \text{Find } u_N = \displaystyle\sum_{k=0}^{N-2} \hat{u}_k \phi_k \in X_N \text{ such that} \\ \mathscr{B}_{N,\omega}(u_N, \phi_j) = \langle g, \phi_j \rangle_{N,\omega}, \quad j = 0,1,\ldots,N-2, \end{cases} \tag{4.54}$$

where

$$\mathscr{B}_{N,\omega}(u_N, v_N) := -\big\langle [I_N(au_N')]', v_N \big\rangle_{N,\omega} + \big\langle bu_N, v_N \big\rangle_{N,\omega}. \tag{4.55}$$

Let us denote

$$b_{jk} = \mathscr{B}_{N,\omega}(\phi_k, \phi_j), \quad B = \big(b_{jk}\big)_{j,k=0,1,\ldots,N-2};$$
$$g_j = \langle g, \phi_j \rangle_{N,\omega}, \quad \mathbf{g} = \big(g_0, g_1, \ldots, g_{N-2}\big)^T;$$
$$\mathbf{u} = \big(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-2}\big)^T.$$

Then, (4.54) is equivalent to the following linear system:

$$B\mathbf{u} = \mathbf{g}. \tag{4.56}$$

We observe that for $u_N = \sum_{k=0}^{N-2} \hat{u}_k \phi_k \in X_N$ and $v_N = \sum_{k=0}^{N-2} \hat{v}_k \phi_k \in X_N$, we have

$$\langle B\mathbf{u}, \mathbf{v} \rangle_{l^2} = \mathscr{B}_{N,\omega}(u_N, v_N), \tag{4.57}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle_{l^2} = \sum_{j=0}^{N-2} a_j b_j$ for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{N-1}$ with components $\{a_j, b_j\}$. It is easy to see that in general B is a full matrix, so we shall resort to an iterative method for which an efficient evaluation of the matrix–vector product $B\mathbf{u}$ is essential.

We now describe how to evaluate

$$(B\mathbf{u})_j = -\big\langle [I_N(au_N')]', \phi_j \big\rangle_{N,\omega} + \big\langle bu_N, \phi_j \big\rangle_{N,\omega}, \quad j = 0, 1, \ldots, N-2$$

without explicitly forming the matrix B. Given $u_N = \sum_{k=0}^{N-2} \hat{u}_k \phi_k$, we compute "$-\big\langle [I_N(au_N')]', \phi_j \big\rangle_{N,\omega}$" as follows:

1. Using (3.206) or (3.234) to determine $\{\tilde{u}_k^{(1)}\}$ from

$$u_N'(x) = \sum_{k=0}^{N-2} \hat{u}_k \phi_k'(x) = \sum_{k=0}^{N} \tilde{u}_k^{(1)} p_k(x);$$

2. (Forward discrete transform) Compute

$$u_N'(x_j) = \sum_{k=0}^{N} \tilde{u}_k^{(1)} p_k(x_j), \quad j = 0, 1, \ldots, N;$$

3. (Backward discrete transform) Determine $\{\tilde{w}_k\}$ from

$$I_N(au_N')(x_j) = \sum_{k=0}^{N} \tilde{w}_k p_k(x_j), \quad j = 0, 1, \ldots, N;$$

4. Using (3.206) or (3.234) to determine $\{\tilde{w}_k^{(1)}\}$ from

$$\big[I_N(au_N')\big]'(x) = \sum_{k=0}^{N} \tilde{w}_k p_k'(x) = \sum_{k=0}^{N} \tilde{w}_k^{(1)} p_k(x);$$

5. For $j = 0, 1, \ldots, N-2$, compute

$$-\big\langle [I_N(au_N')]', \phi_j \big\rangle_{N,\omega} = -\sum_{k=0}^{N} \tilde{w}_k^{(1)} \langle p_k, \phi_j \rangle_{N,\omega}.$$

Note that the main cost in the above procedure is the two discrete transforms in Steps 2 and 3. The cost for each of Steps 1, 4 and 5 is $O(N)$ flops. The term $\langle bu_N, \phi_j \rangle_{N,\omega}$ can also be computed similarly as follows:

1. Compute

$$u_N(x_j) = \sum_{k=0}^{N} \hat{u}_k \phi_k(x_j), \quad j = 0, 1, \ldots, N;$$

2. Determine $\{\tilde{w}_k\}$ from

$$I_N(bu_N)(x_j) = \sum_{k=0}^{N} \tilde{w}_k p_k(x_j), \quad j = 0, 1, \ldots, N;$$

3. Compute

$$-\big\langle bu_N, \phi_j \big\rangle_{N,\omega}, \quad j = 0, 1, \ldots, N-2.$$

Hence, if b is not a constant, two additional discrete transforms are needed. In summary, the total cost for evaluate $B\mathbf{u}$ is dominated by four (only two if b is a constant) discrete transforms, and is $O(N^2)$ (resp. $O(N \log_2 N)$) flops in the Legendre (resp. Chebyshev) case.

4.4.1.1 Legendre Case

Thanks to (3.59), we have for any $u_N, v_N \in X_N$,

$$\begin{aligned} -\big\langle [I_N(au_N')]', v_N \big\rangle_N =& \langle au_N', v_N' \rangle_N + a(1) h_+ u_N(1) v_N(1) \\ & - a(-1) h_- u_N(-1) v_N(-1), \end{aligned} \tag{4.58}$$

where $h_\pm$ are defined in (4.7). Hence,

$$\mathscr{B}_N(u_N, v_N) = \mathscr{B}_N(v_N, u_N), \quad \forall u_N, v_N \in X_N.$$

Consequently, B is symmetric.

To simplify the presentation, we shall assume that $b_+ b_- = 0$ so that the Poincaré inequality is applicable to u_N.

Under the conditions (4.3) and (4.51), we have

$$\begin{aligned} \mathscr{B}_N(u_N, u_N) =& \big\langle au_N', u_N' \big\rangle_N + a(1) h_+ u_N^2(1) - a(-1) h_- u_N^2(-1) \\ & + \langle bu_N, u_N \rangle_N \geq c_1 \big\langle u_N', u_N' \big\rangle_N = c_1 \big(u_N', u_N'\big). \end{aligned} \tag{4.59}$$

On the other hand, using the Poincaré inequality (B.21) and the Sobolev inequality (B.33), it is easy to show that there exists $c_4 > 0$ such that

$$\mathscr{B}_N(u_N, u_N) \le c_4(u'_N, u'_N).$$

Hence, let $s_{ij} = (\phi'_j, \phi'_i)$ and $S = (s_{ij})_{i,j=0,1,\ldots,N-2}$. We have

$$0 < c_1 \le \frac{\langle B\mathbf{u}, \mathbf{u}\rangle_{l^2}}{\langle S\mathbf{u}, \mathbf{u}\rangle_{l^2}} = \frac{\mathscr{B}_N(u_N, u_N)}{(u'_N, u'_N)} \le c_4. \tag{4.60}$$

Since $S^{-1}B$ is symmetric with respect to the inner product $\langle \mathbf{u}, \mathbf{v}\rangle_S := \langle S\mathbf{u}, \mathbf{v}\rangle_{l^2}$, (4.60) implies immediately

$$\mathrm{cond}(S^{-1}B) \le \frac{c_4}{c_1}. \tag{4.61}$$

In other words, S^{-1} is an optimal preconditioner for B in the sense that the convergence rate of the conjugate gradient method applied to the preconditioned system

$$S^{-1}B\mathbf{u} = S^{-1}\mathbf{g} \tag{4.62}$$

will be independent of N. We recall from Sect. 4.1 that S is a diagonal matrix so the cost of applying S^{-1} is negligible. Hence, the main cost in each iteration is the evaluation of $B\mathbf{u}$ for given $\mathbf{u}$.

Remark 4.10. *In the case of Dirichlet boundary conditions, we have $\phi_k = L_k - L_{k+2}$ which, together with* (3.176a)*, implies that $\phi'_k = -(2k+3)L_{k+1}$. Therefore, from $u = \sum_{k=0}^{N-2} \hat{u}_k\phi_k$, we can obtain the derivative $u' = -\sum_{k=0}^{N-2}(2k+3)\hat{u}_kL_{k+1}$ in the modal basis without using* (3.206)*.*

Remark 4.11. *If we use the normalized basis functions*

$$\tilde{\phi}_k := \big(-b_k(4k+6)\big)^{-1/2}\phi_k \ \text{ with } (\tilde{\phi}'_j, \tilde{\phi}'_i) = \delta_{ij},$$

the condition number of the corresponding matrix B with $b_{ij} = \mathscr{B}_N(\tilde{\phi}_j, \tilde{\phi}_i)$ is uniformly bounded. Hence, we can apply the conjugate gradient method directly to this system without preconditioning.

Remark 4.12. *If c_3 in* (4.51) *is large, the condition number in* (4.61) *will be large even though independent of N. In this case, one may improve the situation by replacing the bilinear form (u'_N, v'_N) with $\hat{a}(u'_N, v'_N) + \hat{b}(u_N, v_N)$ where*

$$\hat{a} = \frac{1}{2}\Big(\max_{|x|\le 1} a(x) + \min_{|x|\le 1} a(x)\Big), \quad \hat{b} = \frac{1}{2}\Big(\max_{|x|\le 1} b(x) + \min_{|x|\le 1} b(x)\Big).$$

The matrix corresponding to this new bilinear form is $\hat{a}S + \hat{b}M$ which is positive definite and penta-diagonal (cf. Sect. 4.1)*.*

4.4.1.2 Chebyshev Case

In the Chebyshev case, an appropriate preconditioner for the inner product $\mathscr{B}_{N,\omega}(u_N, v_N)$ in $X_N \times X_N$ is $\left(u'_N, \omega^{-1}(v_N\omega)'\right)_\omega$ for which the associated linear system can be solved in $O(N)$ flops as shown in Sect. 4.1. Ample numerical results indicate that the convergence rate of a conjugate gradient type method for non-symmetric systems such as Conjugate Gradient Square (CGS) or BICGStab methods (see Appendix C) is similar to that in the Legendre case.

The advantage of using the Chebyshev polynomials is of course that the evaluation of $B\mathbf{u}$ can be accelerated by FFT in $O(N\log_2 N)$ operations, instead of $O(N^2)$ in the Legendre case.

A few remarks on the use of modal basis functions are in order.

- For problems with constant coefficients, using appropriate modal basis functions leads to sparse matrices.
- For problems with variable coefficients, one can use a suitable problem with constant coefficients as an effective preconditioner.
- With the modal basis, the choice of collocation points (as long as they are Gauss-type quadrature points) is not important, as it is merely used to define an approximation $I_N f$ to f. Therefore, we can use the same set of Gauss-Lobatto points for almost any problem. On the other hand, with the nodal basis, the choice of quadrature rules/collocation points plays an important role and should be made in accordance with the underlying differential equations and boundary conditions (see Sect. 6.4), particularly for high-order equations and mixed type boundary conditions.

We emphasize that the preconditioning in the modal basis will be less effective if the coefficients $a(x)$ and $b(x)$ have large variations, since the variation of the coefficients is not taken into account in the construction of the preconditioner. However, preconditioners which are robust to the variation in coefficients can be constructed in the nodal basis as shown below.

4.4.2 Preconditioning in the Nodal Basis

For problems with large variations in coefficients $a(x)$ and $b(x)$, it is preferable to construct preconditioners in the physical space, i.e., in the nodal basis. We shall consider two approaches: (a) a finite difference preconditioner (cf. Orszag (1980)) for the collocation method for (4.50) with general boundary conditions (4.5); and (b) a finite element preconditioner (cf. Canuto and Quarteroni (1985), Deville and Mund (1985)) for the Galerkin method with numerical integration for (4.53).

4.4.2.1 Finite Difference Preconditioning

The collocation method in the strong form for (4.50) with (4.5) is

$$\begin{cases} \text{Find } u_N \in P_N \text{ such that} \\ -(au_N')'(x_j) + b(x_j)u_N(x_j) = g(x_j), \quad 1 \le j \le N-1, \\ a_- u_N(-1) + b_- u_N'(-1) = 0, \quad a_+ u_N(1) + b_+ u_N'(1) = 0. \end{cases} \tag{4.63}$$

As in Sect. 4.3, (4.63) can be rewritten as an $(N+1)\times(N+1)$ linear system

$$A\mathbf{w} = \mathbf{b}, \tag{4.64}$$

where the unknowns are $\{w_j := u_N(x_j)\}_{j=0}^N$, and

$$\mathbf{w} = \big(w_0, w_1, \ldots, w_N\big)^T, \quad \mathbf{b} = \big(0, g(x_1), g(x_2), \ldots, g(x_{N-1}), 0\big)^T. \tag{4.65}$$

As suggested by Orszag (1980), we can build a preconditioner for A by using a finite difference approximation to (4.50) with (4.5). Let us denote

$$h_k = x_k - x_{k-1}, \quad \tilde{h}_k = (x_{k+1} - x_{k-1})/2, \quad a_{k+1/2} = a\big((x_{k+1} + x_k)/2\big). \tag{4.66}$$

Then, the second-order finite difference scheme for (4.50) with (4.5) with first-order one-sided difference at the boundaries reads:

$$\begin{cases} -\dfrac{a_{i-1/2}}{\tilde{h}_i h_i} w_{i-1} + \left(\dfrac{a_{i-1/2}}{\tilde{h}_i h_i} + \dfrac{a_{i+1/2}}{\tilde{h}_i h_{i+1}}\right) w_i - \dfrac{a_{i+1/2}}{\tilde{h}_i h_{i+1}} w_{i+1} \\ \qquad\qquad + b(x_i) w_i = g(x_i), \quad 1 \le i \le N-1, \\ a_- w_0 + b_- \dfrac{w_1 - w_0}{h_1} = 0, \quad a_+ w_N + b_+ \dfrac{w_N - w_{N-1}}{h_N} = 0. \end{cases} \tag{4.67}$$

We can rewrite (4.67) in the linear system:

$$A_{fd}\mathbf{w} = \mathbf{b}, \tag{4.68}$$

where A_{fd} is a non-symmetric tridiagonal matrix.

It has been shown (cf. Orszag (1980), Canuto et al. (1987), Kim and Parter (1997)) that in the Dirichlet case, A_{fd}^{-1} is an optimal preconditioner for A, but $\mathrm{cond}(A_{fd}^{-1}A)$ deteriorates with other types of boundary conditions.

Remark 4.13. *The above discussion is valid for both the Legendre and Chebyshev collocation methods.*

4.4.2.2 Finite Element Preconditioning

A more robust preconditioner can be constructed by using a finite element approximation to (4.53).

Let us denote

$$X_h = \left\{u \in H^1_\diamond(I) : u|_{[x_{i+1},x_i]} \in P_1,\ i = 0,1,\ldots,N-1\right\}. \tag{4.69}$$

Then, the piecewise linear finite element approximation to (4.53) is

$$\begin{cases} \text{Find } u_h \in X_h \text{ such that} \\ \mathscr{B}_h(u_h, v_h) = \langle g, v_h\rangle_h, \quad \forall v_h \in X_h, \end{cases} \tag{4.70}$$

where

$$\begin{aligned} \mathscr{B}_h(u_h, v_h) :=& \left\langle a u_h', v_h'\right\rangle_h + a(1)h_+ u_h(1)v_h(1) \\ & - a(-1)h_- u_h(-1)v_h(-1) + \left\langle b u_h, v_h\right\rangle_h, \end{aligned}$$

and $\langle\cdot,\cdot\rangle_h$ is an appropriate discrete inner product associated with the piecewise linear finite element approximation.

To fix the idea, we assume $b_\pm \neq 0$. Let us denote for $k = 1,2,\ldots,N-1$,

$$\hat{h}_k(x) = \begin{cases} \dfrac{x - x_{k+1}}{x_k - x_{k+1}}, & x \in [x_k, x_{k+1}], \\ \dfrac{x_{k-1} - x}{x_{k-1} - x_k}, & x \in [x_{k-1}, x_k], \\ 0, & \text{otherwise}, \end{cases} \tag{4.71}$$

and

$$\begin{aligned} \hat{h}_0(x) &= \begin{cases} \dfrac{x - x_1}{x_0 - x_1}, & x \in [x_0, x_1], \\ 0, & \text{otherwise}, \end{cases} \\ \hat{h}_N(x) &= \begin{cases} \dfrac{x_{N-1} - x}{x_{N-1} - x_N}, & x \in [x_{N-1}, x_N], \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{4.72}$$

Then

$$X_h = \operatorname{span}\left\{\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_N\right\}. \tag{4.73}$$

Setting

$$\begin{aligned} &u_h(x) = \sum_{j=0}^{N} u_h(x_j)\hat{h}_j(x), \quad \mathbf{w} = (u_h(x_0), u_h(x_1), \ldots, u_h(x_N))^T; \\ &b_{kj} = \mathscr{B}_h(\hat{h}_j, \hat{h}_k), \quad B_{fe} = (b_{kj})_{k,j=0,1,\ldots,N}; \\ &m_{kj} = \langle \hat{h}_j, \hat{h}_k \rangle_h, \quad M_{fe} = (m_{kj})_{k,j=0,1,\ldots,N}; \\ &\mathbf{g} = (g(x_0), g(x_1), \ldots, g(x_N))^T, \end{aligned} \tag{4.74}$$

we can reduce (4.70) to the following linear system

$$B_{fe}\mathbf{w} = M_{fe}\mathbf{g} \quad \text{or} \quad M_{fe}^{-1}B_{fe}\mathbf{w} = \mathbf{g}. \tag{4.75}$$

Since both (4.37) and (4.75) provide approximate solutions to (4.53), it is expected that $(M_{fe}^{-1}B_{fe})^{-1}$ (resp. B_{fe}^{-1}) is a good preconditioner for $W^{-1}B$ (resp. B). The optimality of $(M_{fe}^{-1}B_{fe})^{-1}$ as a preconditioner for $W^{-1}B$ has been shown in Franken et al. (1990), while the optimality of B_{fe}^{-1} as a preconditioner for B has been shown in Parter and Rothman (1995).

4.5 Error Estimates

In this section, we perform error analysis for several typical spectral approximation schemes proposed in the previous sections and a spectral-Galerkin method for the 1-D Helmholtz equation.

4.5.1 Legendre-Galerkin Method

We first consider the Legendre-Galerkin method (4.18) (with $f_N = I_N f$ and $\omega \equiv 1$) for (4.10) with homogeneous Dirichlet boundary conditions, i.e., $b_\pm = 0$. In this case, the error analysis is standard. Indeed, applying Theorem 1.3 with $X = H_0^1(I)$, we find immediately

$$\|u - u_N\|_1 \lesssim \inf_{v_N \in X_N} \|u - v_N\|_1 + \|f - I_N f\|.$$

Applying Theorem 3.38 with $\alpha = \beta = 0$ and Theorem 3.44 to the above leads to the following estimate.

Theorem 4.1. *Let u and u_N be the solutions of (4.10) with $b_\pm = 0$ and (4.18), respectively. If $u \in H_0^1(I), \partial_x u \in B_{0,0}^{m-1}(I)$ and $f \in B_{-1,-1}^k(I)$ with $1 \le m \le N+1$ and $1 \le k \le N+1$, we have*

$$\begin{aligned}\|u-u_N\|_1 &\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\|\partial_x^m u\|_{\omega^{m-1,m-1}}\\ &\quad + c\sqrt{\frac{(N-k+1)!}{N!}}(N+k)^{-(k+1)/2}\|\partial_x^k f\|_{\omega^{k-1,k-1}},\end{aligned} \tag{4.76}$$

where c is a positive constant independent of m,k,N,f and u.

Remark 4.14. *Recall from Remark* 3.7 *that the factor*

$$N^{(1-m)/2} \le \sqrt{\frac{(N-m+1)!}{N!}} \le (N-m+2)^{(1-m)/2}, \tag{4.77}$$

and it is of order $O(N^{(1-m)/2})$ for fixed m.

We now consider the Legendre-Galerkin method (4.18) (with $f_N = I_N f$ and $\omega = 1$) with the general boundary conditions (4.5). To handle the boundary conditions involving derivatives, we need to make use of the H_0^2-orthogonal projection: $\Pi_N^{2,0} : H_0^2(I) \to P_N \cap H_0^2(I)$, defined by

$$\big(\partial_x^2(\Pi_N^{2,0}u - u), \partial_x^2 v_N\big) = 0, \quad \forall v_N \in P_N \cap H_0^2(I), \tag{4.78}$$

whose approximation property is stated in the following lemma.

Lemma 4.7. *If $u \in H_0^2(I)$ and $\partial_x^2 u \in B_{0,0}^{m-2}(I)$ with $2 \le m \le N+1$, then we have*

$$\|\Pi_N^{2,0}u - u\|_\mu \le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{\mu-(1+m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}}, \tag{4.79}$$

for $0 \le \mu \le 2$, where c is a positive constant independent of m,N and u.

Proof. We first prove the case: $\mu = 2$. Let $\Pi_N^{1,0}$ be the H_0^1-orthogonal projection operator defined by (3.290) with $\alpha = \beta = 0$, and set

$$\phi(x) = \int_{-1}^{x} \Big(\Pi_{N-1}^{1,0}\partial_y u(y) - \frac{3}{4}(1-y^2)\phi^*\Big)dy,$$

where the constant

$$\phi^* = \int_{-1}^{1} \Pi_{N-1}^{1,0}\partial_x u(x)dx.$$

One verifies readily that $\phi \in P_N$ and $\phi(\pm 1) = \phi'(\pm 1) = 0$. Moreover, thanks to the fact $u(\pm 1) = 0$, we derive from Theorem 3.39 with $\alpha = \beta = 0$ that

$$\begin{aligned}|\phi^*| &\le \int_{-1}^{1} |\Pi_{N-1}^{1,0}\partial_x u(x) - \partial_x u(x)|dx \le \sqrt{2}\|\Pi_{N-1}^{1,0}\partial_x u - \partial_x u\|\\ &\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}},\end{aligned}$$

and

$$\|\partial_x^2(\Pi_N^{2,0}u-u)\| \overset{(4.78)}{\leq} \|\partial_x^2(\phi-u)\| \leq \|\partial_x(\Pi_{N-1}^{1,0}\partial_x u-\partial_x u)\|+c|\phi^*|$$
$$\leq c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(3-m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}},$$

which, together with the Poincaré inequality (B.21), yields the desired result with $\mu=2$.

We now use a duality argument to prove (4.79) with $\mu=0$. Given $f\in L^2(I)$, we consider the following auxiliary problem:

$$\begin{cases}\text{Find } w\in H_0^2(I) \text{ such that}\\ \mathscr{B}(w,z):=(\partial_x^2 w,\partial_x^2 z)=(f,z), \quad \forall z\in H_0^2(I),\end{cases} \tag{4.80}$$

which admits a unique solution in $H_0^2(I)$ satisfying

$$\|w\|_4\leq c\|f\|.$$

Hence, taking $z=\Pi_N^{2,0}u-u$ in (4.80), we have from the shown case (i.e., (4.79) with $\mu=2$) that

$$\begin{aligned}|(f,\Pi_N^{2,0}u-u)| &= |\mathscr{B}(\Pi_N^{2,0}u-u,\Pi_N^{2,0}w-w)|\\ &\leq \|\partial_x^2(\Pi_N^{2,0}u-u)\|\|\partial_x^2(\Pi_N^{2,0}w-w)\|\\ &\leq c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(1+m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}}\|\partial_x^4 w\|_{\omega^{2,2}}\\ &\leq c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(1+m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}}\|f\|.\end{aligned}$$

Consequently,

$$\begin{aligned}\|\Pi_N^{2,0}u-u\| &= \sup_{0\neq f\in L^2(I)}\frac{|(f,\Pi_N^{2,0}u-u)|}{\|f\|}\\ &\leq c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(1+m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}}.\end{aligned}$$

Finally, we prove the cases $0<\mu<2$ by using space interpolation. Let $\theta=1-\mu/2$. Since $H^\mu(I)=[H^2(I),L^2(I)]_\theta$, we have from the Gagliardo-Nirenberg inequality (see Theorem B.7) and (4.79) with $\mu=0,2$ that

$$\begin{aligned}\|\Pi_N^{2,0}u-u\|_\mu &\leq \|\Pi_N^{2,0}u-u\|_2^{1-\theta}\|\Pi_N^{2,0}u-u\|^\theta\\ &\leq c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{\mu-(1+m)/2}\|\partial_x^m u\|_{\omega^{m-2,m-2}}.\end{aligned}$$

This ends the proof. □

Remark 4.15. *We shall provide in Chap. 5 a much simpler proof of the above estimates by using the notion of generalized Jacobi polynomials.*

With the aid of the above lemma, we can derive the following result, which will be useful for the convergence analysis.

Theorem 4.2. *There exists a mapping $\Pi_N^2 : H^2(I) \to P_N$ such that*

$$(\Pi_N^2 u)(\pm 1) = u(\pm 1), \quad (\Pi_N^2 u)'(\pm 1) = u'(\pm 1). \tag{4.81}$$

Moreover, if $u \in H^2(I)$ and $\partial_x^2 u \in B_{0,0}^{m-2}(I)$ with $2 \le m \le N+1$, then for $0 \le \mu \le 2$, we have

$$\begin{aligned} &\|\Pi_N^2 u - u\|_\mu \\ &\quad \le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{\mu-(1+m)/2}\big(\|u\|_2 + \|\partial_x^m u\|_{\omega^{m-2,m-2}}\big), \end{aligned} \tag{4.82}$$

where c is a positive constant independent of m, N and u.

Proof. Recall the Hermite interpolation basis polynomials associated with two points $x_0 = -1$ and $x_1 = 1$:

$$\begin{aligned} H_0(x) &= \frac{(2+x)(1-x)^2}{4}, \quad H_1(x) = H_0(-x), \\ \hat{H}_0(x) &= \frac{(1+x)(1-x)^2}{4}, \quad \hat{H}_1(x) = -\hat{H}_0(-x). \end{aligned}$$

Setting

$$\Phi(x) = u(-1)H_0(x) + u(1)H_1(x) + u'(-1)\hat{H}_0(x) + u'(1)\hat{H}_1(x) \in P_3,$$

we find that $\Phi(\pm 1) = u(\pm 1)$ and $\Phi'(\pm 1) = u'(\pm 1)$. For any $u \in H^2(I)$, we have $u_* := u - \Phi \in H_0^2(I)$. Defining

$$\Pi_N^2 u = \Pi_N^{2,0} u_* + \Phi,$$

we find that $\Pi_N^2 u$ satisfies (4.81), and

$$u - \Pi_N^2 u = u_* - \Pi_N^{2,0} u_*.$$

Therefore, by Lemma 4.7,

$$\begin{aligned} \|u - \Pi_N^2 u\|_\mu &= \|u_* - \Pi_N^{2,0} u_*\|_\mu \\ &\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{\mu-(1+m)/2}\|\partial_x^m u_*\|_{\omega^{m-2,m-2}}. \end{aligned} \tag{4.83}$$

It is clear that for $m \geq 4$, we have $\partial_x^m u_* = \partial_x^m u$. For $m = 2,3$, we obtain from the Sobolev inequality (B.33) that

$$\max_{|x|\leq 1} |\partial_x^m \Phi(x)| \leq c\|u\|_2 \;\;\Rightarrow\;\; \|\partial_x^m u_*\|_{\omega^{m-2,m-2}} \leq c\big(\|u\|_2 + \|\partial_x^m u\|_{\omega^{m-2,m-2}}\big).$$

The estimate (4.82) follows. □

With the above preparations, we are ready to carry out error analysis of the Legendre-Galerkin approximation of (4.10) with general boundary conditions (4.5).

Theorem 4.3. *Let u and u_N be the solutions of (4.10) and (4.18), respectively. If $u \in H^2(I)$, $\partial_x^2 u \in B^{m-2}_{0,0}(I)$ and $f \in B^k_{-1,-1}(I)$ with $2 \leq m \leq N+1$ and $1 \leq k \leq N+1$, we have*

$$\begin{aligned}\|u - u_N\|_1 \leq\; & c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\big(\|u\|_2 + \|\partial_x^m u\|_{\omega^{m-2,m-2}}\big) \\ & + c\sqrt{\frac{(N-k+1)!}{N!}}(N+k)^{-(k+1)/2}\|\partial_x^k f\|_{\omega^{k-1,k-1}},\end{aligned} \tag{4.84}$$

where c is a positive constant independent of m,k,N,f and u.

Proof. We derive from (4.10) and (4.18) that

$$A(u-u_N, v_N) := \alpha(u-u_N, v_N) - ((u-u_N)'', v_N) = (f - I_N f, v_N), \quad \forall v_N \in X_N.$$

Under the assumption (4.3), one verifies the continuity and coercivity:

$$\begin{aligned} & A(v,w) \leq c_1\|v\|_1\|w\|_1, \quad \forall v,w \in H^2(I)\cap H^1_\diamond(I), \\ & A(v,v) \geq c_2|v|_1^2, \quad \forall v \in H^2(I)\cap H^1_\diamond(I).\end{aligned} \tag{4.85}$$

Applying Theorem 1.3 with $X = H^2(I)\cap H^1_\diamond(I)$, and using Theorems 3.44 and 4.2, we find

$$\begin{aligned}\|u-u_N\|_1 &\leq c\big(\|u - \Pi_N^2 u\|_1 + \|I_N f - f\|\big) \\ &\leq c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\big(\|u\|_2 + \|\partial_x^m u\|_{\omega^{m-2,m-2}}\big) \\ &\quad + c\sqrt{\frac{(N-k+1)!}{N!}}(N+k)^{-(k+1)/2}\|\partial_x^k f\|_{\omega^{k-1,k-1}}.\end{aligned}$$

This completes the proof. □

4.5.2 Chebyshev-Collocation Method

We consider in this section the Chebyshev-collocation method for the model equation

$$\gamma u - u_{xx} = f, \quad \text{in } (-1,1),\ \gamma > 0; \quad u(\pm 1) = 0. \tag{4.86}$$

Let $\{x_j\}_{j=0}^N$ be the Chebyshev-Gauss-Lobatto points. As shown previously, the collocation approximation is

$$\begin{cases} \text{Find } u_N \in P_N^0 \text{ such that} \\ \gamma u_N(x_j) - u_N''(x_j) = f(x_j), \quad 1 \le j \le N-1. \end{cases} \tag{4.87}$$

Let $\omega = (1-x^2)^{-1/2}$ be the Chebyshev weight function, and define the bilinear form as in (3.289):

$$a_\omega(u,v) := \big(u_x, \omega^{-1}(v\omega)_x\big)_\omega = \int_{-1}^{1} u_x (v\omega)_x \, dx. \tag{4.88}$$

We find from Lemma 3.5 (with $\alpha = \beta = -1/2$) that $a_\omega(\cdot,\cdot)$ is continuous and coercive in $H^1_{0,\omega}(I) \times H^1_{0,\omega}(I)$. As a special case of Lemma 4.5, we can reformulate the Chebyshev-collocation scheme (4.87) as

$$\begin{cases} \text{Find } u_N \in P_N^0 \text{ such that} \\ \gamma \langle u_N, v_N\rangle_{N,\omega} + a_\omega(u_N, v_N) = \langle f, v_N\rangle_{N,\omega}, \ \ \forall v_N \in P_N^0. \end{cases} \tag{4.89}$$

Then its convergence can be analyzed by using Theorem 1.3 and a standard argument.

Theorem 4.4. *If $u \in H^1_{0,\omega}(I), \partial_x u \in B^{m-1}_{-1/2,-1/2}(I)$ and $f \in B^k_{-1/2,-1/2}(I)$ with $1 \le m,k \le N+1$, then we have*

$$\begin{aligned} \|u - u_N\|_{1,\omega} \le\ & c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2} \|\partial_x^m u\|_{\omega^{m-3/2,m-3/2}} \\ & + c\sqrt{\frac{(N-k+1)!}{N!}} N^{-(1+k)/2} \|\partial_x^k f\|_{\omega^{k-1/2,k-1/2}}, \end{aligned} \tag{4.90}$$

where c is a positive constant independent of m,k,N,f and u.

Proof. Let $\Pi^{1,0}_{N,\omega}$ be the orthogonal projection operator defined in (3.290) with $\alpha = \beta = -1/2$. Applying Theorem 1.3 with $X = H^1_{0,\omega}(I)$ leads to

$$\begin{aligned} \|u - u_N\|_{1,\omega} \le c\Bigg(& \|u - \Pi^{1,0}_{N,\omega} u\|_{1,\omega} + \sup_{0 \ne v_N \in P_N^0} \frac{|(\Pi^{1,0}_{N,\omega} u, v_N)_\omega - \langle \Pi^{1,0}_{N,\omega} u, v_N\rangle_{N,\omega}|}{\|v_N\|_{1,\omega}} \\ & + \sup_{0 \ne v_N \in P_N^0} \frac{|(f, v_N)_\omega - \langle f, v_N\rangle_{N,\omega}|}{\|v_N\|_{1,\omega}} \Bigg). \end{aligned}$$

Therefore, it is necessary to estimate the error between the discrete and inner products. For this purpose, let π_N^c be the L^2_ω-orthogonal projection as defined in (3.249), and I_N^c be the Chebyshev-Gauss-Lobatto interpolation operator. Then we derive from (3.218) and Theorems 3.35 and 3.43 with $\alpha = \beta = -1/2$ that

$$\begin{aligned} |(f, v_N)_\omega - \langle f, v_N\rangle_{N,\omega}| &\le |(f - \pi_{N-1}^c f, v_N)_\omega - \langle I_N^c f - \pi_{N-1}^c f, v_N\rangle_{N,\omega}| \\ &\overset{(3.220)}{\le} c\big(\|f - \pi_{N-1}^c f\|_\omega + \|f - I_N^c f\|_\omega\big)\|v_N\|_\omega \\ &\le c\sqrt{\frac{(N-k+1)!}{N!}} N^{-(1+k)/2} \|\partial_x^k f\|_{\omega^{k-1/2,k-1/2}} \|v_N\|_\omega, \end{aligned} \tag{4.91}$$

and similarly,

$$|(\Pi_{N,\omega}^{1,0} u, v_N)_\omega - \langle \Pi_{N,\omega}^{1,0} u, v_N\rangle_{N,\omega}| \le c\big(\|\Pi_{N,\omega}^{1,0} u - u\|_\omega + \|\pi_{N-1}^c u - u\|_\omega\big)\|v_N\|_\omega.$$

Hence, the estimate (4.90) follows from Theorems 3.35 and 3.39. □

Remark 4.16. *As shown in (4.91), we have the following error estimate between the continuous and discrete inner products relative to the Chebyshev-Gauss-Lobatto setting: If $u \in B^m_{-1/2,-1/2}(I)$ with $1 \le m \le N+1$, then for any $\phi \in P_N$, we have*

$$\begin{aligned} &|(u, \phi)_\omega - \langle u, \phi\rangle_{N,\omega}| \\ &\qquad \le c\sqrt{\frac{(N-m+1)!}{N!}} N^{-(1+m)/2} \|\partial_x^m u\|_{\omega^{m-1/2,m-1/2}} \|\phi\|_\omega, \end{aligned} \tag{4.92}$$

where c is a positive constant independent of m, N, ϕ and u. This result is quite useful for error analysis of Chebyshev spectral methods.

4.5.3 Galerkin Method with Numerical Integration

We considered in previous two sections error analysis of problems with constant coefficients. We now discuss the general variable coefficient problem (4.50) with general boundary conditions (4.5), whose variational formulation is given by (4.52)–(4.53). Correspondingly, the Legendre Galerkin method with numerical integration is given by (4.54)–(4.55) with $\omega \equiv 1$. For clarity of presentation, we recall the formulation. Let

$$X_N = \big\{v \in P_N :\ a_\pm v(\pm 1) + b_\pm v'(\pm 1) = 0\big\}. \tag{4.93}$$

We look for $u_N \in X_N$ such that

$$\mathscr{B}_N(u_N, v_N) = \langle g, v_N\rangle_N, \quad \forall v_N \in X_N, \tag{4.94}$$

where

$$\begin{aligned}\mathscr{B}_N(u_N,v_N) =&\langle au_N',v_N'\rangle_N+\langle bu_N,v_N\rangle_N+a(1)h_+u_N(1)v_N(1)\\ &-a(-1)h_-u_N(-1)v_N(-1),\end{aligned}\tag{4.95}$$

with $h_\pm$ being defined in (4.7). Observe from (4.59) that for any $u_N, v_N \in X_N$,

$$\mathscr{B}_N(u_N,u_N)\ge c\|u_N'\|^2+\langle bu_N,u_N\rangle_N,\tag{4.96}$$

and

$$\big|\mathscr{B}_N(u_N,v_N)\big|\le c\|u_N\|_1\|v_N\|_1.\tag{4.97}$$

For simplicity, we assume $b(x)\ge b_0>0$, if $b_\pm\neq 0$, so we have the coercivity:

$$\mathscr{B}_N(u_N,u_N)\ge c\|u_N\|_1^2.\tag{4.98}$$

As a preparation, we first obtain the following result. As its proof is very similar to that of (4.92), we leave it as an exercise (see Problem 4.3).

Lemma 4.8. *If $u\in B^m_{-1,-1}(I)$ with $1\le m\le N+1$, then for any $\phi\in P_N$,*

$$|(u,\phi)-\langle u,\phi\rangle_N|\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(m+1)/2}\|\partial_x^m u\|_{\omega^{m-1,m-1}}\|\phi\|,\tag{4.99}$$

where c is a positive constant independent of m,N,ϕ and u.

The convergence of the scheme (4.94), under the aforementioned assumptions on $a_\pm, b_\pm$ and the variable coefficients a,b, is presented below.

Theorem 4.5. *Let u and u_N be the solutions of (4.52)–(4.53) and (4.94), respectively. If*

$$\begin{aligned}&a,b,a',b'\in L^\infty(I),\ \ u\in H^2(I),\ \ \partial_x^2u\in B^{m-2}_{0,0}(I),\\ &\partial_x(au')\in B^{m-2}_{0,0}(I),\ \ \partial_x(bu)\in B^{m-1}_{0,0}(I),\ \ g\in B^k_{-1,-1}(I),\end{aligned}\tag{4.100}$$

with $2\le m\le N+1$ and $1\le k\le N+1$, then

$$\begin{aligned}&\|u-u_N\|_1\\ &\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\Big(\|u\|_2+\|\partial_x^m u\|_{\omega^{m-2,m-2}}\\ &\quad+\|\partial_x^{m-1}(au')\|_{\omega^{m-2,m-2}}+\|\partial_x^m(bu)\|_{\omega^{m-1,m-1}}\Big)\\ &\quad+c\sqrt{\frac{(N-k+1)!}{N!}}(N+k)^{-(k+1)/2}\|\partial_x^k g\|_{\omega^{k-1,k-1}},\end{aligned}\tag{4.101}$$

where c is a positive constant only depending on the L^∞-norms of a,b,a' and b'.

Proof. Let Π_N^2 be the same as in Theorem 4.2, and set $\phi = \Pi_N^2 u$ and $e_N = u_N - \phi$. Then by (4.53) and (4.94),

$$\begin{aligned}\mathscr{B}_N(e_N, e_N) &= \mathscr{B}_N(u_N, e_N) - \mathscr{B}_N(\phi, e_N)\\ &= \langle g, e_N\rangle_N - (g, e_N) + \mathscr{B}(u, e_N) - \mathscr{B}_N(\phi, e_N).\end{aligned}$$

Using Lemma 4.8 yields

$$|\langle g, e_N\rangle_N - (g, e_N)| \le c\sqrt{\frac{(N-k+1)!}{N!}}(N+k)^{-(k+1)/2}\|\partial_x^k g\|_{\omega^{k-1,k-1}}\|e_N\|. \quad (4.102)$$

By the definitions (4.52) and (4.95),

$$\begin{aligned}\big|\mathscr{B}(u, e_N) - \mathscr{B}_N(\phi, e_N)\big| &\le \big|\big(au' - I_N(a\phi'), e_N'\big)\big| + |(bu, e_N) - \langle b\phi, e_N\rangle_N|\\ &:= T_a + T_b,\end{aligned}$$

where we used the exactness (3.189) and the property (4.81) to eliminate the boundary values. Using (3.191) and Theorem 3.44, we find that

$$\begin{aligned}T_a &\le \big|\big(au' - I_N(au'), e_N'\big)\big| + \big|\big(I_N(au' - a\phi'), e_N'\big)\big|\\ &\le \big(\|au' - I_N(au')\| + \|I_N(au' - a\phi')\|\big)\|e_N'\|\\ &\le c\Big(\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\|\partial_x^{m-1}(au')\|_{\omega^{m-2,m-2}}\\ &\qquad + \|I_N(au' - a\phi')\|\Big)\|e_N'\|.\end{aligned}$$

Moreover, by (3.191) and Lemma 4.8,

$$\begin{aligned}T_b &\le \big|\big(bu, e_N\big) - \langle bu, e_N\rangle_N\big| + \big|\langle bu - b\phi, e_N\rangle_N\big|\\ &\le c\Big(\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(m+1)/2}\|\partial_x^m(bu)\|_{\omega^{m-1,m-1}}\|e_N\|\\ &\quad + \|I_N(bu - b\phi)\|_N\|e_N\|_N\Big)\\ &\le c\Big(\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(m+1)/2}\|\partial_x^m(bu)\|_{\omega^{m-1,m-1}}\\ &\quad + \|I_N(bu - b\phi)\|\Big)\|e_N\|.\end{aligned}$$

Thanks to (4.81), we have

$$(u-\phi)(\pm 1) = 0, \quad (u-\phi)'(\pm 1) = 0.$$

Thus, we obtain from Lemma 3.11 and Theorem 4.2 that

$$\begin{aligned}\|I_N(au'-a\phi')\| &\le c(\|au'-a\phi'\|+N^{-1}\|(au'-a\phi')'\|_{\omega^{1,1}})\\ &\le c\Big(\big(\|a\|_\infty+N^{-1}\|a'\|_\infty\big)\|(u-\phi)'\|+N^{-1}\|a\|_\infty\|(u-\phi)''\|\Big)\\ &\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\big(\|u\|_2+\|\partial_x^m u\|_{\omega^{m-2,m-2}}\big),\end{aligned}$$

and

$$\begin{aligned}\|I_N(bu-b\phi)\| &\le c(\|bu-b\phi\|+N^{-1}\|(bu-b\phi)'\|_{\omega^{1,1}})\\ &\le c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{-(1+m)/2}\big(\|u\|_2+\|\partial_x^m u\|_{\omega^{m-2,m-2}}\big).\end{aligned}$$

Consequently, we derive from (4.98) and the above estimates that

$$\begin{aligned}\|e_N\|_1 \le\ & c\sqrt{\frac{(N-m+1)!}{N!}}(N+m)^{(1-m)/2}\big(\|u\|_2+\|\partial_x^m u\|_{\omega^{m-2,m-2}}+\\ &+\|\partial_x^{m-1}(au')\|_{\omega^{m-2,m-2}}+\|\partial_x^m(bu)\|_{\omega^{m-1,m-1}}\big)\\ &+c\sqrt{\frac{(N-k+1)!}{N!}}(N+k)^{-(k+1)/2}\|\partial_x^k g\|_{\omega^{k-1,k-1}}.\end{aligned}$$

We complete the proof by using the triangle inequality and Theorem 4.2. □

4.5.4 Helmholtz Equation

As the last example of this chapter, we consider the 1-D Helmholtz equation with complex-valued solution:

$$\begin{aligned}&-u''-k^2u=f, \quad r\in I:=(0,1),\\ &u(0)=0, \quad u'(1)-\mathrm{i}ku(1)=h,\end{aligned}\tag{4.103}$$

where k is called the wave number. We refer to Sect. 9.1 for more details on the background of the Helmholtz equation as well as its spectral approximation in multi-dimensional settings.

Note that this problem does not fit the general framework that we used for previous examples, since the problem is indefinite due to the negative sign in front of k^2.

The solution of (4.103) is increasingly oscillatory as k increases, so the number of unknowns in a numerical approximation should increase properly with k and it is thus important to derive error estimates with explicit dependence on k. The first step is to derive *a priori* estimates for the exact solution and characterize the dependence on k explicitly. To this end, we consider the following weak formulation of (4.103):