Introduction to the Basics

Larry C. Young, tildentechnologies.com

Table of Contents

1. Introduction to the Basics

After having been involved during the early days of orthogonal collocation [Young and Finlayson, 1973, 1976], I had reason to use it again recently. A review of the last 45 years revealed that some aspects of the method appear to have never been explained clearly or never explained at all. Some information in the literature is just plain wrong. This realization prompted the publication of this monograph.

Here we first lay the fundamental framework for the method. Then, we show what works and what does not work using a series of examples. In many cases the examples are solved not only with orthogonal collocation, but also with other methods for comparison, e.g. Galerkin, Moments and finite differences. The examples also cover a variety of problems: boundary value problems, parabolic equations, hyperbolic equations, one and two spatial dimensions. The examples also encompass problems with and without symmetry about x = 0 and with various types of boundary conditions and nonlinearities.

There are four areas where we find explanations in the literature missing, unclear, confusing or just plain wrong, namely:

1. Description of trial functions and the basic method
2. Relationship to other methods, e.g. Galerkin, Moments
3. Accurate representation of boundary fluxes
4. Extensions to finite elements

I hope this monograph improves understanding in these areas.

In previous descriptions of the method, it is often stated the trial functions are orthogonal polynomials, but then sometimes a simple power series of monomials is used in the development. At other times Lagrange interpolating polynomials are used. If the unknowns are the nodal values (at the collocation points), the trial functions are Lagrange interpolating polynomials, plain and simple. The method makes much more sense when developed on this consistent basis. The results in all cases are equivalent, but nodal values are more intuitive.

Of the four areas, the third one is the most important. After all, for engineering calculations, one is most often interested in the exchange of heat, mass, or force between two objects. This invariably involves the calculation of some sort of flux related quantity, e.g. boundary thermal or mass flux, fluid drag due to boundary shear stress, force at the end of a spring or other flexible component. Many problems have boundary conditions involving fluxes, i.e. 2nd or 3rd type conditions. The accurate treatment of fluxes is essential for treatment of these boundary conditions.

The fourth area, extension to finite elements, is directly related to the third. In a finite element context, flux continuity is required at the boundary of adjacent elements. Discussions of orthogonal collocation finite element methods have centered on collocation at Gauss points [Carey and Finlayson, (1975)]. This method generally uses explicit equations to force the continuity of derivatives at the boundary between elements. It would likely be more efficient to use trial functions with built in continuity, e.g. Hermite cubic polynomials. This level of continuity is called C_1 continuity. C_1 methods are difficult (impossible) to extend to irregular shaped elements, e.g. quadrilaterals, in multiple dimensions. In this monograph we show how collocation at Lobatto points naturally extends to a C_0 finite element method with some significant advantages over conventional finite element methods and conventional collocation finite elements (at Gauss points) [Young (1977,1981)].

1.1 Method of Weighted Residuals (MWR)

The collocation method is one of a family of methods collectively called methods of weighted residuals (MWR) [Finlayson (1972)]. These methods start by using a linear combination of trial functions to approximate the solution. Suppose we seek a solution of the equation:

$$u_{xx} + g(x, u, u_x) = 0 \tag{1.1}$$

for $0 < x < 1$, with $u(0) = u_0$ and $u(1) = u_1$. The solution is approximated by:

$$u \cong \tilde{u} = \sum_{j=0}^{n} a_j \, v_j(x) \tag{1.2}$$

Where the functions, $\boldsymbol{v}$, are called trial functions or basis functions. The trial functions are usually but not always required to meet the boundary conditions. The residual is formed by substituting the approximate solution into the differential equation:

$$R(x, \boldsymbol{a}) = \sum_{j=0}^{n} a_j \, v_{jxx} + g(x, \tilde{u}, \tilde{u}_x) \tag{1.3}$$

If the residual is zero for all x, an exact solution has been achieved. In general, this will not be possible, so the parameters, $\boldsymbol{a}$, are selected so the residual will be small in some sense. If the residual is small for all x, Eq. (1.2) will be a good approximation to the true solution. MWR achieve this goal by forcing the function to zero in a weighted average sense:

$$\int_0^1 w_j(x) R(x, \boldsymbol{a}) dx = 0 \ \text{ for } \ j = 0, \ldots, n \tag{1.4}$$

Different choices for the n + 1 weight functions (also called test functions) produce different methods, some of the popular choices are:

1. $w_j = x^j$, Moments method
2. $w_j = v_j(x)$, Galerkin method
3. $w_j = \delta(x - x_j)$, Collocation method

These methods are also called Spectral methods, especially when the trial solution is a Fourier series. However, that terminology is also applied to global polynomial trial functions. The name MWR seems more descriptive of the method. When the method is applied with piecewise polynomial trial functions it is called a Finite Element method or Spectral Element method.

The Moments method forces the residual to zero when weighted by increasingly higher members of a complete set of functions, i.e. the monomials x^j, so as n becomes large the residual must approach zero. For the Moments and Galerkin method, the weight functions can be rearranged or recombined and the result will not change as long as they remain linearly independent. For example, the Moments method does not require use of the monomial powers of x. We could just as easily use Legendre or Chebyshev polynomials, or a linearly independent set of Lagrange interpolating polynomials. This fact is lost on Boyd (2000), as he dismisses the Moments method because monomial powers of x can lead to ill conditioning for large n.

Most texts do not discuss the Moments method, but discuss the Tau method instead [Lanczos (1938, 1956)]. The Tau method is based on a different idea from MWR and the boundary conditions are always imposed through side conditions. However, in most cases it is similar or identical to the Moments method. It makes no difference whether the trial functions are constructed to obey the boundary conditions or whether the boundary conditions are satisfied with side conditions. The end result is the same. As stated above, weighting by Legendre polynomials, $w_j(\mathrm{x}) = P_j(\mathrm{x})$, is equivalent to weighting by monomials. In such cases (e.g. Canuto, et al., (1988), p. 10), the results are identical. Since the Moments method fits better into the MWR framework, we use it instead of the Tau method in this monograph, but the reader should be aware of the similarity.

The Galerkin method weights the residual by the trial functions. Since the trial functions should be chosen from a complete set of functions, this method also guarantees the residual will approach zero for large n. Also, the Galerkin method is related to the Raleigh-Ritz method which is known to produce good results. The Raleigh-Ritz method requires that the solution obey a variational principal such as the minimization of energy. Problems amenable to this treatment are invariably linear. The Galerkin method produces identical approximations for these problems; however, the Galerkin method is a general method, which does not require linearity or a variational principal. Due to this relationship, the method is frequently called the Raleigh-Ritz-Galerkin method.

Some authors take a more general view and refer to virtually all MWR weightings as Galerkin or variational methods. For example, with Chebyshev polynomial trial functions it is common to include the radical $1/\sqrt{1-x^2}$ in the weight function. This term is necessary to exploit the orthogonality of the polynomials and to achieve accurate integration with Gauss-Chebyshev quadrature. Its inclusion is basically a contrivance required by the selection of Chebyshev polynomials and it can lead to some cumbersome mathematics. Furthermore, as discussed in Section 3.1.4 for a method to be conservative, the weight functions must include unity in some combination. When the radical is included, the method cannot be conservative. Methods of this type will not be discussed further in this monograph.

The collocation method sets the residual to zero at selected "collocation" points, so the weight functions may be viewed as dirac delta functions. Since integration is not required, collocation is much simpler and requires less computation than the other MWR. However, the distribution of the collocation points is of critical importance with the method. We will discuss this in much greater detail in subsequent sections and throughout this monograph. Two separate lines of study have developed for this method: (1) *orthogonal collocation* (OC) line following the work of Villadsen and Stewart (1967), Finlayson (1972), and Villadsen and Michelsen (1978) and (2) the *pseudospectral* (PS) line following Orzag (1972) and Gottlieb and Orzag (1977). We say these are separate lines of study because the earlier OC work is rarely referenced in the PS literature, or it is mentioned only in passing. Although there are often differences in implementation details these methods are equivalent when the same collocation points are used.

1.2 Selection of Global Trial Functions

For most problems of practical interest, the general nature of the solution is known. A nice feature of the method of weighted residuals is that the trial or basis functions can be selected to mimic the known characteristics of the solution. Trial function selection should also consider their suitability for calculations with a computer and also to make the method easy to apply and intuitive.

Polynomials are simple functions which are amenable to efficient computer calculations. In this monograph, we are interested in bounded domains, which we take as the interval [0,1]. Some descriptions of orthogonal collocation state that the trial functions are orthogonal polynomials. However, simple monomial powers, x^j, are sometimes used to develop the matrix operators, which work with nodal values, i.e. the $a_j = \tilde{u}(x_j)$ in Eq. (1.2). Some texts imply that the type of representation has an important effect on the accuracy of the result. This is not true. Suppose we choose to approximate the solution with Legendre polynomials, i.e. $v_j(x) = P_j(x)$. As shown in Section 2.5, there is a simple linear transformation between the vector of coefficients, $\boldsymbol{a}$, in this representation and the vector of nodal values, $\tilde{\boldsymbol{u}}$. Excluding machine roundoff errors, the results are identical regardless of the representation. Even for a monomial representation,

roundoff errors are not significant for $n \lesssim 10$. We should then base our selection on efficiency and ease of use.

If the trial functions are orthogonal polynomials, the formulation is called a *modal* formulation, since the trial solution is analogous to a truncated Fourier series and the coefficients, $\boldsymbol{a}$, correspond to the weights given each Fourier mode. Villadsen and Stewart (1967) were the first to propose solving for the values at the points or nodes, $\tilde{\boldsymbol{u}}$, instead of polynomial coefficients. This approach is called a *nodal* method and the trial functions, $v_j(x)$ in Eq. (1.2), are not orthogonal polynomials, but rather the Lagrange interpolating polynomials:

$$v_j(x) = \ell_j(x) = \prod_{\substack{k=0 \\ k\neq j}}^{n+1} \frac{(x - x_k)}{(x_j - x_k)} \tag{1.5}$$

where the first and last nodes are at the boundaries, $x_0 = 0$ and $x_{n+1} = 1$. With this choice of trial function, the approximate solution, Eq. (1.2) is:

$$u(x) \cong \tilde{u}(x) = \sum_{j=0}^{n+1} u_j \, \ell_j(x) \tag{1.6}$$

where $u_j = \tilde{u}(x_j)$.

There are many problems for which the solution is symmetric about $x = 0$. In such cases, symmetry satisfies the condition at $x = 0$, so only one boundary node is needed, $x_{n+1} = 1$. Lagrange interpolating polynomials in x^2 are used for the trial functions:

$$v_j(x^2) = \ell_j(x^2) = \prod_{\substack{k=1 \\ k\neq j}}^{n+1} \frac{(x^2 - x_k^2)}{(x_j^2 - x_k^2)} \tag{1.7}$$

The choice of trial or basis functions is an area where there are implementation differences between the orthogonal collocation (OC) and pseudospectral (PS) literature. Since the pioneering work of Villadsen and Stewart (1967) virtually all OC implementations have used nodal values or Lagrange interpolating polynomial trial functions. PS implementations show a mixture of modal, i.e. orthogonal polynomial, and nodal methods. Modal methods can be more efficient for very large problems, especially for problems which are linear with constant coefficients. They can sometimes be formulated to give banded matrix problems which can be solved more efficiently than the full matrix problems created with nodal methods. When Chebyshev polynomials are used, fast Fourier transforms (FFT) can be used to efficiently solve large problems amenable to explicit time stepping methods. In order for FFT based methods to be more efficient, the problem must be large, i.e. n > 20 to 100 nodes, based on several analyses. Nonlinearities are usually easier to treat with a nodal method. Some references [Canuto, et al. (1988) p. 86 and Boyd (2000), p. 107] recommend switching back and forth between the two representations while solving some nonlinear problems.

Clearly, there are tradeoffs between the modal and nodal representations. The most efficient method depends on the size of the problem, i.e. n, and nonlinearities. Since most practical problems are nonlinear, those will be emphasized here. The nodal method is usually efficient and always more intuitive, since it is finite-difference-like. Apart from some examples shown for comparison, nodal representations are used throughout this monograph.

The discussion above pertains only to global trial functions. If finite element trial functions are also considered, the selection becomes more complicated. It is our opinion, that for $n \gtrsim 10$, a finite element approach based on orthogonal collocation will invariably be more efficient, especially if high order elements are considered, e.g. 4th order or greater. One can vary the number of elements and their order to search for an optimum for a particular class of problem.

1.3 Approximation of Integrals

When solving problems with the Galerkin method or the Method of Moments, one rarely evaluates the integrals analytically, especially if the problem is nonlinear. Instead, the integrals are usually approximated numerically using a quadrature formula of the form:

$$\int_0^1 f(x)x^\gamma dx \cong \sum_{i=1}^{m} W_i f(x_i) \tag{1.8}$$

where $f(x)$ is some function, x_i and W_i are the quadrature base points and weights, respectively. The term x^γ is included, since we are interested in cylindrical (γ = 1) and spherical (γ = 2) coordinates as well as normal planar geometry (γ = 0). Quadrature formulas like this are frequently normalized to the interval x = [-1,1], however we will use the interval [0,1].

Considering the number of function evaluations required, the most efficient method for approximating integrals is Gaussian quadrature [see for example Hildebrand (1987)]. It gives the exact integral for polynomials through degree $2m$ - 1. The Gaussian quadrature base points are the roots of Legendre orthogonal polynomials and all lie within the interval, 0 < x_i < 1, i.e. no points occur on or outside the boundaries. The points and weights are shown in Figs.

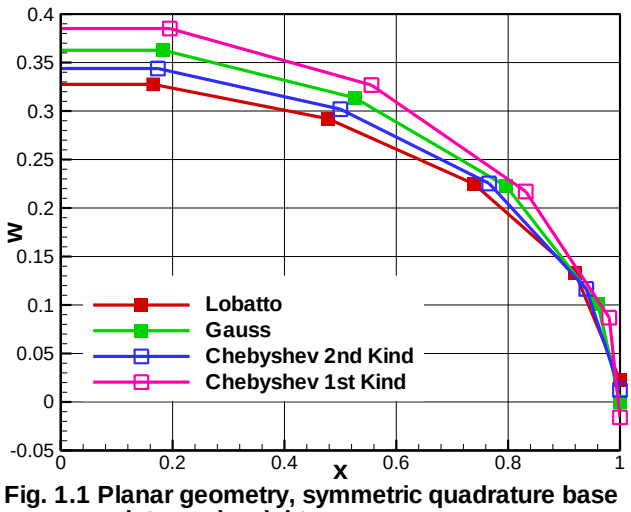

Fig. 1.1 Planar geometry, symmetric quadrature base points and weights

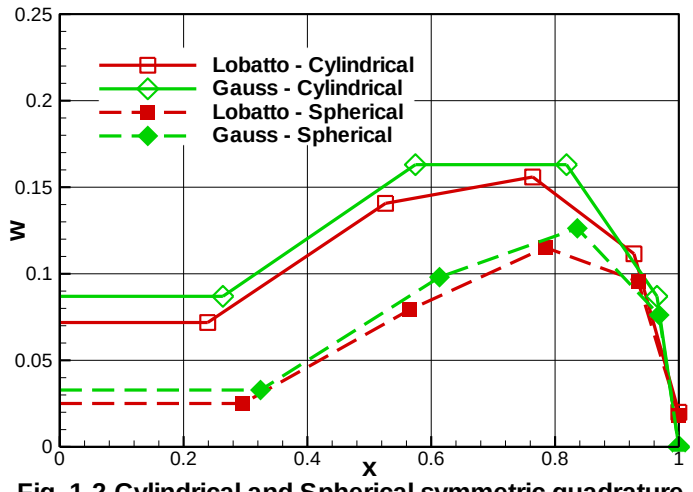

Fig. 1.2 Cylindrical and Spherical symmetric quadrature base points and weights

1.1 and 1.2 for symmetric problems, which correspond to those for the right half of a nonsymmetric problem for planar geometry.

We will also find it useful to use quadrature formulas which include the endpoints, i.e. $x_1 = 0$ and $x_m = 1$. The most efficient method of this type is Lobatto quadrature. It gives the exact integrals for polynomials through degree $2m - 3$. The other base points for Lobatto quadrature are the roots of Jacobi polynomials, and all lie within the interval. These points also correspond to the extrema of Legendre polynomials. Compared to the Gauss points, these points are further from the boundary.

Many articles and texts claim that quadrature weights on the boundaries somehow facilitate approximation of the boundary conditions. In this monograph we will show that in some cases a boundary weight is an asset and in other cases it can be a detriment. The endpoints or boundaries are always included to facilitate approximation of the boundary conditions, regardless of the value of the corresponding quadrature weight.

Lobatto quadrature gives nonzero weights at the endpoints, while the weights are zero there for Gaussian quadrature. In terms of the number of interior points, n, Gaussian quadrature is exact through degree $2n - 1$, while Lobatto quadrature is exact through degree $2n + 1$. We will show that for many cases, the additional two degrees of accuracy are achieved for little or no effort. There is also the related Radau quadrature which uses only one of the endpoints, i.e. $x_1 = 0$ or $x_m = 1$ but not both. It is exact for polynomials through degree $2m - 2$ or $2n$.

Since the work of Lanczos (1938, 1956) many pseudospectral (PS) applications [e.g. see Boyd (2000), Trefethen (2000)] use collocation at the roots of Chebyshev polynomials of the 2nd kind. For a general integral like Eq. (1.8), the most accurate quadrature utilizing Chebyshev roots is Clenshaw-Curtis quadrature [Clenshaw and Curtis (1960)]. With the use of nonzero weights at the endpoints, Chebyshev polynomials of the 1st and 2nd kind both achieve exact integration for polynomials through degree $n + 1$. With this quadrature, the base points are not optimally chosen to give the $O(2n)$ accuracy of Gauss or Lobatto quadrature. The weights plotted in Fig. 1.1 give no indication of this significant difference in accuracy.

1.4 Selection of Collocation Points

Since the collocation method simply sets the residual to zero at the collocation points, it requires much less computation than the other MWR which require integration. However, the distribution of the collocation points is of critical importance. Early applications of collocation using equally spaced points were often disappointing. Lanczos (1938, 1956) was the first to apply collocation at the roots of orthogonal polynomials. Lanczos looked at the problem from the viewpoint of approximation. With MWR one tries to interpolate the residual to coincide with zero. This approach lead Lanczos to choose the roots of Chebyshev polynomials of the 2nd kind. Chebyshev roots are widely used in the Pseudospectral (PS) literature.

Villadsen and Stewart (1967) viewed collocation as an approximation of other integrated MWR. This lead them to use roots of the Jacobi polynomials which are the base points of Lobatto quadrature. They showed this choice can produce an accurate (sometimes exact) approximation of the Galerkin method. Villadsen and Stewart were the first to call the method orthogonal collocation. Finlayson (1972), used the roots of both Jacobi and Legendre polynomials (Gaussian quadrature base points). These references and Villadsen and Michelsen (1978) also cover points for symmetric problems in planar, cylindrical and spherical coordinates. In Chapter 3, we will show that Gaussian quadrature base points produce an accurate approximation of the Moments method. By choosing collocation points that are the base points of accurate quadrature schemes, the method achieves accuracy similar to the Galerkin or Moments method, while maintaining the simplicity of the collocation method.

A variety of orthogonal polynomial roots have been used in the OC and PS literature. The most common choices are the roots of one of the following:

1. ***Chebyshev Polynomials of the 1st Kind*** – also called *Chebyshev Roots* or a *Roots Grid*.
2. ***Chebyshev Polynomials of the 2nd Kind*** – also called *Chebyshev Extrema* points or *Chebyshev-Gauss-Lobatto* or abbreviated CGL points or simply *Gauss-Lobatto* or *Chebyshev* points.
3. ***Legendre Polynomials*** – also called *Gaussian quadrature base points or abscissa* or *Legendre-Gauss* or abbreviated LG points or simply *Legendre* or *Gauss* points.
4. ***Jacobi Polynomials*** – also called *Legendre Extrema* points or *Jacobi* points or *Lobatto quadrature base points or abscissa* or *Legendre-Gauss-Lobatto* abbreviated LGL points or simply *Gauss-Lobatto* or *Lobatto* points

This becomes a veritable minefield of terminology, especially when authors abbreviate “Chebyshev-Gauss-Lobatto” to simply “Gauss-Lobatto” making the name indistinguishable from Lobatto quadrature base points. We will stick to terminology which appears in most mainstream texts on numerical analysis, e.g. Hildebrand (1987). We will label the last three listed above as Chebyshev, Gauss and Lobatto points, respectively. This should not cause any confusion, since we will not use Chebyshev polynomials of the 1st kind in this monograph.

All of the collocation points are roots of a specific Jacobi polynomial which are designated by parameters α and β. When defined on the interval [0,1] these polynomials obey the orthogonality condition:

$$\int_0^1 (1-x)^\alpha x^\beta P_k^{(\alpha,\beta)}(x) P_\ell^{(\alpha,\beta)}(x) dx = 0 \quad \text{for } k \neq \ell \tag{1.9}$$

where the special cases are: Chebyshev 1st Kind, $\alpha = \beta = -½$; Gauss (Legendre), $\alpha = \beta = 0$; Chebyshev 2nd Kind, $\alpha = \beta = ½$; Lobatto, $\alpha = \beta = 1$. Fig. 1.1 shows the distribution of the points along with the quadrature weights used for approximate integration. For nonsymmetric problems the points in Fig. 1.1 are shifted to the interval [½,1] and their reflection would appear

on [0,½]. In all cases the points are more densely spaced near the boundary. By comparing the points in Fig. 1.1 it is apparent that when $\alpha = \beta$, larger values shift the roots toward the center, so they lie in the following order:

$$x_k^T < x_k^G < x_k^C < x_k^L \text{ for } x_k < \tfrac{1}{2} \text{ and } x_k^L < x_k^C < x_k^G < x_k^T \text{ for } x_k > \tfrac{1}{2}$$

where x^T, x^G, x^C and x^L designate Chebyshev 1st kind, Gauss, Chebyshev 2nd kind and Lobatto points, respectively. Note the Chebyshev 2nd kind points are between the Gauss and Lobatto points.

Figs. 1.2 shows the points and weights for cylindrical and spherical geometry. As discussed in Chapter 2, the values of β are -½, 0, and +½ for planar, cylindrical and spherical geometry, while the values are again $\alpha = 0$ for Gauss and $\alpha = 1$ for Lobatto points. Larger values of β shift the points toward the outer boundary where the volume is greater as one would expect from intuition. The main point of this discussion is that the points are denser near the boundary, but the weighting (α and β) shifts their location too.

The points used most frequently in the PS literature differ from those commonly used in OC applications. It appears that in PS applications Chebyshev (2nd kind) points are used in the majority of cases, with Lobatto points used less frequently. Other choices are seldom used. In OC applications, Chebyshev points are rarely used; Gauss points are most frequently used, while Lobatto points are used to a lesser extent.

In early applications of orthogonal collocation, the use of Lobatto points was found to yield somewhat greater accuracy than the use of Gauss points, especially for a small number of collocation points and Dirichlet boundary conditions. However, it was soon discovered that Lobatto points give poor results when a flux boundary condition (2nd or 3rd type) is used [see Finlayson (1972), Table 5.7 and Fig. 5.7; Elnashaie and Cresswell (1973)]. Boundary conditions of this type were satisfied exactly, which is called *boundary collocation*. Boundary collocation is the recommended treatment in most texts in the OC and PS literature [Finlayson (1972), p. 101; Villadsen and Michelsen (1978), p. 137]. After these early results Lobatto points fell from favor and Gauss points became the preferred choice in the OC literature. The recommended use of only Gauss points for flux boundary conditions is buried in Villadsen and Michelsen (1978, p. 249), but was stated earlier in less frequented locations [Ferguson (1971), Young (1974), Young and Finlayson (1976) and probably others].

Apparently, this problem with the treatment of fluxes is not widely known in the PS literature, since boundary collocation is recommended in more recent texts [Trefethen (2000), p. 137; Boyd (2000), p. 111, Peyret (2002), p. 59] despite the fact that similar poor results occur when boundary collocation is used with Chebyshev points, which are popular in the PS literature.

Rather than use boundary collocation, we demonstrate through examples that a simple natural boundary condition treatment works very well when other than Gauss points are used. A natural treatment of flux conditions is standard for applications of the Galerkin method with

both global and finite element trial functions. Others have proposed similar alternatives to boundary collocation [Canuto, *et al*. (1988), Funaro (1992), Shen and Tang (2006)]; but these methods apparently have not gained widespread use.

The accurate representation of fluxes has received relatively little attention in the mathematics literature. Many examples consider only Dirichlet boundary conditions and never bother to determine the accuracy of flux calculations. However, engineers often care nothing about the accuracy of the calculated internal profiles, only the interchange of energy, mass, force, etc. between objects is important, which requires accurate knowledge of the boundary flux.

1.5 Road Map to this Monograph

The primary purpose of this monograph is to use examples to illustrate how to apply Orthogonal Collocation or Pseudospectral methods which work well for nonperiodic problems.

Chapter 2 describes the fundamental relationships and calculation methods which form the building blocks for MWR. These calculations can be successfully applied to solve problems without a complete understanding of the material in this chapter. One may choose to initially scan this material and look at the examples and the available software. Further study is warranted for those seeking a deeper understanding. The chapter describes orthogonal polynomials and interpolating polynomials, calculation of the polynomial roots or collocation points, quadrature weights, differentiation matrices for modal and nodal methods and methods for conversion between the two representations. The computer codes available (in Matlab/Octave, Fortran, C++ and Excel) to implement the methods are also described. Several examples demonstrate some of the basic calculations. This chapter was originally envisioned as an appendix, but grew to such proportions that it was elevated to a chapter.

Chapter 3 treats two boundary value problems. The first is a diffusion problem with a source function and the second is a coupled convection dominated chemical reactor model. Several variations of these problems are covered: linear and nonlinear source, various boundary conditions, constant and variable coefficients, asymmetric and symmetric problems, various geometry, nodal and modal solutions. The problems are solved with collocation, Galerkin and Moments methods. Flux calculations, the treatment of flux boundary conditions and mass conservation are analyzed. This chapter is fundamental to the later ones, since much of the material learned is applied to other more complex problems.

Chapter 4 considers parabolic problems. The first one considered is the falling liquid film problem studied by Villadsen and Michelsen (1978, ch. 4). This problem is solved with the Galerkin method and with collocation using five different choices of collocation points. The no flux condition used at one boundary, provides an opportunity for further testing the treatment of flux boundary conditions. For this problem the first order derivative term is in the axial spatial coordinate, i.e. z, instead of time which is usually the case. The problem is solved both analytically in z and by various stepping methods, Runge-Kutta, etc. The consideration of both

analytical and numerical solutions in z gives valuable insight into the performance of various stepping methods. Nonlinear example to be added.

More to follow …….