

Chapter One: Numerical Methods for Two-Point Boundary Value Problems

1 Finite Difference Methods

1.1 Introduction Consider the second order linear two-point boundary value problem

$$Lu(x) \equiv -u'' + p(x)u' + q(x)u = f(x), \quad x \in I, \quad (1.1)$$

$$u(0) = g_0, \quad u(1) = g_1, \quad (1.2)$$

where $I = [0, 1]$.

• We assume that the functions p, q and f are smooth on I , q is positive, and p^* and q_* are positive constants such that

$$|p(x)| \leq p^*, \quad 0 < q_* \leq q(x), \quad x \in I.$$

• Let $\pi = \{x_j\}_{j=0}^{N+1}$ denote a uniform partition of the interval I such that $x_j = jh, j = 0, 1, \dots, N+1$, and $(N+1)h = 1$. Finite difference method (FD) is to find an approximation to the solution of (1.1)-(1.2) on mesh points. Let $\{u_j\}_{j=0}^{N+1}$ define a *mesh function*. On this partition, the solution u of (1.1)-(1.2) is approximated by the mesh function $\{u_j\}_{j=0}^{N+1}$ defined by the finite difference equations

$$L_h u_j \equiv -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + p_j \frac{u_{j+1} - u_{j-1}}{2h} + q_j u_j = f_j, \quad j = 1, \dots, N, \quad (1.3)$$

$$u_0 = g_0, \quad u_{N+1} = g_1, \quad (1.4)$$

where

$$p_j = p(x_j), \quad q_j = q(x_j), \quad f_j = f(x_j).$$

This mesh function is a FD solution of (1.1)-(1.2) and $u_j \approx u(x_j)$. Equations (1.3) are obtained by replacing the derivatives in (1.1) by basic centered difference quotients.

$$u''(x_j) \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1}))}{h^2} \approx \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}$$

$$u'(x_j) \approx \frac{u(x_{j+1}) - u(x_{j-1}))}{2h} \approx \frac{u_{j+1} - u_{j-1}}{2h}$$

We now show that under certain conditions the difference problem (1.3)-(1.4) has a unique solution $\{u_j\}_{j=0}^{N+1}$.

1.2 The Uniqueness of the difference approximation From (1.3), we obtain

$$h^2 L_h u_j = -\left(1 + \frac{h}{2} p_j\right) u_{j-1} + (2 + h^2 q_j) u_j - \left(1 - \frac{h}{2} p_j\right) u_{j+1} = h^2 f_j, \quad j = 1, \dots, N \quad (1.5)$$

or in matrix form,

$$A\mathbf{u} = \mathbf{b},$$

where

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}, A = \begin{bmatrix} d_1 & e_1 & & & \\ c_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & e_{N-1} \\ & & & c_N & d_N \end{bmatrix}, \mathbf{b} = h^2 \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{bmatrix} - \begin{bmatrix} c_1 g_0 \\ 0 \\ \vdots \\ 0 \\ e_N g_1 \end{bmatrix},$$

and, for $j = 1, \dots, N$,

$$c_j = -\left(1 + \frac{1}{2}hp_j\right), \quad d_j = 2 + h^2q_j, \quad e_j = -\left(1 - \frac{1}{2}hp_j\right).$$

• We only need to prove that the matrix A is nonsingular, or $\det(A) \neq 0$, or $Ax = 0$ has only zero solution.

• A matrix $B = (b_{ij})_{i,j=1}^N$ is *diagonally dominant* if

$$|b_{ii}| \geq \sum_{j \neq i} |b_{ij}|, \quad i = 1, 2, \dots, N$$

and *strictly diagonally dominant* if

$$|b_{ii}| > \sum_{j \neq i} |b_{ij}|, \quad i = 1, 2, \dots, N$$

The matrix A defined above is *tridiagonal* and strictly diagonal dominant if

$$h < 2/p^*$$

which is called a mesh spacing restriction.

Theorem 1.1 *A strictly diagonal dominant matrix is nonsingular.*

Proof Assignment.

The case of $p = 0$ is given in Assignment #1.

1.3 Consistency, Stability and Convergence To study the accuracy and the computability of the difference approximation $\{u_j\}_{j=0}^{N+1}$, we introduce the concepts of consistency, stability and convergence of finite difference methods.

Definition 1.1 (Consistency) *For any smooth function w on I , we define an operator*

$$\tau_{j,\pi}[w] \equiv L_h w(x_j) - Lw(x_j), \quad j = 1, \dots, N,$$

Then the difference problem (1.3)-(1.4) is consistent with the differential problem (1.1)-(1.2) if

$$|\tau_{j,\pi}[w]| \rightarrow 0 \text{ as } h \rightarrow 0.$$

The quantities $\tau_{j,\pi}[w]$, $j = 1, \dots, N$, are called the local truncation (or truncation error).

Definition 1.2 The difference problem (1.1)–(1.2) is locally p^{th} -order accurate if, for sufficiently smooth data, there exists a positive constant C , independent of h , such that

$$\max_{1 \leq j \leq N} |\tau_{j,\pi}[w]| \leq Ch^p.$$

The following lemma demonstrates that the difference problem (1.3)–(1.4) is consistent with (1.1)–(1.2) and is locally second-order accurate.

Lemma 1.1 If $w \in C^4(I)$, then

$$\tau_{j,\pi}[w] = -\frac{h^2}{12}[w^{(4)}(\nu_j) - 2p(x_j)w^{(3)}(\theta_j)],$$

where ν_j and θ_j lie in (x_{j-1}, x_{j+1}) .

Proof — By definition

$$\begin{aligned} \tau_{j,\pi}[w] = & - \left[\frac{w(x_{j+1}) - 2w(x_j) + w(x_{j-1}))}{h^2} - w''(x_j) \right] \\ & + p_j \left[\frac{w(x_{j+1}) - w(x_{j-1}))}{2h} - w'(x_j) \right], \quad j = 1, \dots, N. \end{aligned}$$

It is easy to show using Taylor's theorem that

$$\frac{w(x_{j+1}) - w(x_{j-1}))}{2h} - w'(x_j) = \frac{h^2}{6}w^{(3)}(\theta_j), \quad \theta_j \in (x_{j-1}, x_{j+1}).$$

Also,

$$\frac{w(x_{j+1}) - 2w(x_j) + w(x_{j-1}))}{h^2} - w''(x_j) = \frac{h^2}{12}w^{(4)}(\nu_j), \quad \nu_j \in (x_{j-1}, x_{j+1}).$$

The desired result now follows immediately.

Definition 1.3 (Stability) The linear difference operator L_h is **stable** if, for sufficiently small h , there exists a constant K , independent of h , such that

$$|v_j| \leq K \{ \max(|v_0|, |v_{N+1}|) + \max_{1 \leq i \leq N} |L_h v_i| \} \quad j = 0, \dots, N+1,$$

for any mesh function $\{v_j\}_{j=0}^{N+1}$.

- Stability in general: the solution is bounded by the right-hand side and boundary values.
- Maximal principle for elliptic PDEs. We consider a simple case

$$Lu = -u''(x) = f(x) \quad x \in (0, 1)$$

Assume that $u(x)$ is continuous in $[0, 1]$ and

$$|u(x^*)| = \max |u(x)|$$

Then

$$\begin{aligned} u(1) &= u(x^*) + (1 - x^*)u'(x^*) + (1 - x^*)^2 u''(\xi_1)/2 \\ u(0) &= u(x^*) + (0 - x^*)u'(x^*) + (0 - x^*)^2 u''(\xi)/2 \end{aligned}$$

If $x^* \in (0, 1)$, we have

$$x^*u(1) + (1 - x^*)u(0) = u(x^*) + ((1 - x^*)x^{*2} + x^*(1 - x^*)^2)u''(\xi)/2$$

and

$$|u(x^*)| \leq K \{ \max\{|u(0)|, |u(1)|\} + \max_{x \in (0,1)} Lu(x) \}$$

which is the Maximal principle for the differential operator L . It is also true for more general elliptic differential operator.

Here we study the difference operator L_h and prove that, for h sufficiently small, the difference operator L_h of (1.3) is stable, or equivalently has the discrete maximal principle.

Theorem 1.2 *The difference operator L_h of (1.3) is stable for $h < 2/p^*$, with $K = \max\{1, 1/q_*\}$.*

Proof — If

$$|v_{j^*}| = \max_{0 \leq j \leq N+1} |v_j|, \quad 1 \leq j^* \leq N,$$

then, from (1.5), we obtain

$$d_{j^*}v_{j^*} = -e_{j^*}v_{j^*+1} - c_{j^*}v_{j^*-1} + h^2 L_h v_{j^*}.$$

Thus,

$$d_{j^*}|v_{j^*}| \leq (|e_{j^*}| + |c_{j^*}|) |v_{j^*}| + h^2 \max_{1 \leq j \leq N} |L_h v_j|.$$

If $h < 2/p^*$, then

$$d_{j^*} = |e_{j^*}| + |c_{j^*}| + h^2 q_{j^*},$$

and it follows that

$$h^2 q_{j^*} |v_{j^*}| \leq h^2 \max_{1 \leq j \leq N} |L_h v_j|,$$

or

$$|v_{j^*}| \leq \frac{1}{q_*} \max_{1 \leq j \leq N} |L_h v_j|.$$

Thus, if $\max_{0 \leq j \leq N+1} |v_j|$ occurs for $1 \leq j \leq N$ then

$$\max_{0 \leq j \leq N+1} |v_j| \leq \frac{1}{q_*} \max_{1 \leq j \leq N} |L_h v_j|,$$

and clearly

$$\max_{0 \leq j \leq N+1} |v_j| \leq K \{ \max(|v_0|, |v_{N+1}|) + \max_{1 \leq j \leq N} |L_h v_j| \}$$

with $K = \max\{1, 1/q_*\}$. If $\max_{0 \leq j \leq N+1} |v_j| = \max\{|v_0|, |v_{N+1}|\}$, then the above equation follows immediately. ■

- For $q \geq 0$, see Assignment.
- An immediate consequence of stability is the uniqueness (and hence existence since the problem is linear) of the difference approximation $\{u_j\}_{j=0}^{N+1}$ (which was proved by other means in earlier), for if there were two solutions, their difference $\{v_j\}_{j=0}^{N+1}$, say, would satisfy

$$L_h v_j = 0, \quad j = 1, \dots, N, \quad v_0 = v_{N+1} = 0.$$

Stability then implies that $v_j = 0$, $j = 0, 1, \dots, N + 1$.

Definition 1.4 (Convergence) *Let u be the solution of the boundary value problem (1.1)-(1.2) and $\{u_j\}_{j=0}^{N+1}$ the difference approximation defined by (1.3)-(1.4). The difference approximation converge to u if*

$$\max_{1 \leq j \leq N} |u_j - u(x_j)| \rightarrow 0,$$

as $h \rightarrow 0$. The difference $u_j - u(x_j)$ is the global (or discretization) error at the point x_j , $j = 1, \dots, N$.

Definition 1.5 *The difference approximation $\{u_j\}_{j=0}^{N+1}$ is a p^{th} -order approximation to the solution u of (1.1)-(1.2) if, for h sufficiently small, there exists a constant C independent of h , such that*

$$\max_{0 \leq j \leq N+1} |u_j - u(x_j)| \leq Ch^p.$$

- For a linear problem and a consistent method, stability implies convergence in general. Here we only consider the case (1.1)-(1.4). The basic result connecting consistency, stability and convergence is given in the following theorem. The proof shows that the convergence follows the stability of difference method.

Theorem 1.3 *Suppose $u \in C^4(I)$ and $h < 2/p^*$. Then the difference solution $\{u_j\}_{j=0}^{N+1}$ of (1.3)-(1.4) is convergent to the solution u of (1.1)-(1.2). Moreover,*

$$\max_{0 \leq j \leq N+1} |u_j - u(x_j)| \leq Ch^2.$$

Proof — Under the given conditions, the difference problem (1.3)-(1.4) is consistent with the boundary value problem (1.1)-(1.2) and the operator L_h is stable.

Since

$$L_h[u_j - u(x_j)] = f(x_j) - L_h u(x_j) = Lu(x_j) - L_h u(x_j) = -\tau_{j,\pi}[u],$$

and $u_0 - u(x_0) = u_{N+1} - u(x_{N+1}) = 0$, the stability of L_h implies that

$$|u_j - u(x_j)| \leq \frac{1}{q_*} \max_{1 \leq j \leq N} |\tau_{j,\pi}[u]|.$$

The desired result follows from Lemma \blacksquare

It follows from this theorem that $\{u_j\}_{j=0}^{N+1}$ is a second-order approximation to the solution u of (1.1).

1.4 Numerical estimate for the asymptotic convergence rate If a norm of the global error is $O(h^p)$ then an estimate of p can be determined in the following way. For ease of exposition, we use a different notation in this section and denote by $\{u_j^h\}$ the difference approximation computed with a mesh length of h . Also let

$$e^h \equiv \|u - u^h\| = \max_j |u(x_j) - u_j^h|.$$

If $e^h = O(h^p)$, then, for h sufficiently small, $h < h_0$ say, there exists a constant C , independent of h , such that

$$e^h \approx Ch^p.$$

If we solve the difference problem with two different mesh lengths h_1 and h_2 such that $h_2 < h_1 < h_0$, then

$$e^{h_1} \approx Ch_1^p$$

and

$$e^{h_2} \approx Ch_2^p$$

from which it follows that

$$\ln e^{h_1} \approx p \ln h_1 + \ln C$$

and

$$\ln e^{h_2} \approx p \ln h_2 + \ln C.$$

Therefore an estimate of p can be calculated from

$$p \approx \ln(e^{h_1}/e^{h_2}) / \ln(h_1/h_2)$$

In practice, one usually solves the difference problem for a sequence of values of h , $h_0 > h_1 > h_2 > h_3 > \dots$, and calculates the ratio on the right hand side above for successive pairs of values of h . These ratios converge to the value of p as $h \rightarrow 0$.

1.5 Higher-Order Finite Difference Approximations

- We have used the approximation

$$u''(x_j) \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1}))}{h^2} \approx \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}$$

$$u'(x_j) \approx \frac{u(x_{j+1}) - u(x_{j-1}))}{2h} \approx \frac{u_{j+1} - u_{j-1}}{2h}$$

which is based on low-order interpolation of u . One way to get a high-order FD method is to use high-order interpolation. This makes the method more complicated, particularly near boundary points. We introduce three approaches below.

- Richardson extrapolation. We consider a general problem where there exists an expansion of the local truncation error of the form

$$\tau_{j,\pi}[u] = h^2\tau[u(x_j)] + O(h^4).$$

where $u(x_j)$ is a mesh function. In fact, for the FD method and two-point BVP mentioned above,

$$\tau_{j,\pi}[u] = -\frac{h^2}{12} \left(u^{(4)}(x_j) - 2p(x_j)u^{(3)}(x_j) \right) + O(h^4) \quad (1.6)$$

We **assume** that there exists a mesh function $e(x_j)$ such that

$$u(x_j) = u_j + h^2e(x_j) + O(h^4), \quad j = 0, 1, \dots, N + 1 \quad (1.7)$$

where $u(x)$ is the exact solution and u_j be a finite difference solution. (For the problem (1.3)-(1.4), the next theorem will show that the above equation holds).

Now let $\{u_j^h\}_{j=0}^{N+1}$ denote the solution with a uniform mesh of meshsize h . In Richardson extrapolation, the difference problem (1.3)–(1.4) is solved twice with mesh spacings h and $h/2$ to yield difference solutions $\{u_j^h\}_{j=0}^{N+1}$, and $\{u_j^{h/2}\}_{j=0}^{2(N+1)}$. Then at a point $\hat{x} = jh = (2j)(h/2)$ common to both meshes we have, from (1.7),

$$u(\hat{x}) = u_j^h + h^2e(\hat{x}) + O(h^4)$$

and

$$u(\hat{x}) = u_{2j}^{h/2} + \frac{h^2}{4}e(\hat{x}) + O(h^4),$$

from which it follows that

$$u(\hat{x}) = u_{2j}^{h/2} + \frac{1}{3}(u_{2j}^{h/2} - u_j^h) + O(h^4).$$

Thus

$$u_j^R \equiv u_{2j}^{h/2} + \frac{1}{3}(u_{2j}^{h/2} - u_j^h)$$

is a fourth–order approximation to $u(x_j)$, $j = 0, \dots, N + 1$.

Theorem 1.4 *Suppose $u \in C^6(I)$, and $h < 2/p^*$. Then (1.6) and (1.7) hold with $e(x)$ defined as the solution of the boundary value problem*

$$\begin{aligned} Le(x) &= \tau[u(x)], & x \in I, \\ e(0) &= e(1) = 0, \end{aligned}$$

where

$$\tau[u(x)] = -\frac{1}{12}[u^{(4)}(x) - 2p(x)u^{(3)}(x)].$$

Proof — Since $u \in C^6(I)$, it is easy to show by extending the argument used in Lemma that (1.7) holds with $\tau[u(x)]$ defined above.

As in the proof of Theorem 1.3, we have

$$L_h[u(x_j) - u_j] = \tau_{j,\pi}[u], \quad j = 1, \dots, N.$$

With (1.7) in the above equation, we obtain

$$\begin{aligned} L_h[u(x_j) - u_j] &= h^2\tau[u(x_j)] + O(h^4) \\ &= h^2Le(x_j) + O(h^4) \\ &= h^2[Le(x_j) - L_h e(x_j)] + h^2L_h e(x_j) + O(h^4) \\ &= h^2\tau_{j,\pi}[e] + h^2L_h e(x_j) + O(h^4). \end{aligned}$$

From the smoothness properties of the functions p, q , and τ , it follows that the solution $e(x)$ is unique. Moreover, since $\tau \in C^2(I)$, $e(x) \in C^4(I)$ and $\tau_{j,\pi}[e] = O(h^2)$. Thus, we have

$$L_h[u(x_j) - \{u_j + h^2e(x_j)\}] = O(h^4).$$

The desired result follows from the stability of L_h . ■

- Deferred corrections. Suppose $\hat{\tau}_{j,\pi}[\cdot]$ is a difference operator such that

$$|\tau[u(x_j)] - \hat{\tau}_{j,\pi}[u^h]| = O(h^2),$$

where u^h denotes the solution $\{u_j\}_{j=0}^{N+1}$ of (1.3)-(1.4). We define the mesh function $\{\hat{u}_j\}_{j=0}^{N+1}$ by

$$\begin{aligned} L_h\hat{u}_j &= f_j + h^2\hat{\tau}_{j,\pi}[u^h], \quad j = 1, \dots, N, \\ \hat{u}_0 &= g_0, \quad \hat{u}_{N+1} = g_1. \end{aligned}$$

Then it is easy to show that $\{\hat{u}_j\}_{j=0}^{N+1}$ is a fourth-order approximation to $u(x)$, since

$$\begin{aligned} L_h[\hat{u}_j - u(x_j)] &= f_j + h^2\hat{\tau}_{j,\pi}[u^h] - L_h u(x_j) \\ &= [Lu(x_j) - L_h u(x_j)] + h^2\hat{\tau}_{j,\pi}[u^h] \\ &= -\tau_{j,\pi}[u] + h^2\hat{\tau}_{j,\pi}[u^h] \\ &= -h^2\{\tau[u(x_j)] - \hat{\tau}_{j,\pi}[u^h]\} + O(h^4). \end{aligned}$$

By the stability of L_h , it follows that

$$|\hat{u}_j - u(x_j)| = O(h^4).$$

The main problem in deferred corrections is the construction of second-order difference approximations $\hat{\tau}_{j,\pi}[u^h]$ to the truncation error term $\tau[u(x_j)]$. One way is to replace the

derivatives appearing in $\tau[u(x)]$ by standard difference approximations using the finite difference solution $\{u_j\}_{j=0}^{N+1}$ in place of the exact solution $u(x)$. One problem with this approach is that it requires some modification near the end-points of the interval, or it is necessary to compute the numerical solution outside the interval I . A second approach is to use the differential equation to express $\tau[u(x)]$ in the form

$$\tau[u(x)] = C_2(x)u'(x) + C_1(x)u(x) + C_0(x),$$

where the functions C_0, C_1, C_2 are expressible in terms of the functions p, q, f and their derivatives. Then choose

$$\hat{\tau}_{j,\pi}[u^h] = C_2(x_j) \frac{u_{j+1} - u_{j-1}}{2h} + C_1(x_j)u_j + C_0(x_j).$$

Since

$$u'' = pu' + qu - f,$$

we have

$$\begin{aligned} u^{(3)} &= pu'' + p'u' + qu' + q'u - f' \\ &= p(pu' + qu - f) + (p' + q)u' + q'u - f' \\ &= (p^2 + p' + q)u' + (pq + q')u - (pf + f') \\ &\equiv Pu' + Qu - F. \end{aligned}$$

Similarly,

$$u^{(4)} = (Pp + P' + Q)u' + (Pq + Q')u - (Pf + F').$$

We obtain the desired form.

In deferred corrections, the linear algebraic systems defining the basic difference approximation and the fourth-order approximation have the same coefficient matrix, which simplifies the algebraic problem.

If the solution u of the boundary value problem (1.1)–(1.2) is sufficiently smooth, then it can be shown that the local truncation error $\tau_{j,\pi}[u]$ has an asymptotic expansion of the form

$$\tau_{j,\pi}[u] = \sum_{\nu=1}^m h^{2\nu} \tau_{\nu}[u(x_j)] + O(h^{2m+2}),$$

and, if $m > 1$, deferred corrections can be extended to compute approximations of higher order than 4.

- Numerov's method (compact scheme).

Here we consider the simple case $p = 0$. other cases can be discussed similarly. A higher-order finite difference method is easily constructed in this case that (1.1)–(1.2) becomes

$$\begin{aligned} Lu(x) &= -u''(x) + q(x)u(x) = f(x), \quad x \in I, \\ u(0) &= g_0, \quad u(1) = g_1. \end{aligned}$$

Now let

$$\begin{aligned}\Delta_h u(x_j) &\equiv [u(x_{j+1}) - 2u(x_j) + u(x_{j-1}))]/h^2 \\ &= u''(x_j) + \frac{1}{12}h^2 u^{(4)}(x_j) + O(h^4).\end{aligned}$$

From the differential equation, we have

$$u^{(4)}(x) = [q(x)u(x) - f(x)]''$$

and therefore

$$u^{(4)}(x_j) = \Delta_h [q(x_j)u(x_j) - f(x_j)] + O(h^2).$$

Thus,

$$u''(x_j) = \Delta_h u(x_j) - \frac{1}{12}h^2 \Delta_h [q(x_j)u(x_j) - f(x_j)] + O(h^4).$$

We define $\{\tilde{u}_j\}_{j=0}^{N+1}$ by

$$\begin{aligned}\mathcal{L}_h \tilde{u}_j &\equiv -\Delta_h \tilde{u}_j + \left(1 + \frac{1}{12}h^2 \Delta_h\right) q_j \tilde{u}_j = \left(1 + \frac{1}{12}h^2 \Delta_h\right) f(x_j), \quad j = 1, \dots, N, \\ \tilde{u}_0 &= g_0, \quad \tilde{u}_{N+1} = g_{N+1},\end{aligned}$$

which is commonly known as *Numerov's method*. The coefficient matrix of the linear system is symmetric tridiagonal and may be written in the form

$$\tilde{c}_j \tilde{u}_{j-1} + \tilde{d}_j \tilde{u}_j + \tilde{e}_j \tilde{u}_{j+1} = \frac{1}{12}h^2 [f_{j+1} + 10f_j + f_{j-1}], \quad j = 1, \dots, N,$$

where

$$\tilde{c}_j = -\left(1 - \frac{1}{12}h^2 q_{j-1}\right), \quad \tilde{d}_j = 2 + \frac{5}{6}h^2 q_j, \quad \tilde{e}_j = -\left(1 - \frac{1}{12}h^2 q_{j+1}\right).$$

It is easy to show that, for h sufficiently small, the coefficient matrix of this system is strictly diagonally dominant and hence $\{\tilde{u}_j\}_{j=0}^{N+1}$ is unique. From a similar analysis, it follows that the difference operator \mathcal{L}_h is stable. Also, since

$$\mathcal{L}_h [\tilde{u}_j - u(x_j)] = \left(1 + \frac{1}{12}h^2 \Delta_h\right) f(x_j) - \mathcal{L}_h u(x_j) = O(h^4),$$

the stability of \mathcal{L}_h implies that

$$|\tilde{u}_j - u(x_j)| = O(h^4).$$

1.6 Second-Order Nonlinear Equations We consider the second-order nonlinear two-point boundary value problem

$$\begin{aligned}\mathcal{L}u(x) &\equiv -u'' + f(x, u) = 0, \quad x \in I, \\ u(0) &= g_0, \quad u(1) = g_1.\end{aligned}\tag{1.8}$$

The basic second-order finite difference approximation to (1.8) takes the form

$$\begin{aligned}\mathcal{L}_h u_j &\equiv -\Delta_h u_j + f(x_j, u_j) = 0, \quad j = 1, \dots, N. \\ u_0 &= g_0, \quad u_{N+1} = g_1.\end{aligned}$$

If

$$\tau_{j,\pi}[u] \equiv \mathcal{L}_h u(x_j) - \mathcal{L}u(x_j)$$

then clearly

$$\tau_{j,\pi}[u] = -\frac{h^2}{12}u^{(4)}(\xi_j), \quad \xi_j \in (x_{j-1}, x_{j+1}),$$

if $u \in C^4(I)$.

- Stability of the nonlinear difference problem is defined in the following way.

Definition 1.6 *A difference problem defined by the nonlinear difference operator \mathcal{L}_h is stable if, for sufficiently small h , there exists a positive constant K , independent of h , such that, for all mesh functions $\{v_j\}_{j=0}^{N+1}$ and $\{w_j\}_{j=0}^{N+1}$,*

$$|v_j - w_j| \leq K \left\{ \max(|v_0 - w_0|, |v_{N+1} - w_{N+1}|) + \max_{1 \leq i \leq N} |\mathcal{L}_h v_j - \mathcal{L}_h w_j| \right\}.$$

For a linear operator, this definition reduces to the definition in section 1.3 applied to the mesh function $\{v_j - w_j\}_{j=0}^{N+1}$.

Theorem 1.5 *If $f_u \equiv \frac{\partial f}{\partial u}$ is continuous on $I \times (-\infty, \infty)$ such that*

$$0 < q_* \leq f_u,$$

then the difference problem is stable, with $K = \max(1, 1/q_)$*

Proof — For mesh functions $\{v_j\}$ and $\{w_j\}$, we have

$$\begin{aligned}\mathcal{L}_h v_j - \mathcal{L}_h w_j &= -\Delta_h(v_j - w_j) + f(x_j, v_j) - f(x_j, w_j) \\ &= -\Delta_h(v_j - w_j) + f_u(x_j, \hat{v}_j)(v_j - w_j),\end{aligned}\tag{1.9}$$

on using the mean value theorem, where \hat{v}_j lies between v_j and w_j . Then

$$\begin{aligned}h^2[\mathcal{L}_h v_j - \mathcal{L}_h w_j] &= c_j[v_{j-1} - w_{j-1}] + d_j[v_j - w_j] \\ &\quad + e_j[v_{j+1} - w_{j+1}]\end{aligned}\tag{1.10}$$

where $c_j = e_j = -1$ and

$$d_j = 2 + h^2 f_u(x_j, \hat{v}_j).$$

Clearly

$$|c_j| + |e_j| = 2 < |d_j|.$$

The remainder of the proof is similar to that of Theorem 1.2. ■

- Consider now a system of the N equations

$$\phi_i(u_1, u_2, \dots, u_N) = 0, \quad i = 1, \dots, N,$$

for the unknowns u_1, u_2, \dots, u_N , which we may write in vector form as

$$\mathbf{\Phi}(\mathbf{u}) = \mathbf{0}.$$

If we linearize the i^{th} equation, we obtain

$$\phi_i(u_1^{(0)}, u_2^{(0)}, \dots, u_N^{(0)}) + \sum_{j=1}^N \frac{\partial \phi_i}{\partial u_j}(u_1^{(0)}, \dots, u_N^{(0)})(u_j^{(1)} - u_j^{(0)}) = 0, \quad i = 1, \dots, N.$$

If $\mathcal{J}(\mathbf{u})$ denotes the matrix with (i, j) element $\frac{\partial \phi_i}{\partial u_j}(\mathbf{u})$, then the nonlinear system can be written in the form

$$\mathcal{J}(\mathbf{u}^{(0)}) \Delta \mathbf{u} = -\mathbf{\Phi}(\mathbf{u}^{(0)}).$$

If $\mathcal{J}(\mathbf{u}^{(0)})$ is nonsingular, then

$$\Delta \mathbf{u} = -[\mathcal{J}(\mathbf{u}^{(0)})]^{-1} \mathbf{\Phi}(\mathbf{u}^{(0)}),$$

and

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + \Delta \mathbf{u}$$

is taken as the new approximation. If the matrices $\mathcal{J}(\mathbf{u}^{(\nu)})$, $\nu = 1, 2, \dots$, are nonsingular, one hopes to determine a sequence of successively better approximations $\mathbf{u}^{(\nu)}$, $\nu = 1, 2, \dots$ from the algorithm

$$\mathbf{u}^{(\nu+1)} = \mathbf{u}^{(\nu)} + \Delta \mathbf{u}^{(\nu)}, \quad \nu = 0, 1, 2, \dots,$$

where $\Delta \mathbf{u}^{(\nu)}$ is obtained by solving the system of linear equations

$$\mathcal{J}(\mathbf{u}^{(\nu)}) \Delta \mathbf{u}^{(\nu)} = -\mathbf{\Phi}(\mathbf{u}^{(\nu)}).$$

This procedure is known as Newton's method for the solution of the system of nonlinear equations. It can be shown that as in the scalar case this procedure converges quadratically if $\mathbf{u}^{(0)}$ is chosen sufficiently close to \mathbf{u} .

- Now consider the system of the nonlinear FD equations

$$\phi_i(u_1, \dots, u_N) = -u_{i-1} + 2u_i - u_{i+1} + h^2 f(x_i, u_i) \quad i = 1, \dots, N,$$

In this case, the (i, j) element of the Jacobian \mathcal{J} is given by

$$\mathcal{J}(\mathbf{u}) = J + h^2 F(\mathbf{u})$$

where

$$F(\mathbf{u}) = \text{diag}(f_u(x_i, u_i)).$$

In this case, Newton's method becomes

$$\begin{aligned} [J + h^2 F(\mathbf{u}^{(\nu)})] \Delta \mathbf{u}^{(\nu)} &= - [J \mathbf{u}^{(\nu)} + h^2 \mathbf{f}(\mathbf{u}^{(\nu)}) - \mathbf{g}], \\ \mathbf{u}^{(\nu+1)} &= \mathbf{u}^{(\nu)} + \Delta \mathbf{u}^{(\nu)}, \quad \nu = 0, 1, 2, \dots \end{aligned}$$

1.8. Other boundary conditions There are several different boundary conditions

$$\begin{aligned} u(0) &= 0 \quad u'(1) = 0, \\ u'(0) &= 0 \quad u'(1) = 0, \\ \alpha u(0) + \beta u'(0) &= \gamma \quad u(1) = 0, \end{aligned}$$

nonlocal boundary condition

$$\int_0^1 u(x) dx = g_0, \quad u(0) = g_1$$

and the periodic boundary condition

$$u(0) = u(1) \quad u'(0) = u'(1).$$

Example

Let's consider such a boundary value problem

$$\begin{aligned} Lu(x) &\equiv -u'' + p(x)u' + q(x)u = f(x), \quad x \in I, \\ u(0) &= 0, \quad u'(1) = 0, \end{aligned}$$

where $I = [0, 1]$.

And the finite difference equations are:

$$\begin{aligned} L_h u_j &\equiv -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + p_j \frac{u_{j+1} - u_{j-1}}{2h} + q_j u_j = f_j, \quad j = 1, \dots, N, \\ u_0 &= 0, \quad -2u_N + (2 + h^2 q_{N+1})u_{N+1} = h^2 f_{N+1}, \end{aligned}$$

where

$$p_j = p(x_j), \quad q_j = q(x_j), \quad f_j = f(x_j).$$

The whole system can also be written in a matrix form.

$$A\mathbf{u} = \mathbf{b},$$

where

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \\ u_{N+1} \end{bmatrix}, \quad A = \begin{bmatrix} d_1 & e_1 & & & & \\ c_2 & \ddots & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & e_N & \\ & & & -2 & 2 + h^2 q_{N+1} & \end{bmatrix}, \quad \mathbf{b} = h^2 \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ f_{N+1} \end{bmatrix},$$

where

$$(i)\alpha_1 = d_1,$$

$$(ii)\gamma_i = e_i/\alpha_i, \quad \alpha_{i+1} = d_{i+1} - c_{i+1}\gamma_i, \quad i = 1, \dots, N-1.$$

• We now show that conditions (1.11) ensure that the quantities α_i and γ_i are bounded and that the bounds are independent of the order of the matrix A .

Theorem 1.6 *If the elements of A satisfies conditions (1.11), then*

$$|\gamma_i| < 1,$$

and

$$0 < |d_i| - |c_i| < |\alpha_i| < |d_i| + |c_i|, \quad i = 2, \dots, N.$$

Proof —

Since $\gamma_1 = e_1/\alpha_1 = e_1/d_1$,

$$|\gamma_1| = |e_1|/|d_1| < 1,$$

from (1.11(i)). Now suppose that

$$|\gamma_j| < 1, \quad j = 1, \dots, i-1.$$

Then, with (1.11 (iii)) in (1.11 (ii)), we have

$$\gamma_i = e_i/(d_i - c_i\gamma_{i-1})$$

and

$$|\gamma_i| = |e_i|/(|d_i| - |c_i||\gamma_{i-1}|) < |e_i|/(|d_i| - |c_i|).$$

Thus, using (1.11 (ii)), it follows that $|\gamma_i| < 1$, and the theorem follows by induction.

• Procedure of algorithm

$$v_1 = b_1/\alpha_1$$

For $i = 2$ **to** N **do**

$$v_i = (b_i - c_i v_{i-1})/\alpha_i$$

end

$$u_N = v_N$$

For $i = 1$ **to** $N - 1$ **do**

$$u_{N-i} = v_{N-i} - \gamma_{N-i} u_{N-i+1}$$

end

2 The Finite Element Method (FEM)

2.1 Introduction

Consider a simple two-point boundary value problem of the form

$$\begin{aligned} Lu &\equiv -u'' + q(x)u = f(x), & x \in I := [0, 1]. \\ u(0) &= u(1) = 0. \end{aligned} \tag{2.1}$$

where the function $q > 0$ and f are smooth on I . The above boundary value problem has a unique solution. Let $L^2(I)$ be the space of all functions satisfying

$$\int_0^1 (u(x))^2 dx \leq C$$

where C is a constant. For $u, v \in L^2(I)$, let

$$\int_0^1 uv dx = (u, v).$$

Let $H_0^1(I)$ denote the space of all piecewise continuously differentiable functions on I which vanish at 0 and 1. If $v \in H_0^1(I)$, then

$$-u''v + quv = fv,$$

and

$$\int_0^1 [-u''v + quv] dx = \int_0^1 fvd x.$$

On integrating by parts, we obtain

$$\int_0^1 u'v' dx - [u'v]_0^1 + \int_0^1 quv dx = \int_0^1 fvd x.$$

Since $v(0) = v(1) = 0$, we have

$$(u', v') + (qu, v) = (f, v), \quad v \in H_0^1(I), \tag{2.2}$$

where

$$(\varphi, \psi) = \int_0^1 \varphi(x)\psi(x) dx.$$

• A variational model (V). The equation (2.2) is called the **weak form** of the boundary value problem (2.1), and is written in the form

$$a(u, v) = (f, v), \quad v \in H_0^1(I),$$

where

$$a(\phi, \psi) = (\phi', \psi') + (q\phi, \psi), \quad \phi, \psi \in H_0^1(I).$$

A **variational model** of (1.1) (Galerkin method) is to find $u \in H_0^1(I)$ such that for any $v \in H_0^1(I)$,

$$a(u, v) = (f, v).$$

- Minimization model (M). Let

$$F(u) = \frac{1}{2}a(u, u) - (f, u)$$

A **minimization model** (Ritz model) is to find the solution

$$\min_{u \in H_0^1(I)} F(u).$$

- Under certain assumptions, there three models, (D), (M) and (V) are equivalent mathematically.

Proof – (i) $(D) \rightarrow (V)$

Assume that $u(x)$ is a solution of (D). Then

$$-u''(x) = f(x) \quad \text{and} \quad u(0) = u(1) = 0.$$

For any $v \in H_0^1$,

$$-(u'', v) = (f, v).$$

Using integration by part,

$$-(u'', v) = -\int_0^1 u''v \, dx = -u(x)v(x)|_0^1 + \int_0^1 u'v' \, dx.$$

Since $v(0) = v(1) = 0$, we have

$$(u', v') = (f, v),$$

i.e., $u(x)$ is a solution of (V).

(ii) $(V) \rightarrow (M)$

Assume that $u(x)$ is a solution of (V). Then for any given $v \in H_0^1$, we set $w = v - u$ so that $v = w + u$ and therefore,

$$\begin{aligned} F(v) &= F(w + u) = \frac{1}{2}(u' + w', u' + w') - (f, u + w) \\ &= \frac{1}{2}(u', u') - (f, u) + (u', w') - (f, w) + \frac{1}{2}(w', w'). \end{aligned}$$

Since u is a solution of (V), $(u', w') - (f, w) = 0$ and the last equation becomes

$$F(v) = F(u) + \frac{1}{2}(w', w') \geq F(u)$$

which implies that $u(x)$ is a solution of (M).

(iii) $(M) \rightarrow (D)$ (We follow $(M) \rightarrow (V) \rightarrow (D)$)

Assume that u is a solution of (M) and $u \in C^2(I)$. Then for any $v \in H_0^1$ and real number ϵ , $u + \epsilon v \in H_0^1$ and

$$F(u) \leq F(u + \epsilon v).$$

Let $g(\epsilon) := F(u + \epsilon v)$. Then

$$g(\epsilon) = F(u + \epsilon v) = \frac{1}{2}(u', u') + \epsilon(u', v') + \frac{\epsilon^2}{2}(v', v') - (f, u) - \epsilon(f, v)$$

and $g(0) \leq g(\epsilon)$, *i.e.*, $g(\epsilon)$ has the minimum at $\epsilon = 0$. The necessary condition $g'(0) = 0$ is satisfied and is written by

$$g'(0) = (u', v') - (f, v) = 0 \quad \text{for any } v \in V_0.$$

Thus u is a solution of (V). By integration by part,

$$-(u'', v) - (f, v) = 0$$

i.e.,

$$\int_0^1 (u'' + f)v \, dx = 0 \quad \text{for any } v \in V_0.$$

Hence,

$$-u''(x) = f(x) \quad \text{and} \quad u(0) = u(1) = 0.$$

u is a solution of (D).

• All three models need to find a solution in $H_0^1(I)$. This minimization model is different from those in calculus course where one needs to find a minimal solution in a finite dimensional space. For example,

Example 1.

$$\min_{x \in \mathbb{R}} f(x) := x^2 + 2x$$

Example 2.

$$\min_{X \in \mathbb{R}^N} F(X)$$

FEM method is based on a finite dimensional approximation to the infinite dimensional space $H_0^1(I)$. There are many such approximations. For example, if $u(x)$ is smooth enough, we have its Taylor's series

$$u(x) = u(x_0) + (x - x_0)u'(x_0) + \dots + \frac{(x - x_0)^n}{n!}u^{(n)}(x_0) + \dots$$

Here $u(x)$ is a linear combination of $(x - x_0)^i$, $i = 0, 1, \dots$, or $u \in S = \{1, (x - x_0), \dots, (x - x_0)^n, \dots\}$. A simple approximation to $u(x)$ is its Taylor's polynomial,

$$u_N = u(x_0) + (x - x_0)u'(x_0) + \dots + \frac{(x - x_0)^N}{N!}u^{(N)}(x_0)$$

Then $u_N \in S_N = \{1, (x - x_0), \dots, (x - x_0)^N\}$ where S_N define a finite dimensional space. Based on the idea, we have a class of numerical approximations to the above models. Let S_N be a finite dimensional space. The discrete variational model (V_N) is to find $u \in S_N$ such that

$$a(u, v) = (f, v), \quad v \in \tilde{S}_N,$$

Here we always assume that $\tilde{S}_N = S_N$. This method is called Galerkin method.

The corresponding discrete minimization model (M_N) is to find $u \in S_N$ such that

$$\min_{u \in S_N} F(u).$$

• The approximation solution in a finite dimensional space. Let S_N be an N -dimensional space and $\{\phi_j\}_{j=1}^N$ be the base. For any $u \in S_N$, we have the expression

$$u = \sum_{j=1}^N \alpha_j \phi_j(x).$$

Equivalently, (M_N) model and (V_N) model are to find the coefficients α_j . For the minimization model

$$\min_{u \in S_N} F(u) = \min_{\alpha \in R^N} F\left(\sum_{j=1}^N \alpha_j \phi_j(x)\right)$$

the solution satisfies the equations

$$\frac{\partial F}{\partial \alpha_i} = \sum_{j=1}^N a(\phi_i, \phi_j) \alpha_j \quad i = 1, 2, \dots, N,$$

and in matrix form,

$$A\alpha = F$$

where $A = (a_{ij})_{i,j=1}^N$, $F = (f_j)$ and

$$a_{ij} = a(\phi_i, \phi_j), \quad f_j = (f, \phi_j).$$

For the variational model, taking $v = \phi_i$, $i = 1, 2, \dots, N$, respectively, gives

$$a\left(\sum_{j=1}^N \alpha_j \phi_j, \phi_i\right) = (f, \phi_i), \quad i = 1, 2, \dots, N$$

or the linear system

$$\sum_{j=1}^N \alpha_j a(\phi_j, \phi_i) = (f, \phi_i), \quad i = 1, 2, \dots, N$$

Example.

$$\min_{u \in P_1} \int_0^1 ((u')^2 + 2u^2 - 4u) dx$$

- For more general boundary value problem

$$\begin{aligned} Lu &\equiv -u'' + p(x)u' + q(x)u = f(x), \quad x \in I. \\ u(0) &= u(1) = 0. \end{aligned}$$

the variational model can be obtained by integration by part, which gives

$$(u', v') + (pu', v) + (qu, v) = (f, v) \quad v \in H_0^1(I).$$

The variational model is to find $u \in H_0^1(I)$ such that

$$a(u, v) = (f, v), \quad v \in H_0^1(I)$$

where $a(u, v) = (u', v') + (pu', v) + (qu, v)$.

2.2 Spaces of Piecewise Polynomial Functions The choice of the subspace S_h is a key element in the success of the Galerkin method. It is essential that S_h be chosen so that the Galerkin approximation can be computed efficiently. Secondly, S_h should possess good approximation properties as the accuracy of the Galerkin approximation u_h depends on how well u can be approximated by elements of S_h . The subspace S_h is usually chosen to be a space of piecewise polynomial functions. To define such spaces, let P_r denote the set of polynomials of degree $\leq r$, let

$$\pi : 0 = x_0 < x_1 < \dots < x_N < x_{N+1} = 1$$

denote a partition of \bar{I} , and set

$$I_j = [x_{j-1}, x_j], \quad j = 1, \dots, N+1,$$

$h_j = x_j - x_{j-1}$ and $h = \max_j h_j$. We define

$$\mathcal{M}_k^r(\pi) = \{v \in C^k(\bar{I}) : v|_{I_j} \in P_r, j = 1, \dots, N+1\},$$

where $C^k(\bar{I})$ denotes the space of functions which are k times continuously differentiable on \bar{I} , $0 \leq k < r$, $v|_{I_j}$ denotes the restriction of the function v to the interval I_j , and

$$\mathcal{M}_k^{r,0}(\pi) := \mathcal{M}_k^r(\pi) \cap \{v | v(0) = v(1) = 0\}.$$

It is easy to see that $\mathcal{M}_k^r(\pi)$ and $\mathcal{M}_k^{r,0}(\pi)$ are linear spaces of dimensions $N(r-k) + r + 1$ and $N(r-k) + r - 1$, respectively. These spaces have the following approximation properties.

Theorem 2.1 *For any $u \in W_p^j(I)$, there exists a $\bar{u} \in \mathcal{M}_k^r(\pi)$ and a constant C independent of h and u such that*

$$\|(u - \bar{u})^{(\ell)}\|_{L^p(I)} \leq Ch^{j-\ell} \|u^{(j)}\|_{L^p(I)}, \quad (2.3)$$

for all integers ℓ and j such that $0 \leq \ell \leq k+1$, $\ell \leq j \leq r+1$. If $u \in W_p^j(I) \cap \{v | v(0) = v(1) = 0\}$ then there exists a $\bar{u} \in \mathcal{M}_k^{r,0}(\pi)$ such that the above equation holds.

where L^* denotes the formal adjoint of L ,

$$L^*u = -u'' - (pu)' + qu.$$

It follows from the hypotheses on the smoothness of p and q that $\psi \in H^2(I) \cap H_0^1(I)$ and

$$\|\psi\|_{H^2} \leq C\|Y\|_{L^2}. \quad (2^*)$$

since

$$L^*\psi = -\psi'' - (p\psi)' + q\psi = -\left(\left(\psi e^{\int p dx}\right)' e^{-\int p dx}\right)' + q\psi = -\left((\psi\bar{p})'\bar{p}^{-1}\right)' + q\psi$$

where $\bar{p} = e^{\int p dx}$, and

$$(Y, \bar{p}\psi) = (L^*\psi, \bar{p}\psi) = \left(-((\psi\bar{p})'\bar{p}^{-1})', \psi\bar{p}\right) + (q\psi, \psi\bar{p}).$$

By integration by part,

$$(L^*\psi, \bar{p}\psi) = \left((\psi\bar{p})'\bar{p}^{-1}, (\psi\bar{p})'\right) + (q\psi, \psi\bar{p}) \geq C\|(\psi\bar{p})'\|_{L^2}^2 + C\|\psi\|_{L^2}^2.$$

By Schwartz's inequality,

$$\|\psi\|_{L^2}^2 \leq C(Y, \bar{p}\psi) \leq C\|\psi\bar{p}\|_{L^2}\|Y\|_{L^2} \leq C\|\psi\|_{L^2}\|Y\|_{L^2}$$

and

$$(L^*\psi, \bar{p}\psi) \geq C\|(\psi\bar{p})'\|_{L^2}^2 = C\|\psi'\bar{p} + \psi\bar{p}'\|_{L^2}^2 \geq C\|\psi'\|_{L^2}^2 - C\|\psi\|_{L^2}^2$$

and therefore,

$$\|\psi'\|_{L^2}^2 \leq C(Y, \bar{p}\psi) + C\|\psi\|_{L^2}^2 \leq C\|\psi\|_{L^2}\|Y\|_{L^2} + \|\psi\|_{L^2}^2 \leq C\|Y\|_{L^2}^2$$

Similarly,

$$\|Y\|_{L^2} = \|- \psi'' + (p\psi)' + q\psi\|_{L^2} \geq \|\psi''\|_{L^2} - \|(p\psi)'\|_{L^2} - \|q\psi\|_{L^2} \geq \|\psi''\|_{L^2} - C\|\psi'\|_{L^2} - C\|\psi\|_{L^2}$$

which implies that

$$\|\psi''\|_{L^2} \leq C\|Y\|_{L^2} \quad \blacksquare$$

Thus, using integration by parts and (1*), we have

$$\|Y\|_{L^2}^2 = (L^*\psi, Y) = a(Y, \psi) = a(Y, \psi - \chi),$$

where $\chi \in S_h$, and since

$$a(\phi, \psi) \leq C\|\phi\|_{H^1}\|\psi\|_{H^1}, \quad (3^*)$$

since

$$\begin{aligned} a(\phi, \psi) &= (\phi', \psi') + (p\phi', \psi) + (q\phi, \psi) \\ &\leq \|\phi'\|_{L^2}\|\psi'\|_{L^2} + C\|\phi'\|_{L^2}\|\psi\|_{L^2} + C\|\phi\|_{L^2}\|\psi\|_{L^2} \leq C\|\phi\|_{H^1}\|\psi\|_{H^1} \quad \blacksquare \end{aligned}$$

where we have used the Schwartz's inequality.

For $\phi, \psi \in H^1(I)$, it follows that

$$\|Y\|_{L^2}^2 \leq C\|Y\|_{H^1}\|\psi - \chi\|_{H^1}.$$

From Theorem 2.1, we can choose $\eta \in S_h$ such that

$$\|\psi - \eta\|_{H^1} \leq Ch\|\psi\|_{H^2}.$$

since we can take $j = 2$ and let $l = 1$ and $l = 0$, respectively

$$\begin{aligned} \|\phi' - \eta'\|_{L^2} &\leq Ch\|\psi''\|_{L^2} \leq Ch\|\psi\|_{H^2} \\ \|\phi - \eta\|_{L^2} &\leq Ch^2\|\psi''\|_{L^2} \leq Ch^2\|\psi\|_{H^2} \quad \blacksquare \end{aligned}$$

Thus

$$\|Y\|_{L^2}^2 \leq Ch\|Y\|_{H^1}\|\psi\|_{H^2} \leq Ch\|Y\|_{H^1}\|Y\|_{L^2},$$

where in the last inequality we have used (2*), and we obtain

$$\|Y\|_{L^2} \leq Ch\|Y\|_{H^1}. \quad (4^*)$$

Since $Y \in S_h$,

$$a(Y, Y) = 0,$$

from which it follows that

$$\|Y'\|_{L^2}^2 = -(pY', Y) - (qY, Y) \leq C\{\|Y'\|_{L^2} + \|Y\|_{L^2}\}\|Y\|_{L^2} \leq Ch\|Y\|_{H^1}^2$$

since $a(Y, Y) = (Y', Y') + (pY', Y) + (qY, Y) = 0$. Using this inequality and (4*), we obtain

$$\|Y\|_{H^1}^2 \leq Ch\|Y\|_{H^1}^2.$$

Thus, for h sufficiently small,

$$\|Y\|_{H^1} = 0,$$

and hence from Sobolev's inequality,

$$Y = 0. \quad \blacksquare$$

For the self-adjoint boundary value problem

$$\begin{aligned} Lu &\equiv -(pu')' + q(x)u = f(x), \quad x \in I. \\ u(0) &= u(1) = 0, \end{aligned}$$

the uniqueness of the Galerkin approximation $u_h \in S_h$ satisfying

$$(pu'_h, v') + (qu_h, v) = (f, v), \quad v \in S_h,$$

is much easier to prove. In this case, the coefficient matrix \mathcal{A} of the Galerkin equations is positive definite, from which it follows immediately that the Galerkin approximation is unique. This result is proved in the following theorem.

Theorem 2.2 The matrix \mathcal{A} is positive definite.

Proof — Suppose $\boldsymbol{\beta} \in \mathbf{R}^s$ and $\boldsymbol{\beta} \neq \mathbf{0}$. Then, if $\mathcal{A} = (a_{ij})$ and $w = \sum_{j=1}^s \beta_j w_j$, then

$$\begin{aligned} \boldsymbol{\beta}^T \mathcal{A} \boldsymbol{\beta} &= \sum_{i,j=1}^s a_{ij} \beta_i \beta_j \\ &= \sum_{i,j=1}^s \{(p\beta_i w'_i, \beta_j w'_j) + (q\beta_i w_i, \beta_j w_j)\} \\ &= (pw', w') + (qw, w) \\ &= \|p^{1/2} w'\|_{L^2}^2 + \|q^{1/2} w\|_{L^2}^2 \\ &\geq p_* \|w'\|_{L^2}^2. \end{aligned}$$

The proof is complete if $\|w'\| > 0$. Suppose $\|w'\| = 0$. Then $w' = 0$ and $w = C$, where C is a constant. Since $w \in S_h$, $w(0) = w(1) = 0$, from which it follows that $C = 0$. Therefore $w = 0$; that is $\sum_{j=1}^s \beta_j w_j = 0$. Since $\{w_1, \dots, w_s\}$ is a basis for S_h and is therefore linearly independent, this implies that $\beta_j = 0$, $j = 1, \dots, N$, which is a contradiction. Therefore $\|w'\| > 0$ and \mathcal{A} is positive definite. ■

2.5 The Accuracy of the Galerkin Approximation

- Optimal H^1 and L^2 error estimates.

In the following, an estimate of the error $u - u_h$ in an H^k -norm (or an L^p -norm) will be called *optimal in H^k (or L^p)* if the estimate has the same power of h as is possible by the approximation properties of the subspace S_h with the same smoothness assumptions on u .

Theorem 2.3 *Suppose $u \in H^{r+1}(I) \cap H_0^1(I)$. Then, for h sufficiently small,*

$$\|u - u_h\|_{L^2} + h\|u - u_h\|_{H^1} \leq Ch^{r+1}\|u\|_{H^{r+1}}.$$

Proof — Let $\phi \in H_0^1(I)$ satisfy

$$\begin{aligned} L^*\phi &= e(x), \quad x \in I, \\ \phi(0) &= \phi(1) = 0, \end{aligned}$$

where $e = u - u_h$. Then, as in section 2.4 but with ϕ and e replacing ψ and Y , respectively, we have

$$\|e\|_{L^2} \leq Ch\|e\|_{H^1}, \quad (5^*)$$

since

$$a(e, v) = 0, \quad v \in S_h.$$

Since

$$a(e, e) = a(e, u - \chi), \quad \chi \in S_h,$$

it follows that

$$\begin{aligned} \|e'\|_{L^2}^2 &= -(pe', e) - (qe, e) + a(e, u - \chi) \\ &\leq C\left\{[\|e'\|_{L^2} + \|e\|_{L^2}]\|e\|_{L^2} + \|e\|_{H^1}\|u - \chi\|_{H^1}\right\} \end{aligned}$$

on using Schwarz's inequality and (3*). On using (5*) we have, for h sufficiently small,

$$\|e\|_{H^1} \leq C\|u - \chi\|_{H^1}.$$

Since $u \in H^{r+1}(I) \cap H_0^1(I)$, from Theorem 2.1, χ can be chosen so that

$$\|u - \chi\|_{H^1} \leq Ch^r\|u\|_{H^{r+1}},$$

and hence

$$\|e\|_{H^1} \leq Ch^r\|u\|_{H^{r+1}}.$$

The use of this estimate in (5*) completes the proof. ■

- Optimal L^∞ error estimate.

In this section, we derive an optimal L^∞ error estimate when $S_h = \mathcal{M}_k^{r,0}(\pi)$ and the partition π of I is from a *quasi-uniform collection of partitions*, which we now define.

Definition Let

$$\pi : 0 = x_0 < x_1 < \dots < x_N < x_{N+1} = 1$$

denote a partition of I and let $\Pi(I)$ denote the collection of all such partitions π of I . As before, set

$$h_i = x_i - x_{i-1}, \quad i = 1, \dots, N+1,$$

and $h = \max_i h_i$. A collection of partitions $\mathcal{C} \subset \Pi(I)$ is called *quasi-uniform* if there exists a constant $\sigma \geq 1$ such that, for all $\pi \in \mathcal{C}$,

$$\max_{1 \leq j \leq N+1} hh_j^{-1} \leq \sigma.$$

The following lemma, which is proved in [?], plays an important role in the derivation of the L^∞ estimate.

Lemma 2.1 *Let Pu be the L^2 projection of u into $\mathcal{M}_k^r(\pi)$, that is,*

$$(Pu - u, v) = 0 \quad v \in \mathcal{M}_k^r(\pi),$$

where $0 \leq k \leq r-1$. Then, if $u \in W_\infty^{r+1}(I)$,

$$\|Pu - u\|_{L^\infty(I)} \leq Ch^{r+1} \|u\|_{W_\infty^{r+1}(I)},$$

provided π is chosen from a quasi-uniform collection of partitions of I .

We now prove the following theorem.

Theorem 2.4 *Suppose $S_h = \mathcal{M}_k^{r,0}(\pi)$, with π chosen from a quasi-uniform collection of partitions. If $u \in W_\infty^{r+1}(I) \cap H_0^1(I)$, then*

$$\|u - u_h\|_{L^\infty} \leq Ch^{r+1} \|u\|_{W_\infty^{r+1}}.$$

Proof — Let $W \in S_h$ satisfy

$$(W', v') = (u', v'), \quad v \in S_h.$$

Since

$$a(u - u_h, v) = 0, \quad v \in S_h,$$

it follows that

$$((W - u_h)', v') + (u - u_h, qv - (pv)') = 0,$$

for all $v \in S_h$. If we set $v = W - u_h$, then

$$\|(W - u_h)'\|_{L^2}^2 \leq C \|u - u_h\|_{L^2} \|W - u_h\|_{H^1}.$$

Therefore, since $\|\cdot\|_{H_0^1}$ and $\|\cdot\|_{H^1}$ are equivalent norms on H_0^1 ,

$$\|W - u_h\|_{H^1} \leq C \|u - u_h\|_{L^2}$$

and from Sobolev's inequality, we obtain

$$\|W - u_h\|_{L^\infty} \leq C \|u - u_h\|_{L^2}.$$

Then, from Theorem 2.3, it follows that

$$\|W - u_h\|_{L^\infty} \leq Ch^{r+1}\|u\|_{H^{r+1}}.$$

Since

$$L^\infty,$$

we need to estimate $\|u - W\|_{L^\infty}$ to complete the proof.

Note that since

$$((u - W)', 1) = 0,$$

it follows that W' is the L^2 projection of u' into $\mathcal{M}_{k-1}^{r-1}(\pi)$, and hence, from Lemma 2.1, we obtain

$$\|(u - W)'\|_{L^\infty} \leq Ch^r \|u'\|_{W_\infty^r} \leq Ch^r \|u\|_{W_\infty^{r+1}}. \quad (6^*)$$

Now suppose $g \in L^1$ and define G by

$$\begin{aligned} G'' &= -g(x), \quad x \in I, \\ G(0) &= G(1) = 0. \end{aligned}$$

Then

$$\|G\|_{W_1^2} \leq C\|g\|_{L^1}, \quad (7^*)$$

and, for $\chi \in S_h$,

$$(u - W, g) = -(u - W, G'') = ((u - W)', (G - \chi)').$$

On using Hölder's inequality, we obtain

$$(u - W, g) \leq \|(u - W)'\|_{L^\infty} \|(G - \chi)'\|_{L^1}.$$

From Theorem (2.1), we can choose χ so that

$$\|(G - \chi)'\|_{L^1} \leq Ch\|G\|_{W_1^2}.$$

Hence, on using (7*), it follows that

$$|(u - W, g)| \leq Ch\|(u - W)'\|_{L^\infty}\|g\|_{L^1}.$$

On using (6*) and duality, we have

$$\|u - W\|_{L^\infty} \leq Ch^{r+1}\|u\|_{W_\infty^{r+1}}.$$

The desired result now follows. \blacksquare

2.6 Superconvergence results The error estimates of Theorems ?? and ?? are optimal and consequently no better global rates of convergence are possible. However, there can be identifiable points at which the approximate solution converges at rates that exceed the optimal global rate. In the following theorem, we derive one such *superconvergence result*.

Theorem 3.5 *If $S_h = \mathcal{M}_0^{r,0}(\pi)$ and $u \in H^{r+1}(I) \cap H_0^1(I)$, then, for h sufficiently small,*

$$|(u - u_h)(x_i)| \leq Ch^{2r} \|u\|_{H^{r+1}}, \quad i = 0, \dots, N + 1.$$

Proof — Let $G(x, \xi)$ denote the Green's function for (??); that is,

$$u(x) = -(Lu, G(x, \cdot)) = a(u, G(x, \cdot))$$

for sufficiently smooth u . This representation is valid for $u \in H_0^1(I)$ and hence it can be applied to $e = u - u_h$. Thus, for $\chi \in S_h$,

$$e(x_i) = a(e, G(x_i, \cdot)) = a(e, G(x_i, \cdot) - \chi),$$

since

$$a(e, \chi) = 0, \quad \chi \in S_h.$$

Thus,

$$|e(x_i)| \leq C \|e\|_{H^1} \|G(x_i, \cdot) - \chi\|_{H^1}.$$

From the smoothness assumptions on p and q , it follows that

$$G(x_i, \cdot) \in H^{r+1}([0, x_i]) \cap H^{r+1}([x_i, 1]),$$

and

$$\|G(x_i, \cdot)\|_{H^{r+1}([0, x_i])} + \|G(x_i, \cdot)\|_{H^{r+1}([x_i, 1])} \leq C.$$

Hence there exists $\chi \in S_h$ (obtained, for example, by Lagrange interpolation on each subinterval) such that

$$\|G(x_i, \cdot) - \chi\|_{H^1} \leq Ch^r.$$

From Theorem ??, we have

$$\|e\|_{H^1} \leq Ch^r \|u\|_{H^{r+1}},$$

for h sufficiently small, and hence combining (??)–(??), we obtain

$$|e(x_i)| \leq Ch^{2r} \|u\|_{H^{r+1}},$$

as desired. ■

A method which involves very simple auxiliary computations using the Galerkin solution can be used to produce superconvergent approximations to the derivative, [?]. First we consider approximations to $u'(0)$ and $u'(1)$. Motivated by the fact that

$$(f, (1 - x)) = (-u'' + pu' + qu, 1 - x) = u'(0) + a(u, 1 - x),$$

we define an approximation Γ_0 to $u'(0)$ by

$$\Gamma_0 = (f, 1 - x) - a(u_h, 1 - x),$$

where u_h is the solution to (??). Also, with $1 - x$ replaced by x in the above, we find that

$$u'(1) = a(u, x) - (f, x),$$

and hence we define an approximation Γ_{N+1} to $u'(1)$ by

$$\Gamma_{N+1} = a(u_h, x) - (f, x).$$

It can be shown that that if $u \in H^{r+1}(I)$, then

$$|\Gamma_j - u'(x_j)| \leq Ch^{2r} \|u\|_{H^{r+1}},$$

$j = 0, N + 1$, when for example, $S_h = \mathcal{M}_k^{r,0}(\pi)$.

If $S_h = \mathcal{M}_0^{r,0}(\pi)$, a procedure can be defined which at the nodes x_i , $i = 1, \dots, N$, gives superconvergence results similar to (??). Specifically, if Γ_j , an approximation to $u'(x_j)$, $j = 1, \dots, N$, is defined by

$$\Gamma_j = \frac{a(u_h, x)_{I'_j} - (f, x)_{I'_j}}{x_j},$$

where the subscript I'_j denotes that the inner products are taken over $I'_j = (0, x_j)$, then (??) holds for $j = 1, \dots, N$. The approximation Γ_j is motivated by the fact that

$$(Lu, x)_{I'_j} = (f, x)_{I'_j},$$

and, after integration by parts,

$$-u'(x_j)x_j + a(u, x)_{I'_j} = (f, x)_{I'_j}.$$

2.7 Remarks

- non-uniform mesh
- Quadrature Galerkin methods.
- Linear solver.
- Nonlinear Problems. Consider the boundary value problem

$$\begin{aligned} -u'' + f(x, u) &= 0, & x \in I, \\ u(0) &= u(1) = 0. \end{aligned}$$

where we assume $f_u > 0$. It is easy to show that the weak form of this boundary value problem is

$$(u', v') + (f(u), v) = 0, \quad v \in H_0^1(I).$$

As before, let S_h be a finite dimensional subspace of $H_0^1(I)$ with basis $\{w_1, \dots, w_s\}$. The Galerkin approximation to u is the element $u_h \in S_h$ such that

$$(u'_h, v') + (f(u_h), v) = 0, \quad v \in S_h, \quad (5^*)$$

and if

$$u_h(x) = \sum_{j=1}^s \alpha_j w_j(x),$$

we obtain with $v = w_i$, the nonlinear system

$$\sum_{j=1}^s (w'_i, w'_j) \alpha_j + \left(f\left(\sum_{\nu=1}^s \alpha_\nu w_\nu\right), w_i \right) = 0, \quad i = 1, \dots, s,$$

for $\alpha_1, \dots, \alpha_s$. Newton's method for the solution above can be easily derived by linearizing (5*) to obtain

$$(u_h^{(n)'} , v') + (f(u_h^{(n-1)}), v) + (f_u(u_h^{(n-1)})(u_h^{(n)} - u_h^{(n-1)}), v) = 0, \quad v \in S_h,$$

where $u_h^{(0)}$ is arbitrary. If

$$u_h^{(k)} = \sum_{j=1}^s \alpha_j^{(k)} w_j,$$

then

$$(A + B_n)\boldsymbol{\alpha}^{(n)} = -\mathbf{F}_n + B_n\boldsymbol{\alpha}^{(n-1)}, \quad (2.5)$$

where

$$\begin{aligned} A &= ((w'_i, w'_j)), \quad \boldsymbol{\alpha}^{(n)} = (\alpha_1^{(n)}, \dots, \alpha_N^{(n)})^T, \\ B_n &= \left((f_u\left(\sum_{\nu=1}^s \alpha_\nu^{(n-1)} w_\nu\right) w_i, w_j) \right), \\ \mathbf{F}_n &= ((f(u_h^{(n-1)}), w_i)). \end{aligned}$$

Note that (2.4) may be written in the form

$$(A + B_n)(\boldsymbol{\alpha}^{(n)} - \boldsymbol{\alpha}^{(n-1)}) = -(A\boldsymbol{\alpha}^{(n-1)} + \mathbf{F}_n).$$

A comprehensive account of the Galerkin method for second order nonlinear problems is given by Fairweather (1978).

A different way is to use Newton method (linearization) for the boundary value problem before to change it to a weak form. The linearized equation is given by

$$\begin{aligned} -u^{(n+1)''} + f(x, u^{(n)}) + f_u(x, u^{(n)})(u^{(n+1)} - u^{(n)}) &= 0 \\ u^{(n+1)}(0) = u^{(n+1)}(1) &= 0. \end{aligned}$$

3 Collocation Methods

3.1 Introduction. Consider the linear second order two-point boundary value problem

$$\begin{aligned} Lu &\equiv -u'' + p(x)u' + q(x)u = f(x), \quad x \in I, \\ u(0) &= u(1) = 0, \end{aligned} \tag{3.6}$$

where the functions p , q and f are smooth on I . Let

$$\pi : 0 = x_0 < x_1 < \dots < x_N < x_{N+1} = 1$$

denote a partition of \bar{I} , and set $h_i = x_i - x_{i-1}$.

Let V_N be a finite dimensional space approximating to $H_0^1(I)$ and we hope to find an approximation in V_N . Let $u_h = \sum \alpha_j \phi_j(x)$ where $\phi_j(x)$, $j = 1, 2, \dots, N$, are the basis functions of V_N . The collocation solution $u_h(x)$ (or α_j) satisfies the system

$$Lu_h(x_i^c) = f(x_i^c), \quad i = 1, 2, \dots, N,$$

where $x_i^c \in I$ denote the collocation points. Moreover, we have

$$\sum_{j=1}^N L\phi_j(x_i^c)\alpha_j = f(x_i^c), \quad i = 1, 2, \dots, N,$$

or

$$\sum_{j=1}^N a_{ij}\alpha_j = f(x_i^c), \quad i = 1, 2, \dots, N,$$

where $a_{ij} = L\phi_j(x_i^c)$.

Variational form: find a solution $u_h \in V_N$ such that

$$(Lu_h, v) = (f, v) \quad v \in U_N$$

If we choose

$$v = \delta(x - x_i^c) \quad i = 1, 2, \dots, N$$

we have

$$Lu_h(x_i^c) = f(x_i^c) \quad i = 1, 2, \dots,$$

which is the collocation system.

Collocation methods: simple, clear and no numerical integration, but linear system is symmetric.

3.2 spectral collocation methods. Collocation methods depend upon: (i) V_N and (ii) collocation points x_i^c , $i = 1, 2, \dots, N$.

- Spectral collocation uses: Global polynomial of degree $N + 1$ with the conditions $u(0) = u(1) = 0$, and zeros or extremal points of some special polynomials, Legendre, Chebyshev, or Radau-type and Lobatto-type.

Example 3.1. For the case $p = 0$, $q = 1$ and $f(x) = \sin \pi x$, find the spectral collocation solution in P_3 . Let

$$u_h = x(1-x)(\alpha_0 + \alpha_1 x)$$

and choose the collocation points to be the zeros of Gauss-Legendre polynomial of degree 2,

$$x_1^c = (1 - \sqrt{3}/3)/2, \quad x_2^c = (1 + \sqrt{3}/3)/2$$

. Then

$$-u_h''(x_i^c) + u_h(x_i^c) = \sin(\pi x_i^c), \quad i = 1, 2 \quad (3.7)$$

- Linear system: condition number $= O(N^4)$, ill-conditioning and nonsymmetric.
- exponential accuracy: $O(N^r)$.

3.3 Orthogonal Hermite cubic spline collocation methods.

In the orthogonal Hermite cubic spline collocation method for (3.6), the approximate solution $u_h \in \mathcal{M}_1^3(\pi)$. If $\{\phi_j\}_{j=1}^s$ is a basis for $\mathcal{M}_1^3(\pi)$, where $s = 2(N+1) + 2$, we may write

$$u_h(x) = \sum_{j=1}^s u_j \phi_j(x). \quad (3.8)$$

Then the coefficients $\{u_j\}_{j=1}^s$ are determined by requiring that u_h satisfy (3.6) at the points $\{x_j^c\}_{j=1}^{s-2}$, and the boundary conditions in (3.1):

$$\begin{aligned} Lu_h(x_j^c) &= f(\xi_j), \quad j = 1, 2, \dots, s-2, \\ u_h(0) &= g_0, u_h(1) = g_1, \end{aligned}$$

where

$$x_{(i-1)(r-1)+k}^c = x_{i-1} + h\sigma_k, \quad i = 1, 2, \dots, N+1, \quad k = 1, 2,$$

and $\{\sigma_k\}_{k=1}^2$ are the nodes for the 2-point Gauss-Legendre quadrature rule on the interval $[0, 1]$, *i.e.*,

$$x_{2i-1}^c = x_{i-1} + \frac{1}{2} \left(1 - \frac{1}{\sqrt{3}}\right) h, \quad x_{2i}^c = x_{i-1} + \frac{1}{2} \left(1 + \frac{1}{\sqrt{3}}\right) h, \quad i = 1, 2, \dots, N+1.$$

In a standard element $[0, 1]$,

$$\begin{aligned} u_h &= u_1(1+2t)(1-t)^2 + u_1' t(1-t)^2 + u_2(3-2t)t^2 - u_2' t^2(1-t) \\ &= u_1 \eta_0(t) + u_1' \eta_1(t) + u_2 \eta(1-t) - u_2' \eta_1(1-t). \end{aligned}$$

Then in the element $[x_j, x_{j+1}]$,

$$u_h(x) = u_h(x_j + th) = u_j \eta_0(t) + u_j' h \eta_1(t) + u_{j+1} \eta_0(1-t) - u_2' h \eta_1(1-t)$$

Example 3.2 We consider the same problem as in example 3.1. We need to solve the boundary value problem

$$\begin{aligned} -u''(x) + u &= \sin \pi x \\ u(0) &= u(1) = 0 \end{aligned} \quad (3.9)$$

by Hermite cubic spline collocation method with two elements and two collocation points at each element. Let $u_h = \sum_{j=1}^4 u_j \phi_j(x)$. We have a system of four equations with two boundary conditions.

$$\sum_{j=1}^4 v_j (-\phi_j''(x_i^c) + \phi_j(x_i^c)) = f(x_i^c), \quad i = 1, 2, 3, 4$$

We have a system of four equations with two boundary conditions. In the element $[x_0, x_1]$, The linear system is given by

$$\begin{bmatrix} \times & \times & \times & \\ \times & \times & \times & \\ & \times & \times & \times \\ & \times & \times & \times \end{bmatrix} \begin{bmatrix} u'_0 \\ u_1 \\ u'_1 \\ u'_2 \end{bmatrix} = \begin{bmatrix} f(x_1^c) \\ f(x_2^c) \\ f(x_3^c) \\ f(x_4^c) \end{bmatrix}.$$

- Linear solver.
- Convergence.