Solving Nonlinear Algebraic Equations I

Discretization using Polynomial Interpolation

Consider a function $u(z)$ to be a continuous function defined over $z \in [a,b]$ and let $\{u_1, u_2, \ldots u_{n+1}\}$ represent the values of the function at an arbitrary set of points $\{z_1, z_2, \ldots, z_{n+1}\}$ in the domain $[a,b]$. Another function, say $\tilde{u}(z)$ in $C[a,b]$ that assumes values $\{u_1, u_2, \ldots u_{n+1}\}$ exactly at $\{z_1, z_2, \ldots, z_{n+1}\}$ is called an interpolation function. Most popular form of interpolating functions are polynomials. Polynomial interpolation has many important applications. It is one of the primary tool used in the approximation of the infinite dimensional operators and generating computationally tractable approximate forms. In this section, we examine applications of polynomial interpolation to discretization. In the development that follows, for the sake of notational convenience, it is assumed that

$$z_1 = a < z_2 < z_3 < \ldots < z_{n+1} = b \tag{84}$$

1 Lagrange Interpolation

In Lagrange interpolation, it is desired to find an interpolating polynomial $p(z)$ of the form

$$p(z) = \alpha_0 + \alpha_1 z + \ldots + \alpha_n z^n \tag{85}$$

such that

$$p(z_i) = u_i \quad \text{for } i = 1, 2, \ldots n+1$$

To find coefficients of the polynomial that passes exactly through $\{u_i : i = 1, 2, \ldots, n+1\}$, consider (n+1) equations

$$\begin{aligned}
\alpha_0 + \alpha_1 z_1 + \ldots + \alpha_n z_1^n &= u_1 \\
\alpha_0 + \alpha_1 z_2 + \ldots + \alpha_n z_2^n &= u_2 \\
\ldots &= \ldots \\
\alpha_0 + \alpha_1 z_{n+1} + \ldots + \alpha_n z_{n+1}^n &= u_{n+1}
\end{aligned}$$

which can be rearranged as follows

$$\mathbf{A}\theta = \mathbf{u} \tag{86}$$

where

$$\theta = \begin{bmatrix} \alpha_0 & \alpha_1 & \ldots & \alpha_n \end{bmatrix}^T \tag{87}$$

$$\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \ldots & u_{n+1} \end{bmatrix}^T \tag{88}$$

$$\mathbf{A} = \begin{bmatrix} 1 & z_1 & \ldots & (z_1)^n \\ 1 & z_2 & \ldots & (z_2)^n \\ \ldots & \ldots & \ldots & \ldots \\ 1 & z_{n+1} & \ldots & (z_{n+1})^n \end{bmatrix}$$

Since matrix $\mathbf{A}$ and vector u are known, the coefficients of the Lagrange interpolation polynomial can be found by solving for vector θ

2 Piecewise Polynomial Interpolation [2]

Matrix $\mathbf{A}$ appearing in equation (86) is known as Vandermond matrix. Larger dimensional Vandermond matrices tend to become numerically ill-conditioned (Refer to Section 7 in module on Solving Linear Algebraic Equations). Also, if the number of data points is large, fitting a large order polynomial can result in a polynomial which exhibits unexpected oscillatory behavior. In order to avoid such oscillations and the difficulties arising from ill conditioning of the Vandermond matrix, the data is divided into sub-intervals and a lower order spline approximation is developed on each sub-interval. Let [a,b] be a finite interval. We introduce a partition of the interval by placing points

$$a \leq \mathcal{Z}_1 < \mathcal{Z}_2 < \mathcal{Z}_3 \ldots < \mathcal{Z}_{n+1} \leq b$$

where $\mathcal{Z}_i$ are called *nodes*. A function is said to be a piecewise polynomial of degree k on this partition if in each subinterval $\mathcal{Z}_i \leq z \leq \mathcal{Z}_{i+1}$ we develop a k'th degree polynomial. For example, a piecewise polynomial of degree one consists of straight line segments. Such an approximation is continuous at the nodes but will have discontinuous derivatives. In some applications it is important to have a smooth approximation with continuous derivatives. A piecewise k'th degree polynomial, which has continuous derivatives up to order $k-1$ is called a spline of degree k. In particular, the case $k = 3,$ i.e. cubic spline, has been studied extensively in the literature. In this section, we restrict our discussion to the development of cubic splines. Thus, given a set of points $z_1 = a < z_2 < z_3 < \ldots < z_{n+1} = b,$ the nodes are chosen as

$$\mathcal{Z}_i = z_i \quad \text{for} \quad i = 1, 2, \ldots, n+1$$

and n cubic splines that fit $(n+1)$ data points can be expressed as

---
---
(89)

$$p_1(z) = \alpha_{0,1} + \alpha_{1,1}(z - z_1) + \alpha_{2,1}(z - z_1)^2 + \alpha_{3,1}(z - z_1)^3$$

---
(90)

$$(z_1 \leq z \leq z_2)$$

---
(91)

$$p_2(z) = \alpha_{0,2} + \alpha_{1,2}(z - z_2) + \alpha_{2,2}(z - z_2)^2 + \alpha_{3,2}(z - z_2)^3$$

---
(92)

$$(z_2 \leq z \leq z_3)$$

$$\ldots\ldots = \ldots\ldots\ldots\ldots\ldots\ldots$$

$$p_n(z) = \alpha_{0,n} + \alpha_{1,n}(z - z_n) + \alpha_{2,n}(z - z_n)^2 + \alpha_{3,n}(z - z_n)^3$$

$$(z_n \leq z \leq z_{n+1})$$

---
(93)

There are total $4n$ unknown coefficients $\{\alpha_{0,1}, \alpha_{1,1} \ldots\ldots \alpha_{3,n}\}$ to be determined. In order to ensure continuity and smoothness of the approximation, the following conditions are imposed

- Initial point of each polynomial

---
(94)

$$p_i(z_i) = u_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

- Terminal point of the last polynomial

---
(95)

$$p_n(z_{n+1}) = u_{n+1}$$

- Conditions for ensuring continuity between two neighboring polynomials

---
---
(96)

$$p_i(z_{i+1}) = p_{i+1}(z_{i+1}) \quad ; \; i = 1, 2, \ldots . n-1$$

$$\frac{dp_i(z_{i+1})}{dz} = \frac{dp_{i+1}(z_{i+1})}{dz} \quad ; \; i = 1, 2, \ldots . n-1 \qquad (97)$$

$$\frac{d^2p_i(z_{i+1})}{dz^2} = \frac{d^2p_{i+1}(z_{i+1})}{dz^2} \quad ; \; i = 1, 2, \ldots . n-1 \qquad (98)$$

which result in $4n-2$ conditions including earlier conditions.

- Two additional conditions are imposed at the boundary points

$$\frac{d^2p_1(z_1)}{dz^2} = \frac{d^2p_n(z_{n+1})}{dz^2} = 0 \qquad (99)$$

which are referred to as *free* boundary conditions. If the first derivatives at the boundary points are known,

$$\frac{dp_1(z_1)}{dz} = d_1 \; ; \; \frac{dp_n(z_{n+1})}{dz} = d_{n+1} \qquad (100)$$

then we get the *clamped* boundary conditions.

Using constraints (94-98) and defining $\Delta z_i = z_{i+1} - z_i$, we get the following set of coupled linear algebraic equations

$$\alpha_{0,i} = u_i \quad ; \quad (i = 1, 2, \ldots, n) \qquad (101)$$

$$\alpha_{0,n} + \alpha_{1,n}(\Delta z_n) + \alpha_{2,n}(\Delta z_n)^2 + \alpha_{3,n}(\Delta z_n)^3 = u_{n+1} \qquad (102)$$

$$\alpha_{0,i} + \alpha_{1,i}(\Delta z_i) + \alpha_{2,i}(\Delta z_i)^2 + \alpha_{3,i}(\Delta z_i)^3 = \alpha_{0,i+1} \qquad (103)$$

$$\alpha_{1,i} + 2\alpha_{2,i}(\Delta z_i) + 3\alpha_{3,i}(\Delta z_i)^2 = \alpha_{1,i+1} \qquad (104)$$

$$\alpha_{2,i} + 3\alpha_{3,i}(\Delta z_i) = \alpha_{2,i+1} \qquad (105)$$

$$for \quad i = 1, 2, \ldots, n-1$$

In addition, using the free boundary conditions, we have

$$\alpha_{2,1} = 0 \qquad (106)$$

$$\alpha_{2,n} + 3\alpha_{3,n}(\Delta z_n) = 0 \qquad (107)$$

Eliminating $\alpha_{3,i}$ using equation (105 and 107), we have

$$\alpha_{3,i} = \frac{\alpha_{2,i+1} - \alpha_{2,i}}{3(\Delta z_i)} \quad \text{for} \quad i = 1, 2, \ldots, n-1 \qquad (108)$$

$$\alpha_{3,n} = \frac{-\alpha_{2,n}}{3(\Delta z_n)} \qquad (109)$$

and eliminating $\alpha_{1,i}$ using equations (102,103), we have

$$\alpha_{1,i} = \frac{1}{\Delta z_i}(\alpha_{0,i+1} - \alpha_{0,i}) - \frac{\Delta z_i}{3}(2\alpha_{2,i} + \alpha_{2,i+1}) \quad (110)$$

$$\text{for } i = 1, 2, \ldots, n-1$$

$$\alpha_{1,n} = \frac{u_{n+1} - \alpha_{0,n}}{\Delta z_n} - (\Delta z_n)\alpha_{2,n} - \alpha_{3,n}(\Delta z_n)^2 \quad (111)$$

Thus, we get only $\{\alpha_{2,i} : i = 1, \ldots, n\}$ as unknowns and the resulting set of linear equations assume the form

$$\alpha_{2,1} = 0 \quad (112)$$

$$(\Delta z_{i-1})\alpha_{2,i-1} + 2(\Delta z_i + \Delta z_{i-1})\alpha_{2,i} + (\Delta z_i)\alpha_{2,i+1} = b_i \quad (113)$$

$$\text{for } i = 2, \ldots, n-1$$

where

$$b_i = \frac{3(\alpha_{0,i+1} - \alpha_{0,i})}{\Delta z_i} - \frac{3(\alpha_{0,i} - \alpha_{0,i-1})}{\Delta z_{i-1}}$$
$$= \frac{3(u_{i+1} - u_i)}{\Delta z_i} - \frac{3(u_i - u_{i-1})}{\Delta z_{i-1}}$$

for $i = 2, \ldots, n-1$.

$$\frac{1}{3}(\Delta z_{n-1})\alpha_{2,n-1} + \frac{2}{3}(\Delta z_{n-1} + \Delta z_n)\alpha_{2,n} = b_n \quad (114)$$

$$b_n = \frac{u_{n+1}}{\Delta z_n} - \left(\frac{1}{\Delta z_n} + \frac{1}{\Delta z_{n-1}}\right)u_n + \frac{u_{n-1}}{\Delta z_{n-1}}$$

Defining vector $\boldsymbol{\alpha}_2$ as

$$\boldsymbol{\alpha}_2 = \begin{bmatrix} \alpha_{2,1} & \alpha_{2,2} & \ldots\ldots & \alpha_{2,n} \end{bmatrix}^T$$

the above set of n equations can be rearranged as

$$\mathbf{A}\boldsymbol{\alpha}_2 = \mathbf{b} \quad (115)$$

where $\mathbf{A}$ is a $(n \times n)$ matrix and $\mathbf{b}$ is $(n \times 1)$ vector. Elements of $\mathbf{A}$ and $\mathbf{b}$ can be obtained from equations (112-114). Note that matrix $\mathbf{A}$ will be a near tridiagonal matrix, i.e. a sparse matrix.Once all the $\alpha_{2,i}$ are obtained, $\alpha_{1,i}$ and $\alpha_{2,i}$ can be easily obtained.

3 Interpolation using Linearly Independent Functions

While polynomial is a popular choice as basis for interpolation, any set of linearly independent functions defined on [a,b] can be used for developing an interpolating function. Let $\{f_0(z), f_1(z), \ldots f_n(z)\}$ represent a set of linearly independent functions in $C[a,b]$. Then, we can construct an interpolating function, $g(z)$, as follows

$$g(z) = \alpha_0 f_1(z) + \ldots\ldots + \alpha_n f_n(z) \quad (116)$$

Forcing the interpolating function to have values u_i at $z = z_i$ leads to the following set of linear algebraic equations

$$\alpha_0 f_0(z_i) + \ldots\ldots + \alpha_n f_n(z_i) = u_i \qquad (117)$$

$$i = 0, 1, \ldots, n$$

which can be further rearranged as $\mathbf{A}\theta = \mathbf{u}$ where [with $z_0 = 0$ and $z_n = 1$]

$$\mathbf{A} = \begin{bmatrix} f_0(0) & f_1(0) & \ldots. & f_n(0) \\ f_0(z_1) & f_1(z_1) & \ldots. & f_n(z_1) \\ \ldots. & \ldots. & \ldots. & \ldots. \\ f_0(1) & f_1(1) & \ldots. & f_n(1) \end{bmatrix} \qquad (118)$$

and vectors θ and $\mathbf{u}$ are defined by equations (87) and (88), respectively. Commonly used interpolating functions are

- Shifted Legandre polynomials
- Chebysheve polynomials
- Trigonometric functions, i.e. sines and cosines
- Exponential functions $\{e^{\alpha_i z} : i = 0, 1, \ldots n\}$ with $\alpha_0 \ldots . \alpha_n$ specified i.e.

$$g(z) = \theta_1 e^{\alpha_1 z} + \theta_2 e^{\alpha_2 z} + \ldots\ldots\ldots\ldots\ldots\ldots + \theta_n e^{\alpha_n z} \qquad (119)$$

4 Discretization using Orthogonal Collocations [2]

One of the important applications of polynomial interpolation is the method of orthogonal collocations. By this approach, the differential operator over a spatial / temporal domain is approximated using an interpolation polynomial.

4.1 Discretization of ODE-BVP

Consider the second order ODE-BVP given by equations (32), (33) and (34a). To see how the problem discretization can be carried out using Lagrange interpolation, consider a selected set of collocation (grid) points $\{z_i : i = 1, \ldots, n+1\}$ in the domain $[0, 1]$ such that $z_1 = 0$ and $z_{n+1} = 1$ and $\{z_2, z_3, \ldots . z_n\} \in (0, 1)$ such that

$$z_1 = 0 < z_2 < z_3 < \ldots . < z_{n+1} = 1$$

Let $\{u_i = u(z_i) : i = 1, 2, \ldots, n+1\}$ represent the values of the dependent variable at these collocation points. Given these points, we can propose an approximate solution, $u(z)$, of the form

$$u(z) = \alpha_0 + \alpha_1 z + \ldots . . + \alpha_n z^n$$

to the ODE-BVP as an interpolation polynomial that passes exactly through $\{u_i : i = 1, \ldots n+1\}$. This requires that the following set of equations hold

$$u(z_i) = \alpha_0 + \alpha_1 z_i + \ldots . . + \alpha_n z_i^n = u_i$$
$$i = 1, 2, \ldots, n+1$$

at the collocation points. The unknown polynomial coefficients $\{\alpha_i : i = 0, 1, \ldots n\}$ can be expressed in terms of unknowns $\{u_i : i = 1, \ldots, n+1\}$ as follows

$$\theta = \mathbf{A}^{-1}\mathbf{u}$$

where matrix $\mathbf{A}$ is defined in equation (86). To approximate the OBE-BVP in $(0,1)$, we force the residuals at the collocation points to zero using the approximate solution $u(z)$, i.e.

$$R_i = \Psi\left[\frac{d^2u(z_i)}{dz^2}, \frac{du(z_i)}{dz}, u(z_i), z_i\right] = 0 \quad (120)$$

for $i = 2,3,\ldots n$. Thus, we need to compute the first and second derivatives of the approximate solution $\widetilde{u}(z)$ at the collocation points. The first derivative at i'th collocation point can be computed as follows

$$\frac{d\widetilde{u}(z_i)}{dz} = 0\alpha_0 + \alpha_1 + 2\alpha_2 z_i + \ldots\ldots + n\alpha_n z_i^{n-1} \quad (121)$$

$$= \begin{bmatrix} 0 & 1 & 2z_i & \ldots & nz_i^{n-1} \end{bmatrix}\theta \quad (122)$$

$$= \begin{bmatrix} 0 & 1 & 2z_i & \ldots & nz_i^{n-1} \end{bmatrix}\mathbf{A}^{-1}\mathbf{u} \quad (123)$$

Defining vector

$$\left[\mathbf{s}^{(i)}\right]^T = \begin{bmatrix} 0 & 1 & 2z_i & \ldots & nz_i^{n-1} \end{bmatrix}\mathbf{A}^{-1}$$

we have

$$\frac{d\widetilde{u}(z_i)}{dz} = \left[\mathbf{s}^{(i)}\right]^T\mathbf{u}$$

Similarly, the second derivative can be expressed in terms of vector $\mathbf{u}$ as follows:

$$\frac{d^2\widetilde{u}(z_i)}{dz^2} = 0\alpha_0 + 0\alpha_1 + 2\alpha_2 + \ldots\ldots + n(n-1)\alpha_n z_i^{n-2} \quad (124)$$

$$= \begin{bmatrix} 0 & 0 & 2 & \ldots & n(n-1)z_i^{n-2} \end{bmatrix}\theta \quad (125)$$

$$= \begin{bmatrix} 0 & 0 & 2 & \ldots & n(n-1)z_i^{n-2} \end{bmatrix}\mathbf{A}^{-1}\mathbf{u} \quad (126)$$

Defining vector

$$\left[\mathbf{t}^{(i)}\right]^T = \begin{bmatrix} 0 & 0 & 2 & \ldots & n(n-1)z_i^{n-2} \end{bmatrix}\mathbf{A}^{-1}$$

we have

$$\frac{d^2\widetilde{u}(z_i)}{dz^2} = \left[\mathbf{t}^{(i)}\right]^T\mathbf{u}$$

Substituting for the first and the second derivatives of $\widetilde{u}(z_i)$ in equations in (120), we have

$$\Psi\left[\left[\mathbf{t}^{(i)}\right]^T\mathbf{u}, \left[\mathbf{s}^{(i)}\right]^T\mathbf{u}, \mathbf{u}_i, z_i\right] = 0 \quad (127)$$

for $i = 2,3\ldots,n$. At the boundary points, we have two additional constraints

$$f_1\left[\frac{d\tilde{u}(0)}{dz}, u_1, 0\right] = f_1\left[\left[\mathbf{s}^{(1)}\right]^T \mathbf{u}, u_1, 0\right] = 0$$

$$f_2\left[\frac{d\tilde{u}(1)}{dz}, u_{n+1}, 1\right] = f_2\left[\left[\mathbf{s}^{(n+1)}\right]^T \mathbf{u}, u_{n+1}, 1\right] = 0 \qquad \text{----- ----- (128)}$$

Thus, we have $(n+1)$ algebraic equations to be solved simultaneously in $(n+1)$ unknowns, i.e. $\{u_i : i = 1, \ldots, n+1\}$.

It may be noted that the collocation points need not be chosen equispaced. It has been shown that, if these collocation points are chosen at the roots of n^{th} order orthogonal polynomial, then the error $|u^*(z) - u(z)|$ is evenly distributed in the entire domain of z [2]. For example, one possibility is to choose the orthogonal collocation points at the roots of shifted Legandre polynomials (see Table1). In fact, the name *orthogonal collocation* can be attributed to the choice the collocation points at the roots of orthogonal polynomials.

Discretization using orthogonal collocation technique requires computation of vectors $\{(\mathbf{s}^{(i)}, \mathbf{t}^{(i)}) : i = 1, 2, \ldots n+1\}$, which can be accomplished by solving the following matrix equations: Let us define matrices $\mathbf{S}$ and $\mathbf{T}$ such that these vectors form rows of these matrices, i.e.

$$\mathbf{S} = \begin{bmatrix} \left[\mathbf{s}^{(1)}\right]^T \\ \left[\mathbf{s}^{(2)}\right]^T \\ \ldots. \\ \left[\mathbf{s}^{(n+1)}\right]^T \end{bmatrix} \quad ; \quad \mathbf{T} = \begin{bmatrix} \left[\mathbf{t}^{(1)}\right]^T \\ \left[\mathbf{t}^{(2)}\right]^T \\ \ldots. \\ \left[\mathbf{t}^{(n+1)}\right]^T \end{bmatrix} \qquad \text{----- ----- (129)}$$

Table 1: Roots of Shifted Legandre Polynomials

Order (m)	Roots
1	0.5
2	0.21132, 0.78868
3	0.1127, 0.5, 0.8873
4	0.9305,0.6703, 0.3297, 0.0695
5	0.9543, 0.7662, 0.5034, 0.2286, 0.0475
6	0.9698, 0.8221, 0.6262, 0.3792, 0.1681, 0.0346
7	0.9740, 0.8667, 0.7151, 0.4853, 0.3076, 0.1246, 0.0267

In addition, let us define matrices $\mathbf{C}$ and $\mathbf{D}$ as follows

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & .... & (n)(z_0)^{n-1} \\ 0 & 1 & .... & (n)(z_1)^{n-1} \\ .... & .... & .... & .... \\ 0 & 1 & .... & (n)(z_n)^{n-1} \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 2 & 6z_0 & .. & n(n-1)(z_0)^{n-2} \\ 0 & 0 & 2 & 6z_1 & .. & n(n-1)(z_1)^{n-2} \\ .... & .... & .... & .... & .. & .... \\ 0 & 0 & 2 & 6z_n & .. & n(n-1)(z_n)^{n-2} \end{bmatrix}$$

It is easy to see that

$$\mathbf{S} = \mathbf{C}\mathbf{A}^{-1} \quad ; \quad \mathbf{T} = \mathbf{D}\mathbf{A}^{-1} \qquad (130)$$

where matrix $\mathbf{A}$ is defined by equation (86).

Example 15 [2] Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out. Using method of orthogonal collocation with $n = 4$ and defining vector

$$\mathbf{C} = \begin{bmatrix} C_1 & C_2 & \dots & C_5 \end{bmatrix}^T$$

at

$$z_1 = 0, z_2 = 0.1127, z_3 = 0.5, z_4 = 0.8873 \; and \; z_5 = 1$$

the matrices $\mathbf{A}, \mathbf{S}$ and $\mathbf{T}$ for the selected set of collocation points are as follows

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0.1127 & (0.1127)^2 & (0.1127)^3 & (0.1127)^4 \\ 1 & 0.5 & (0.5)^2 & (0.5)^3 & (0.5)^4 \\ 1 & 0.8873 & (0.8873)^2 & (0.8873)^3 & (0.8873)^4 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad (131)$$

$$\mathbf{S} = \begin{bmatrix} \left[\mathbf{s}^{(1)}\right]^T \\ \left[\mathbf{s}^{(2)}\right]^T \\ \left[\mathbf{s}^{(3)}\right]^T \\ \left[\mathbf{s}^{(4)}\right]^T \\ \left[\mathbf{s}^{(5)}\right]^T \end{bmatrix} = \begin{bmatrix} -13 & 14.79 & -2.67 & 1.88 & -1 \\ -5.32 & 3.87 & 2.07 & -1.29 & 0.68 \\ 1.5 & -3.23 & 0 & 3.23 & -1.5 \\ -0.68 & 1.29 & -2.07 & -3.87 & 5.32 \\ 1 & -1.88 & 2.67 & -14.79 & 13 \end{bmatrix} \qquad (132)$$

$$\mathbf{T} = \begin{bmatrix} \left[\mathbf{t}^{(1)}\right]^T \\ \left[\mathbf{t}^{(2)}\right]^T \\ \left[\mathbf{t}^{(3)}\right]^T \\ \left[\mathbf{t}^{(4)}\right]^T \\ \left[\mathbf{t}^{(5)}\right]^T \end{bmatrix} = \begin{bmatrix} 84 & -122.06 & 58.67 & -44.60 & 24 \\ 53.24 & -73.33 & 26.67 & -13.33 & 6.76 \\ -6 & 16.67 & -21.33 & 16.67 & -6 \\ 6.76 & -13.33 & 26.67 & -73.33 & 53.24 \\ 24 & -44.60 & 58.67 & -122.06 & 84 \end{bmatrix} \qquad \text{----- ----- (133)}$$

Forcing the residual to zero at the internal grid points and using the two boundary conditions we get following set of five simultaneous nonlinear algebraic equations:

$$\frac{1}{Pe}\left[\left[\mathbf{t}^{(i)}\right]^T\mathbf{C}\right] - \left[\left(\mathbf{s}^{(i)}\right)^T\mathbf{C}\right] - DaC_i^2 = 0$$

$$i = 2,3,4$$

These equations can be expanded as follows

$$\begin{bmatrix} \frac{53.24}{Pe} + 5.32 & \frac{-73.33}{Pe} - 3.87 & \frac{26.67}{Pe} - 2.07 & \frac{-13.33}{Pe} + 1.29 & \frac{6.76}{Pe} - 0.68 \\ \frac{-6}{Pe} - 1.5 & \frac{16.67}{Pe} + 3.23 & \frac{-21.33}{Pe} & \frac{16.67}{Pe} - 3.23 & \frac{-6}{Pe} + 1.5 \\ \frac{6.76}{Pe} + 0.68 & \frac{-13.33}{Pe} - 1.29 & \frac{26.67}{Pe} + 2.07 & \frac{-73.33}{Pe} + 3.87 & \frac{53.24}{Pe} - 5.32 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \end{bmatrix}$$

$$- Da \begin{bmatrix} C_2^2 \\ C_3^2 \\ C_4^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The remaining two equations are obtained by discretization of the bourdary consitions.

$$B.C.1 : \left[\left[\mathbf{s}^{(1)}\right]^T\mathbf{C}\right] - Pe(C_1 - 1) = 0$$
$$B.C.2 : \left[\left[\mathbf{s}^{(5)}\right]^T\mathbf{C}\right] = 0$$

or in the expanded form, we have

$$(-13 - Pe)C_1 + 14.79C_2 - 2.67C_3 + 1.88C_4 - C_5 + P_e = 0$$

$$C_1 - 1.88C_2 + 2.67C_3 - 14.79C_4 + 13C_5 = 0$$

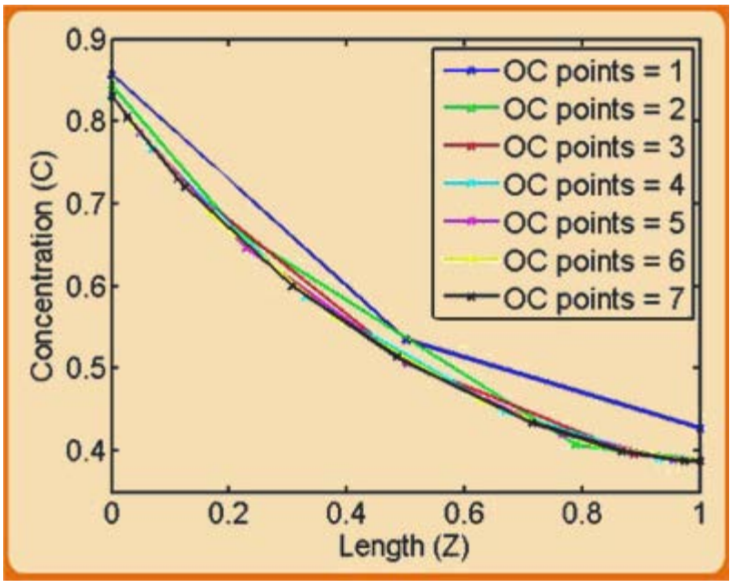

Figure 3: TRAM Problem: Comparison of approximate solutions constructed using different number of orthogonal collocation points.

Thus, the discretization yields a set of five nonlinear algebraic equations in five unknowns, which have to be solved simultaneously.

To provide some insights into how the approximate solutions change as a function of the choice number of collocation points, we have carried out studies on the TRAM problem (with Pe = 6 and Da = 2). Figure 3 demonstrates how the approximate solutions behave as a function of number of collocation points. As evident from this figure, better soltions are obtained as the number of collocations points increase.

Remark 16

Are the two methods presented above, i.e. finite difference and collocation methods, doing something fundamentally different? Let us compare the following two cases (a) finite difference method with 3 internal grid points (b) collocation with 3 internal grid points on the basis of expressions used for approximating the first and second order derivatives computed at one of the grid points. For the sake of comparison, we have taken equi-spaced grid points for the collocation method instead of taking them at the roots of 3'rd order orthogonal polynomial. Thus, for both collocation and finite difference method, the grid (or collocation) points are at $\{z_1 = 0, z_2 = 1/4, z_3 = 1/2, z_4 = 3/4, z_5 = 1\}$. Let us compare expressions for approximate derivatives at $z = z_3$ used in both the approaches.

- **Finite Difference**

$$\frac{du(z_3)}{dz} = \frac{(u_4 - u_2)}{2(\Delta z)} = 2u_4 - 2u_2 \;\; ; \;\; \Delta z = 1/4$$

$$\frac{d^2u(z_3)}{dz^2} = \frac{(u_4 - 2u_3 + u_2)}{(\Delta z)^2} = 16u_4 - 32u_3 + 16u_2$$

- **Collocation**

$$\frac{du(z_3)}{dz} = 0.33u_1 - 2.67u_2 + 2.67u_4 - 0.33u_5$$

$$\frac{d^2u(z_3)}{dz^2} = -1.33u_1 + 21.33u_2 - 40u_3 + 21.33u_4 - 1.33u_5$$

It is clear from the above expressions that the essential difference between the two approaches is the way the derivatives at any grid (or collocation) point is approximated. The finite difference method takes only immediate neighboring points for approximating the derivatives while the collocation method finds derivatives as weighted sum of all the collocation (grid) points. As a consequence, the approximate solutions generated by these approaches will be different.

4.2 Discretization of PDE's [2]

Example 17 Consider the PDE describing unsteady state conditions in a tubular reactor with axial mixing (TRAM) given earlier. Using method of orthogonal collocation with $n-1$ internal collocation points, we get

$$\frac{dC_i(t)}{dt} = \frac{1}{Pe}\left[\left[\mathbf{t}^{(i)}\right]^T\mathbf{C}(t)\right] - \left[\left(\mathbf{s}^{(i)}\right)^T\mathbf{C}(t)\right] - DaC_i(t)^2$$

$$i = 2,3,\ldots n$$

where

$$\mathbf{C}(t) = \begin{bmatrix} C_1(t) & C_2(t) & \ldots & C_{n+1}(t) \end{bmatrix}$$

$C_i(t)$ represents time varying concentration at the i'th collocation point, $C(z_i,t)$, and the vectors $\left[\mathbf{t}^{(i)}\right]^T$ and $\left(\mathbf{s}^{(i)}\right)^T$ represent row vectors of matrices $\mathbf{T}$ and $\mathbf{S}$. defined by equation (129). The two boundary conditions yield the following algebraic constraints

$$\left[\left[\mathbf{s}^{(1)}\right]^T\mathbf{C}(t)\right] = Pe(C_1(t)-1)$$
$$\left[\left[\mathbf{s}^{(n+1)}\right]^T\mathbf{C}(t)\right] = 0$$

Thus, the process of discretization in this case yields a set of differential algebraic equations of the form

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x},\mathbf{z})$$
$$\overline{0} = G(\mathbf{x},\mathbf{z})$$

which have to be solved simultaneously subject to the specified initial conditions on $(\mathbf{x},\mathbf{z})$. In the present case, since the algebraic constraints are linear, they can be used to eliminate variables $C_1(t)$ and $C_{n+1}(t)$ from the set of ODEs resulting from discretization. For example, when we select 3 internal grid points as discussed in Example 15, the boundary constraints can be stated as follows

$$-(13+Pe)C_1(t) + 14.79C_2(t) - 2.67C_3(t) + 1.88C_4(t) - C_5(t) = -Pe$$

$$C_1(t) - 1.88C_2(t) + 2.67C_3(t) - 14.79C_4(t) + 13C_5(t) = 0$$

These equations can be used to eliminate variables $C_0(t)$ and $C_4(t)$ from the three ODEs $\{C_1(t),C_2(t),C_3(t)\}$ by solving the following linear algebraic equation

$$\begin{bmatrix} -(13+Pe) & -1 \\ 1 & 13 \end{bmatrix}\begin{bmatrix} C_1(t) \\ C_5(t) \end{bmatrix} = \begin{bmatrix} -14.79C_2(t) + 2.67C_3(t) - 1.88C_4(t) - Pe \\ 1.88C_2(t) - 2.67C_3(t) + 14.79C_4(t) \end{bmatrix}$$

Thus,the resulting set of (n-1) ODEs together with initial conditions

$$C_2(0) = f(z_2),\ldots\ldots C_n(0) = f(z_n) \qquad (134)$$

is the discretized problem.

Example 18[2]

Consider the 2-dimensional Laplace equation given in Example 12. We consider a scenario where the thermal diffusivity α is function of temperature. To begin with, we choose $(n_x - 1)$ internal collocation points along x-axis and $(n_y - 1)$ internal collocation points along the y-axis. Using $n_x - 1$ internal grid lines parallel to y axis and $n_y - 1$ grid lines parallel to x-axis, we get $(n_x - 1) \times (n_y - 1)$ internal collocation points. Corresponding to the chosen collocation points, we can compute matrices $(\mathbf{S}_x, \mathbf{T}_x)$ and $(\mathbf{S}_y, \mathbf{T}_y)$ using equations (130). Using these matrices, the PDE can be transformed as to a set of coupled algebraic equations as follows

$$\alpha(T_{i,j})\left[\left(\mathbf{t}_x^{(i)}\right)^T T_x^{(j)} + \left(\mathbf{t}_y^{(j)}\right)^T T_y^{(i)} \right] = f(x_i, y_j)$$

$$i = 2, \ldots n_x \; ; \; j = 2, \ldots n_y$$

where vectors $T_x^{(j)}$ and $T_y^{(i)}$ are defined as

$$T_x^{(j)} = \begin{bmatrix} T_{1,j} & T_{2,j} & \ldots & T_{n_x+1,j} \end{bmatrix}$$

$$T_y^{(i)} = \begin{bmatrix} T_{i,1} & T_{i,2} & \ldots & T_{i,n_y+1} \end{bmatrix}$$

At the boundaries, we have

$$T_{0,j} = T^* \; ; \; (j = 1, \ldots n_y + 1)$$
$$T_{1,j} = T^* \; ; \; (j = 1, \ldots n_y + 1)$$
$$T_{i,0} = T^* \; ; \; (i = 1, \ldots n_x + 1)$$
$$k\left[\mathbf{s}_{n_x+1}^{(i)} \right]^T T_x^{(n_y+1)} = h(T_\infty - T_{x,i}^{(n_y+1)}) \quad for \quad (i = 2, \ldots n_x)$$

The above discretization procedure yields a set of $(n_x + 1) \times (n_y + 1)$ nonlinear algebraic equations in $(n_x + 1) \times (n_y + 1)$ unknowns, which have to be solved simultaneously.

To get better insight into discretization, let us consider scenario where we choose three internal collocation points each along x and y directions. This implies that $(\mathbf{S}_x = \mathbf{S}_y = \mathbf{S})$ and $(\mathbf{T}_y = \mathbf{T}_x = \mathbf{T})$ where $\mathbf{S}$ and $\mathbf{T}$ matrices are given in Example 15. Now, at an internal collocation point, say (x_2, y_3), the residual can be stated as follows

$$\alpha(T_{2,3})\left[\left(\mathbf{t}^{(2)}\right)^T T_x^{(3)} + \left(\mathbf{t}^{(3)}\right)^T T_y^{(2)} \right] = f(x_2, y_3)$$

$$T_x^{(3)} = \begin{bmatrix} T_{1,3} & T_{2,3} & T_{3,3} & T_{4,3} & T_{5,3} \end{bmatrix}$$

$$T_y^{(2)} = \begin{bmatrix} T_{2,1} & T_{2,2} & T_{2,3} & T_{2,4} & T_{2,5} \end{bmatrix}$$

$$\alpha(T_{2,3})\left\{\ 53.24T_{1,3}\ -73.33T_{2,3}\ +26.67T_{3,3}\ -13.33T_{4,3}\ +6.76T_{5,3}\ \right\}$$
$$+\alpha(T_{2,3})\left\{\ -6T_{2,1}\ +16.67T_{2,2}\ -21.33T_{2,3}\ \ 16.67T_{2,4}\ -6T_{2,5}\ \right\}$$
$$=f(x_2,y_3)$$

5 Orthogonal Collocations on Finite Elements (OCFE)

The main difficulty with polynomial interpolation is that Vandermond matrix becomes ill conditioned when the order of interpolation polynomial is selected to be large. A remedy to this problem is to sub-divide the region into finite elements and assume a lower order polynomial spline solution. The collocation points are then selected within each finite element, where the residuals are forced to zero. The continuity conditions (equal slopes) at the boundaries of neighboring finite elements gives rise to additional constraints. We illustrate this method by taking a specific example.

Example 19 [2] Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out. It is desired to solve this problem by OCFE approach.

Step 1: The first step is to create finite elements in the domain. Let us assume that we create 3 sub-domains. Finite Element 1: $0 \le z \le 0.3$, Finite Element 2: $0.3 \le z \le 0.7$, Finite Element 3: $0.7 \le z \le 1$. It may be noted that these sub-domains need not be equi-sized.

Step 2: On each finite element, we define a scaled spacial variable as follows

$$\zeta_1 = \frac{z - Z_1}{Z_2 - Z_1}, \zeta_2 = \frac{z - Z_2}{Z_3 - Z_2} \text{ and } \zeta_3 = \frac{z - Z_3}{Z_4 - Z_3}$$

where $Z_1 = 0, Z_2 = 0.3$, $Z_3 = 0.7$ and $Z_4 = 1$ represent the boundary points of the finite elements. It is desired to develop a polynomial spline solution such that polynomial on each finite element is 4'th order. Thus, within each element, we select 3 collocation points at the root of the 3'rd order shifted Legandre polynomial, i.e.,

$$\zeta_{i,1} = 0.1127, \zeta_{i,2} = 0.5 \text{ and } \zeta_{i,3} = 0.8873 \text{ for } i = 1,2,3$$

In other words, collocation points are placed at

$$Z_i + 0.1127(Z_{i+1} - Z_i),\ Z_i + 0.5(Z_{i+1} - Z_i), \text{ and } Z_i + 0.8873(Z_{i+1} - Z_i) \quad \text{for } i = 1,2,3$$

in the i'th element $Z_i \le z \le Z_{i+1}$. Thus, in the present case, we have total of 9 collocation points. In addition, we have two points where the neighboring polynomials meet, i.e. at $Z_1 = 0.3$ and $Z_2 = 0.7$. Thus, there are total of 11 internal points and two boundary points, i.e. $Z_1 = 0$ and $Z_2 = 1$.

Step 3: Let the total set of points created in the previous step be denoted as $\{z_1, z_1, \ldots z_{13}\}$ and let the corresponding values of the independent variables be denoted as $\{C_1, C_1, \ldots, C_{13}\}$. Note that variables associate with each of the finite elements are as follows

$$\text{Finite Element 1} \quad \mathbf{C}^{(1)} = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 \end{bmatrix}^T$$
$$\text{Finite Element 2} \quad \mathbf{C}^{(2)} = \begin{bmatrix} C_5 & C_6 & C_7 & C_8 & C_9 \end{bmatrix}^T$$
$$\text{Finite Element 3} \quad \mathbf{C}^{(3)} = \begin{bmatrix} C_9 & C_{10} & C_{11} & C_{12} & C_{13} \end{bmatrix}^T$$

Now, we force residuals to zero at all the internal collocation points within a finite element. Let h_1, h_2 and

h_3 denote length of individual finite elements, i.e.

$$h_1 = Z_2 - Z_1,\ h_2 = Z_3 - Z_2 \text{ and } h_3 = Z_4 - Z_3 \quad (135)$$

Defining scaled spatial variables

$$\zeta_i = \frac{z - Z_i}{Z_{i+1} - Z_i} = \frac{z - Z_i}{h_i}$$

for i = 1,2,3, the ODE in each finite element is modified as follows

$$\frac{1}{Pe}\left(\frac{1}{h_i^2}\right)\frac{d^2C}{d\zeta_i^2} - \left(\frac{1}{h_i}\right)\frac{dC}{d\zeta_i} - DaC^2 = 0 \qquad for \quad Z_i \le z \le Z_{i+1} \text{ and } i = 1,2,3 \quad (136)$$

The main difference here is that only the variables associated within an element are used while discretizing the derivatives. Thus, at the collocation point z_2 in finite element 1, the residual is computed as follows

$$R_2 = \frac{1}{Pe}\left(\frac{1}{h_1^2}\right)\left[\mathbf{t}^{(2)}\right]^T\mathbf{C}^{(2)} - \left(\frac{1}{h_1}\right)\left[\mathbf{s}^{(2)}\right]^T\mathbf{C}^{(1)} - Da(C_2)^2 = 0 \quad (137)$$

$$\left[\mathbf{t}^{(2)}\right]^T\mathbf{C}^{(1)} = (53.24C_1 - 73.33C_2 + 26.27C_3 - 13.33C_4 + 6.67C_5)$$

$$\left[\mathbf{s}^{(2)}\right]^T\mathbf{C}^{(1)} = (-5.32C_1 + 3.87C_2 + 2.07C_3 - 1.29C_4 + 0.68C_5)$$

where vectors $\left[\mathbf{s}^{(2)}\right]^T$ and $\left[\mathbf{t}^{(2)}\right]^T$ are 2^{nd} rows of matrices (132) and (133), respectively. Similarly, at the collocation point $z = z_8$, which corresponds to $\zeta_{i,3} = 0.8873$ in finite element 2, the residual is computed as follows

$$R_8 = \frac{1}{Pe}\left(\frac{1}{h_2^2}\right)\left[\mathbf{t}^{(3)}\right]^T\mathbf{C}^{(2)} - \left(\frac{1}{h_2}\right)\left[\mathbf{s}^{(3)}\right]^T\mathbf{C}^{(2)} - Da(C_8)^2 = 0 \quad (138)$$

$$\left[\mathbf{t}^{(3)}\right]^T\mathbf{C}^{(2)} = 6.76C_5 - 13.33C_6 + 26.67C_7 - 73.33C_8 + 53.24C_9$$

$$\left[\mathbf{s}^{(2)}\right]^T\mathbf{C}^{(2)} = -0.68C_5 + 1.29C_6 - 2.07C_7 - 3.87C_8 + 5.32C_9$$

Other equations arising from the forcing the residuals to zero are

$$\text{Finite Element 1: } R_3 = R_4 = 0$$

$$\text{Finite Element 2: } R_6 = R_7 = 0$$

$$\text{Finite Element 3: } R_{10} = R_{11} = R_{12} = 0$$

In addition to these 9 equations arising from the residuals at the collocation points, there are two constraints at the collocation points z_4 and z_8, which ensure smoothness between the the two neighboring polynomials, i.e.

$$\left(\frac{1}{h_1}\right)\left[\mathbf{s}^{(5)}\right]^T\mathbf{C}^{(1)} = \left(\frac{1}{h_2}\right)\left[\mathbf{s}^{(1)}\right]^T\mathbf{C}^{(2)}$$

$$\left(\frac{1}{h_2}\right)\left[\mathbf{s}^{(5)}\right]^T\mathbf{C}^{(2)} = \left(\frac{1}{h_3}\right)\left[\mathbf{s}^{(1)}\right]^T\mathbf{C}^{(3)}$$

The remaining two equations come from discretization of the boundary conditions.

$$\left(\frac{1}{h_1}\right)\left[\left[\mathbf{s}^{(1)}\right]^T\mathbf{C}^{(1)}\right] = Pe(C_0 - 1)$$

$$\left(\frac{1}{h_3}\right)\left[\left[\mathbf{s}^{(5)}\right]^T\mathbf{C}^{(3)}\right] = 0$$

Thus, we have 13 equations in 13 unknowns. It may be noted that, when we collect all the equations together, we get the following form of equation

$$\mathbf{AC} = \mathbf{F}(\mathbf{C})$$

$$\mathbf{A} = \begin{bmatrix} A_1 & [\mathbf{0}] & [\mathbf{0}] \\ [\mathbf{0}] & A_2 & [\mathbf{0}] \\ [\mathbf{0}] & [\mathbf{0}] & A_3 \end{bmatrix}_{13\times13}$$

$$\mathbf{C} = \begin{bmatrix} C_0 & C_1 & \dots & C_{12} \end{bmatrix}^T$$

and $\mathbf{F}(\mathbf{C})$ is a 13×1 function vector containing all the nonlinear terms. Here, A_1,A_1 and A_3 are each 5×5 matrices and matrix $\mathbf{A}$ is a sparse *block diagonal* matrix.

The method described above can be easily generalized to any number of finite elements. Also, the method can be extended to the discretization of PDEs in a similar way. These extensions are left to the reader as an exercise and are not discussed separately. Note that block diagonal and sparse matrices naturally arise when we apply this method.

To provide insights into how the approximate solutions change as a function of the choice number of collocation points and finite element, we have carried out studies on the TRAM problem (with Pe = 6 and Da = 2). Figure 4 demonstrates how the approximate solutions behave as a function of number of collocation points when different number of finite elements are constructed such that each segment has three internal collocation points. Finally, solutions obtained using finite difference (FD), orthogonal collocation (OC) and OC on finite elements (OCFE) are compared in Figure 5. This figure demonstrates that orthogonal collocation based approach is able to generate an approximate solution, which is comparable to FD solution with large number of grid points, using significantly less number of collocation points and hence significantly less computational cost.

References and cited materials

1. Gilbert Strang, *Linear Algebra and Its Applications* (*4th Ed.*), Wellesley Cambridge Press (2009).

2. Philips, G. M.,Taylor, P. J. ; Theory and Applications of Numerical Analysis (2nd Ed.),Academic Press, 1996.

3. Gourdin,A. and M Boumhrat; Applied Numerical Methods. Prentice Hall (2000).

4. Gupta, S. K.; Numerical Methods for Engineers. Wiley Eastern, New Delhi, 1995.