UNIVERSITÉ PARIS XI
U.E.R. MATHÉMATIQUE
91405 ORSAY FRANCE

N° 219 - 7667

THE COMPUTATIONAL THEORY OF STIFF DIFFERENTIAL EQUATIONS

par

Willard L. MIRANKER

(Publication Mathématique d'Orsay)

PREFACE

Ces notes de cours correspondent à deux cours enseignés par l'auteur durant l'année 1974-75 à l'Université de Paris-Sud puis à l'Istituto per le Applicazioni del Calcolo "Mauro Picone", à Rome.

La typographie du texte a été assurée par l'Institut Mauro Picone.

Table des Matières

§ 1. INTRODUCTION

1.1 Motivation

Stiff differential equations are equations which are ill-conditioned in a computational sense. To reveal the nature of the ill-conditioning and to motivate the need to study numerical methods for stiff differential equations, let us consider an elementary error analysis for the initial value problem

1.1)
$$\dot{y} = -Ay, \qquad 0 < t \leqslant \bar{t},$$
$$y(0) = y_0$$

Here y is an m-vector and A is a constant $m \times m$ matrix. The dot denotes time differentiation. Corresponding to the increment $h > 0$, we introduce the mesh points $t_n = nh$, $n = 0, 1, \ldots$. If

$$y_n = y(t_n),$$

the solution to (1.1) obeys the recurrence relation,

1.2)
$$y_{n+1} = e^{-Ah} y_n.$$

For convenience we introduce the function $S(z) = e^{-z}$, and we rewrite (1.2) as

1.3)
$$y_{n+1} = S(Ah) y_n.$$

The simplest numerical procedure for determining an approximation u_n to y_n, $n = 1, 2, \ldots,$ is furnished by Euler's method,

1.4)
$$u_{n+1} - u_n = -hAu_n, \qquad n = 1, 2, \ldots,$$
$$u_0 = y_0.$$

Using the function $K(z) = 1 - z$ we may rewrite (1.4) as

1.5)
$$u_{n+1} = K(Ah) u_n.$$

By subtracting (1.3) and (1.5), we find that the global error,

$$e_n = u_n - y_n$$

obeys the recurrence relation

1.6) $$e_{n+1} = Ke_n + Ty_n.$$

Here T is the truncation operator $T = K - S$. (1.6) may be solved to yield

$$e_{n+1} = \sum_{j=0}^{n} K^j Ty_{n-j},$$

from which we obtain the bound

1.7) $$||e_n|| \leqslant n \max_{0\leqslant j\leqslant n-1} ||K||^j \max_{0\leqslant j\leqslant n-1} ||Ty_j||.$$

Note that $nh \leqslant \bar{t}$.

If the numerical method is stable, i.e.,

1.8) $$||K|| \leqslant 1$$

and accurate of order p, i.e.,

1.9) $$||Ty|| = O(h^{p+1}),$$

then the bound (1.7) shows that $||e_n|| = O(h^p)$. (Of course for Euler's method $p = 1$, to which case we restrict ourselves.)

To demonstrate (1.9) we note that $||y||$ is bounded for $0 \leqslant t \leqslant \bar{t}$ and we show that $||T|| = O(h^2)$. For the latter we use the spectral representation theorem which tells us that

1.10) $$T(hA) = \sum_{j=1}^{m} T(h\lambda_j)P_j(A).$$

Here we have assumed that the eigenvalues λ_j, $j = 1,\ldots,m$ of A are distinct.The $P_j(z)$, $j=1,\ldots,m$ are the fundamental polynomials on the spectrum of A. (i.e. $P_j(z)$ is the polynomial of minimal degree such that $P_j(\lambda_i) = \delta_{ij}$, $i,j = 1,\ldots,m.$) We have chosen $T(z) = K(z) - S(z)$ to be small at a single point, $z = 0$. Indeed

$$T(z) = O(z^2).$$

This and (1.10) assures us that $||T|| = O(h^2)$. More precisely we have that

1.11 $$||T|| = O(|\lambda_{max}|^2 h^2)$$

where

$$|\lambda_{max}| = \max_{1 \leqslant j \leqslant m} |\lambda_j| .$$

One proceeds similarly, using the spectral representation theorem to deal with the requirment of stability. For Euler's method we obtain stability if

1.12) $$|1-h\lambda_j| \leqslant 1, \qquad j = 1, \ldots, m.$$

For the usual equations one encounters in numerical analysis, $|\lambda_{max}|$ is not too large and (1.12) is achieved with a reasonable restriction on the size of h. In turn (1.11) combined with the bound (1.7) for $||e_n||$ gives us an acceptable error size for a reasonable restriction on the size of h.

1.2. Stiffness

A stiff system of equations is one for which $|\lambda_{max}|$ is enormous, so that either the stability or the error bound or both can only be assured by unreasonable restrictions on h. (i.e., an excessively small h requiring too may steps to solve our problem.) Enormous means enormous relative to a scale which here is $\bar{t}$. Thus an equation with $|\lambda_{max}|$ small may also be stiff if we must solve it for great values of time.

In the literature one usually finds stiffness in a system of differential equations to be defined as the case where the ratio of the eigenvalues of largest and smallest magnitude, respectively is large. This definition is unduly restrictive. Indeed as we may see, a single equation can be stiff. Moreover this usual definition excludes the obviously stiff system corresponding to a high frequency harmonic oscillator, viz

1.13) $$\ddot{y} + \omega^2 y = 0, \qquad \omega^2 \text{ large.}$$

Indeed neither definition is entirely useful in the nonautonomous or nonlinear case. While stiffness is an informal notion we can include most of the problems which are of interest by using the idea of ill conditioning. Suppose we develope the numerical approximation to the solution of a differential equation along the points of a mesh, for example, by means of a relationship of the type (1.5). Then if small changes in u_n in (1.5) result in large changes in u_{n+1}, then the numerical method represented by (1.5) is ill conditioned. To exclude the difficulty wherein this unstable behavior is caused by the numerical method and is not an intrinsic difficulty to the differential equations, we will say that a system of differential equations is stiff if this unstable behavior occurs in the solutions of the differential equations. More formally we have the following definition.

Def. 1.1 - A system of differential equations is said to be stiff on the interval $[0,\bar{t}]$ if there exists a solution of that system a component of which has a variation on that interval which is large compared to $1/\bar{t}$.

The following example shows how treacherous the reliance on eigenvalues to characterize stiffness can be: even in the linear case:

1.14)
$$\dot{y} = A(t)y.$$

where

1.15)
$$A(t) = \begin{pmatrix} \sin \omega t & \cos \omega t \\ \cos \omega t & -\sin \omega t \end{pmatrix}.$$

The eigenvalues of $A(t)$ are ± 1. The matrizant of (1.14) is

1.16)
$$\Phi(t) = B(t)\, \frac{\sinh \sigma}{\sigma} + I \cosh \sigma.$$

Here I is the 2×2 identity matrix,

1.17)
$$\sigma = \sqrt{2}(1 - \cos \omega t)^{1/2}$$

and

1.18)
$$B(t) = \frac{1}{\omega} \begin{pmatrix} 1 - \cos \omega t & \sin \omega t \\ \sin \omega t & \cos \omega t - 1 \end{pmatrix}.$$

Thus

1.19)
$$\Phi(t) = (\cosh \sqrt{2 - 2\cos \omega t})\,(1 + O(\omega^{-1}))I$$

uniformly for $t \in [0,\bar{t}]$.

Thus is spite of the eigenvalues of $A(t)$, the solution of (1.14) varies with frequency ω, a quantity at our disposal.

These lectures will deal with the computational theory of stiff equations.

1.3 Warning

The various methods which are presented and discussed here have been selected because of the ideas and properties of a mathematical nature which the expose. No inference concerning the efficacity of a method should be drawn solely from its inclusion here and inversely.

1.4 References

References will be given at the end of each section. Although there is a large bibliography for our subject, we will not display one. Rather we refer to the references at the end of this chapter. These references are of a general nature and contain large bibliographies.

REFERENCES

[1.1] Bjurel, G., Dahlquist, G., Lindberg, B., Linde, S., and Oden, L., "Survey of Stiff Ordinary Differential Equations", Report NA 70.11, Department of Information Processing Computer Science, the Royal Institute of Technology, Stockholm, Sweden.

[1.2] Liniger, W., "Lecture Notes on Stiff Differential Equations" A course given at the University of Lausanne, (1973-74).

[1.3] "Stiff Differential Equations" Proceedings of the IBM Research Symposia Series, Edited by R.A. Willoughby, Plenum Press (1974).

§ 2. REVIEW OF THE CLASSICAL LINEAR MULTISTEP THEORY

2.1 The Initial Value Problem

We begin by considering the nonlinear initial value problem

$$\dot{x} = f(t,x), \tag{2.1}$$
$$x(a) = s,$$

where x, f and $s \in C_m$ (i.e. are m-tuples of complex numbers). We seek a solution to (2.1) on the interval I;

$$I = \{t \mid a \leqslant t \leqslant b; \; -\infty < a < b < \infty\}.$$

Def. 2.1: f is said to be an $\mathbb{L}$-function if for all $t \in I$ and x and $y \in C_m$, there exists a constant L such that

$$||f(t,x) - f(t,y)|| \leqslant L||x-y||.$$

Here $||x||$ denotes any norm of $x=(x^1,\dots,x^m)$. For example $||x|| = \sum_{i=1}^{m} |x^i|$.

We may now state the following existence and uniqueness theorem for the problem (2.1).

Theorem 2.1: If f is continuous in t for $t \in I$ and if f is an $\mathbb{L}$-function, the problem (2.1) has one and only one solution in I.

2.2 Linear Multistep Operators

The best known numerical methods used to generate approximate solutions are based on the linear multistep operator $\mathcal{L}$ given by

$$\mathcal{L} \equiv \sum_{j=0}^{k} \alpha_j E^j - h \sum_{j=0}^{k} \beta_j E^j \frac{d}{dt}.$$

Here E is the shift operator

$$Ex(t) = x(t+h)$$

and the α_j and β_j are given scalars with $(\alpha_0^2 + \beta_0^2) \cdot \alpha_k \neq 0$. k is called the number of steps of $\mathcal{L}$.

Def. 2.2: $\mathcal{L}$ is said to have degree of precision p, if $\mathcal{L}$ annihilates all monomials t^n, $n \leqslant p$ and p is maximal with respect to this property.

Now let us suppose that $x(t) \in C^\infty$ and let us express $\mathcal{L}\,x(t)$ in the form of a Taylor series.

$$2.2)\qquad \mathcal{L}\,x(t) = \sum_{\nu=0}^{\infty} c_\nu h^\nu x^{(\nu)}(t).$$

An alternate definition of the degree of precision of $\mathcal{L}$ is given in the following definition.

Def. 2.3: $\mathcal{L}$ is said to have degree of precision p if the coefficients α_j and β_j may be chosen so that $c_\nu = 0,\ \nu = 0,1,\dots,p$ and p is maximal with respect to this property. Clearly $p \leqslant 2k$.

2.3 Approximate Solutions

To construct an approximate solution to (2.1), we begin by introducing the mesh $t_n = a + nh,\ h > 0,\ n \in J_h \equiv \{0,1,\dots,n_{max}\}$. J_h is the set of integers such that $t_{n+i} \in I,\ i = 0,1,\dots,k$.

An approximate solution is a sequence $\{x_n\},\ n \in J_h$ where x_n is considered as an approximation to $x(t_n), n \in J_h$. We define an approximate solution by means of $\mathcal{L}$ through the linear multistep method,

$$2.3)\qquad F(x_n) \equiv \sum_{j=0}^{k} \alpha_j x_{n+j} - h \sum_{j=0}^{\infty} \beta_j f_{n+j} = 0, \qquad n \in J_h.$$

Here $f_n \equiv f(t_n, x_n)$.

The linear multistep method is said to be explicit if $\beta_k = 0$. Otherwise it is implicit. Each $x_{n+k},\ n \in J_h$ is obtained from (2.3) through transposing and solving an equation of the form

$$\alpha_k x_{n+k} - h\beta_k f(t_{n+k}, x_{n+k}) = constant.$$

In the explicit case solving this equation requires only division by α_k.

The linear multistep formula allows the step by step determination of $x_n, n \in J_h$, provided that the values of $x_0, \dots, x_{k-1}$ are known. These so called starting values are determined by some independent procedure which may be called the starting procedure. As a notation for the starting procedure we will write

$$2.4)\qquad x_m = S_m(h), \qquad m = 0,1,\dots,k-1.$$

The following two definitions are basic.

Def. 2.4: The starting procedure is said to be bounded if there exists aconstan $M>0$, such that $||S_m(h)||\leq M$ for all sufficiently small h.

Def. 2.5: The starting procedure is said to be compatible if

$$\lim_{h\to 0} S_m(h) = s, \qquad m = 0, 1, \ldots, k-1.$$

Let (C.f.(2.1))

$$h_0 = \alpha_k \ (\beta_k \mathbb{L})^{-1}. \tag{2.5}$$

The existence and uniqueness of the numerical procedure is the subject of the following theorem.

Theorem 2.2: A linear multistep formula has one and only one solution x_n, $n\in J_h$ for all starting procedures $S_m(h)$ if $0\leq h<h_0$.

2.4 Examples of Linear Multistep Methods

The following are some of the well known linear multistep methods:

i) Adams' method

$$x_{n+k} - x_{n+k-1} - h \sum_{j=0}^{k} \beta_j f_{n+j} = 0$$

$\beta_k \neq 0$: Adams-Moulton, $k=1$: Trapezoidal formula

$\beta_k = 0$: Adams-Bashforth, $k=1$: Euler's formula

ii) Nystrom's method

$$x_{n+k} - x_{n+k-2} - h \sum_{j=0}^{k-1} \beta_j f_{n+j}$$

$k = 2$: mid-point formula

iii) Method of Newton-Cotes

$$x_{n+k} - x_n - h \sum_{j=0}^{k} \beta_j f_{n+j} = 0$$

$k = 2$: Simpson's formula

iv) Backward differentiation formula

$$\sum_{j=0}^{k} \alpha_j x_{n+j} - h\beta_k f_{n+k} = 0$$

2.5 Stability, Constistency and Convergence

A linear multistep formula is consistent if its order $p \geqslant 1$. This is explicitly characterized in the following definition.

Def. 2.5: A linear multistep method is said to be consistent if

$$||F(x(t_n))|| = O(h), \qquad n \in J_h ,$$

where $x(t)$ is any solution of $x' = f(t,x)$. (C.f.(2.3).)

We now introduce the ρ and σ polynomials.

2.5)
$$\rho(\omega) = \sum_{j=0}^{k} \alpha_j \omega^j ,$$
$$\sigma(\omega) = \sum_{j=0}^{k} \beta_j \omega^j$$

and we suppose that $(\rho|\sigma) = 1$. We now easily conclude the following theorem:

Theorem 2.3: A linear multistep method is consistent if and only if

$$\mathcal{L}(1) = \rho(1) = 0$$

and

$$\mathcal{L}(t) = h(\rho'(1) - \sigma(1)) = 0$$

The stability of a linear multistep method is characterized in the following definition.

Def. 2.6: Let M be a constant. A linear multistep formula is said to be stable if

$$\max_{n \in J_h} ||x_n|| \leqslant M$$

uniformly in h, $h \in (0, h_0]$ for all bounded starting procedures and for all $f \in \mathbb{L}$.

The study of stability makes use of the root condition given in the following definition.

Def. 2.7: A polynomial $p(\omega)$ is said to satisfy the root condition if all of its roots lie in the closed unit disc while those on the boundary of the disc are simple.

With this we have the following theorem.

Theorem 2.4: A linear multistep method is stable if and only if $\rho(\omega)$ obeys the root condition.

The global or cumulative error of the linear multistep method is

$$e_n = x_n - x(t_n), \qquad n \in J_h. \tag{2.6}$$

A convergent method is characterized in the following definition.

Def. 2.8: A linear multistep method is convergent if for all $f \in L$ and all compatible starting procedures, we have

$$\lim_{h \to 0} \max_{n \in J_h} ||e_n|| = 0.$$

Finally, the main theorem of this subject is the following.

Theorem 2.5: A linear multistep method is convergent if and only if it is stable and consistent.

REFERENCES

[2.1] Henrici,P., "Discrete Variable Methods in Ordinary Differential Equations", Wiley, New York (1962).

§ 3. THE METHOD OF ABSOLUTE STABILITY

3.1 Stiff Systems

Consider the linear case

$$\dot{x} = Ax, \qquad t \in (0,\bar{t}], \tag{3.1}$$

where A is an $m \times m$ constant matrix. Let λ_j, $j = 1,\dots,m$ be the eigenvalue of A. The following definition characterizes a stiff system.

Def. 3.1: The linear system (3.1) is said to be stiff if

$$\max_{1\leqslant j\leqslant m} |\lambda_j \bar{t}| >> 1.$$

As we may see this is not a precisely defined notion.

Remark 3.1: A system consisting of a single equation may be stiff.

To motivate the first method for dealing with stiff systems, consider the case $m = 2$ with $\lambda_2 << \lambda_1 < 0$ and with the solution

$$F(t) = e^{\lambda_1 t} + e^{\lambda_m t}.$$

As t increases from zero there is a transitory stage during which $F(t)$ varies extremely rapidly. After a time of the order λ_m^{-1} the component $e^{\lambda_m t}$ of $F(t)$ becomes negligible and a new permanent stage developes. To determine a numerical approximation to $F(t)$ in the transitory stage we would use a mesh increment, h_1, such that $|h_1\lambda_m|$ is small. For the permanent stage we would like to use a much larger mesh increment h_2 and one such that

$$|\lambda_1 h_2| << 1 << |\lambda_m h_2|.$$

In this case the numerical theory is applicable for the component $e^{\lambda_1 t}$. We do not expect the same to be true for the other component. However, if the method is stable no matter how large $|\lambda_m h_2|$ is, we may expect the component $e^{\lambda_m t}$ to remain negligible. This technique calls for methods of an extraordinary stable character, indeed it calls for methods with a form of absolute stability.

We give three criticisms of the idea.

i) Getting through the transitory stage requires a number of steps proportional to λ_m^{-1} and this may not be acceptable.

ii) If λ_m is large in magnitude because it has a large imaginary part, the transitory stage is permanent.

iii) Absolutely stable methods of simple types are rare. (This will be seen presently.)

For the time being we exclude eigenvalues with a large imaginary part and we will return to this type of problem in §§12.14.

3.3 *A*-stability

Now we formalize the celebrated notion of absolute stability called *A*-stability.

Def. 3,2: A linear multistep method is *A*-stable if all solutions of the difference equation generated by the application of this method to the test equation (scalar)

3.2) $\dot{x} = \lambda x, \qquad \lambda$ a complex constant,

tend to zero as $n \to \infty$ for all λ with $Re\ \lambda < 0$ and for all $h > 0$ fixed.

To determine which linear multistep methods are *A*-stable, we note that when the test equation (3.2) is inserted into the linear multistep formula, a linear difference equation results:

3.3) $$\sum_{j=0}^{k} (\alpha_j - q\beta_j) y_{n-j} = 0, \qquad q = \lambda h$$

The characteristic equation corresponding to (3.3) is

3.4) $$X(\omega; q) \equiv \rho(\omega) - q\sigma(\omega) = 0.$$

(cf. (2.5)),

X defines a k-valued mapping of q into ω. The inverse of this mapping,

3.5) $$q(\omega) = \rho(\omega)/\sigma(\omega),$$

defines a single valued mapping of ω into q.

With these observations we may state the following proposition:

Proposition 3.1: Let ω_i, $i=1,\ldots,k$ be the roots of $X(\omega; q)=0$. Then the following three statements are equivalent

a) a linear multistep method is *A*-stable

3.6) b) $Re\ q < 0 \Rightarrow |\omega_i| < 1, \qquad i=1,\ldots,k$

c) $|\omega| > 1 \Rightarrow Re\ q(\omega) > 0.$

Using this proposition we may state and prove the following lemma.

Lemma 3.1: The linear multistep method $\sum_{j=0}^{k} (\alpha_j - q\beta_j) x_{n+j} = 0$ is

A-stable if and only if

i) The roots σ_i of $\sigma(\omega)$ satisfy $|\sigma_i| \leqslant 1$, $i=1,\ldots,k$

and

ii) $Re\ \rho(\omega)/\sigma(\omega) \geqslant 0$, for all ω in $W \equiv \{\omega/|\omega|>1\}$.

Proof A) We first show that A-stability implies (i) and (ii).

That A-stability implies (ii) is obvious. We proceed to verify (i) Since $(\rho|\sigma)=1$, $\rho(\sigma_i)\neq 0$. Thus under the mapping of $\omega \rightarrow q$ generated by $X(\omega;q)=0$; each σ_i is mapped into the north pole of the q-Riemann sphere, the latter being a point on the imaginary axis of that sphere. Similarly each neighborhood of σ_i is mapped onto a neighborhood of the north pole. Now every neighborhood of the north pole contains values of q such that $Re\ q<0$.

Suppose (ii) were not true. Then one of the roots σ_i is such that $|\sigma_i|>1$. Then there exists a sufficiently small neighborhood of this σ_i contained in W. (c f. Figure 3.1).

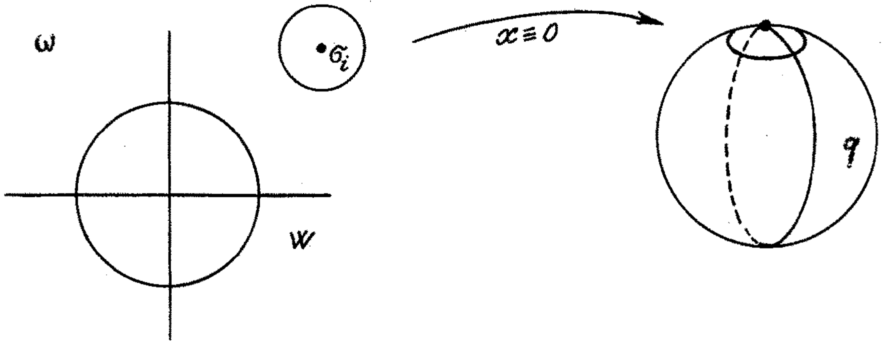

Figure 3.1

Thus $X=0$ would have solutions in W for values of q with $Re\ q<0$. This contradicts the A-stability, completing part (A) of this proof.

B) (ii) implies (3.6c) in W. Thus there remains only to verify (3.6c) for $|\omega|=1$. Then let ω_0 be such that $|\omega_0|=1$ and consider two cases; case (a) $\sigma(\omega_0)\neq 0$ and case (b) $\sigma(\omega_0)=0$.

Case a: $\sigma(\omega_0)\neq 0$

In this case $q(\omega)$ is analytic in a neighborhood of ω_0. Suppose to the contrary that $Re\ q(\omega_0)<0$. Then a sufficiently small neighborhood of ω_0 will be mapped onto a neighborhood of $q(\omega_0)$, the latter neighborhood being entirely contained in $Re\ q<0$. (c f. Figure 3.2).

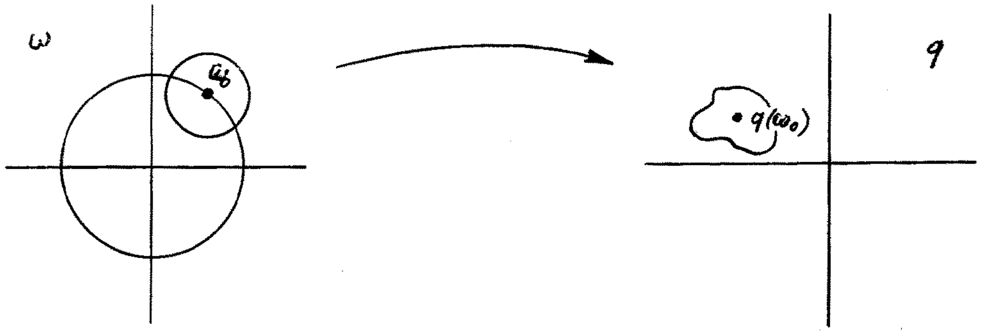

Figure 3.2

This neighborhood of ω_0 contains points ω of W whose image, $q(\omega)$ satisfies $Re\ q < 0$. This contradicts (ii) completing the proof of case (a).

Case b: $\sigma(\omega_0)=0$

In this case $q(\omega_0)$ is the north pole of the q-Riemann sphere, a point on the imaginary axis. Thus (3.6c) is obviously satisfied. This completes the proof of case b and the lemma.

The following proposition is interesting because it increases the similarity of conditions on $\sigma(\omega_0)$ for A-stability to the root condition for $\rho(\omega)$ for ordinary stability of the linear multistep method.

Proposition 3.2: If a root ω_0 of $\sigma(\omega)$ has magnitude unity and is not a simple root, then the linear multistep method is not A-stable.

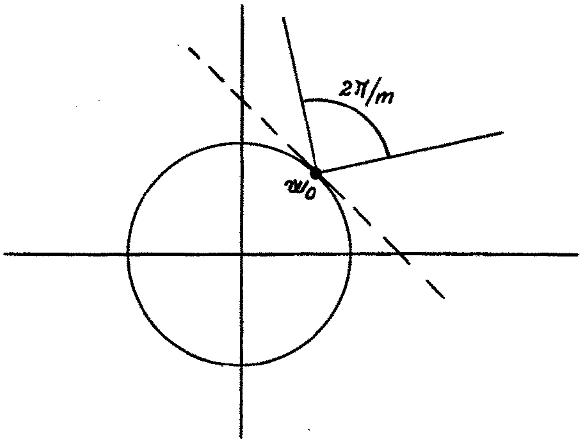

Figure 3.3

Proof. Let $m \geqslant 2$ be the multiplicity of the root ω_0. Then $q(\omega)=$ $=const(\omega-\omega_0)^m(1+o(1))$. Thus the sectors of a neighborhood of ω_0 which are of angle $2\pi/m$ (at most a half plane) are mapped onto a neighborhood of the north pole of the q-sphere. Since the sectors are at most a half plane, we may choose one which lies entirely in W (except of course for the vertex ω_0 of this sec tor). (c f. Figure 3.3).
Thus there exists points of W whose images satisfy $Re\ q < 0$.
Thus the corresponding linear multistep method is not A-stable.

3.3 Examples of A-stable Methods

We now give several examples of A-stable methods.

1. The trapezoidal formula:

$$x_{n+1} - x_n - \frac{1}{2} h(f_{n+1}+f_n) = 0.$$

$$\rho(\omega) = \omega-1, \qquad \sigma(\omega) = \frac{1}{2}(\omega+1).$$

$$Re\ q(\omega) = \frac{|\omega|^2-1}{|\omega+1|^2}.$$

Thus $Re\ q(\omega) > 0$ in W and the root of σ on $|\omega|=1$ is simple.

2. The backward Euler formula:

$$x_{n+1} - x_n - hf_{n+1} = 0.$$

$$\rho(\omega) = \omega-1, \qquad \sigma(\omega) = \omega.$$

$$Re\ q(\omega) = \frac{|\omega|^2 - Re\ \omega}{|\omega|^2} > 0, \qquad |\omega| > 1.$$

3.
$$x_{n+k} - x_n - \frac{1}{2} hk(f_{n+k}+f_n) = 0.$$

$$\rho(\omega) = \omega^k - 1, \qquad \sigma(\omega) = \frac{1}{2} k(\omega^k+1).$$

The roots of $\sigma(\omega)$ are the k-th roots of unity.

$$Re\ q(\omega) = \frac{1}{2}\, k\, \frac{|\omega|^{2k} - 1}{|\omega^k + 1|^2} > 0, \qquad |\omega| > 1.$$

Note further that $\rho(1)=0$, $\rho'(1)=\sigma(1)=k$, implying the consistency of this method. This example shows the occurrence of linear multistep methods which are consistent and A-stable for any k (i.e. any number of steps).

3.4 Properties of A-stable Methods

Achieving A-stability is costly in terms of the restrictions this property imposes on the class of linear multistep methods. The first restriction is the loss of explicit schemes which requires a greater amount of computation in each step of the method. This restriction is characterized by the following theorem.

Theorem 3.1: An explicit linear multistep method can not be A-stable.

Proof: Assume to the contrary that the method is both explicit and A-stable. Then $\beta_k=0$ and $q(\omega)=\rho(\omega)/\sigma(\omega)$ has a pole at the point, ω_0, at infinity on the ω-sphere. But ω_0 as well as neighborhood of ω_0 lie in W. The image of such a neighborhood under the mapping $q=q(\omega)$ is a neighborhood of the point, q_0, at infinity on the q-sphere. Such a neighborhood contains points for which $Re\ q<0$. This contradicts (3.6c) completing the proof of the theorem.

If a linear multistep method is of order p, we have from (2.3) that

$$\mathcal{L}(x(t)) = c_{p+1} h^{p+1} x^{(p+1)}(t)(1+O(h)). \tag{3.7}$$

If $p \geqslant 1$, $\rho(1)=0$ and since $(\rho|\sigma)=1$ then $\sigma(1)\neq 0$. Now consider the following definition which introduces the so-called error constant c^*, which serves as a measure of quality of linear multistep methods of the same order.

Def. 3.2: $c^*=-c_{p+1}/\sigma(1)$ is called the error constant of a linear multistep method of order $p \geqslant 1$.

Remark 3.2:

$$c^* = \lim_{\omega\to 1} [\log\omega - \rho(\omega)/\sigma(\omega)]/(\omega-1)^{p+1}. \tag{3.8}$$

The following theorem characterizes the key restriction on A-stable methods.

Theorem 3.2 The order p of an A-stable linear multistep method can not exceed 2. The trapezoidal formula is the A-stable method of order 2 which gives the smallest error constant, $c^*=1/12$.

Proof: The proof begins with a side calculation.

Let $z=\dfrac{\omega+1}{\omega-1}$, the well known 1-1 Moebious transformation carrying $\omega=1$ into the point z at infinity. Let the transformation Γ be defined by

$$\Gamma f(\omega)=2^{-\frac{1}{2}k}(\omega-1)^k f\left(\frac{\omega+1}{\omega-1}\right)$$

and let

$$r(z)=\Gamma\rho(\omega), \qquad s(z)=\Gamma\sigma(\omega).$$

Now apply Γ to (3.8). We get

$$\log\frac{z+1}{z-1}-\frac{r(z)}{s(z)}=c^*\left(\frac{2}{z}\right)^{p+1}(1+o(1)), \qquad z\to\infty.$$

Since $\log\dfrac{z+1}{z-1}=2z^{-1}+\dfrac{2}{3}z^{-3}+O(z^{-4})$, this becomes

3.9) $$\frac{r(z)}{s(z)}=2z^{-1}+\left(\frac{2}{3}-8c'\right)z^{-3}+O(z^{-4}),$$

where

$$c'=\begin{cases} c^*, & p=2 \\ 0, & p\geqslant 3. \end{cases}$$

Thus we may note that the coefficient of z^{-3} in (3.9) is strictly positive if $p\geqslant 3$.

Next we translate the conditions (i) and (ii) of Lemma 3.1. By using properties of the Moebious transformation, we see that this lemma asserts the equivalence of A-stability of a linear multistep method and

i) The roots s_i of $s(z)$ satisfy $Re\ s_i \leqslant 0,\quad i=1,\dots,k$

3.10)

ii) $Re\,\dfrac{r(z)}{s(z)} \geqslant 0$ for all z in $Re\ z \geqslant 0.$

Next we make use of the following variant of the Riesz-Herglotz theorem (c.f. [3.1], p.152):

Theorem: An analytic function $\phi(z)$ which satisfies

a) $\sup\limits_{0<x<\infty} |x\phi(x)| < \infty$

b) $\phi(z)$ regular in $Re\ z > 0$

c) $Re\ \phi(z) \geqslant 0$ in $Re\ z > 0$

may be represented as follows:

$$\phi(z) = \int_{-\infty}^{\infty} \frac{d\omega(t)}{z-it},$$

where $\omega(t)$ is a bounded nondecreasing function.

Now we will show that $z\phi(z) = z\,\dfrac{r(z)}{s(z)}$ is bounded for all $x\epsilon[0,\infty]$. We note first that (3.9) implies that $x\,r(x)/s(x)$ is bounded as $x\to\infty$. By hypothesis the linear multistep method is A-stable. Then from proposition 3.2, $\sigma(\omega)$ has a zero of order at most unity at $\omega=-1$. The same then is true for $s(z)$ at $z=0$. Then $xr(x)/s(x)$ is bounded at $x=0$. Using (3.10) (i) we may conclude that $xr(x)/s(x)$ is bounded for all x on the positive real axis. Thus $z\phi(z)$ is indeed bounded as claimed.

Now (3.10) (i) and (ii) imply that $\phi(z)$ is regular and that $Re\ \phi(z) \geqslant 0$ in the half plane $Re\ z > 0.$

Thus the hypotheses a), b) and c) of the cited theorem are verified and for $x > 0$, we have

$$x\,\frac{r(x)}{s(x)} = \int_{-\infty}^{\infty} \frac{x}{x-it}\,d\omega(t) = \int_{-\infty}^{\infty} \frac{x^2}{x^2+t^2}\,d\omega(t).$$

Since $\dfrac{d}{dx}\,\dfrac{x^2}{x^2+t^2} = \dfrac{2xt^2}{(x^2+t^2)^2} \geqslant 0$ for $x \geqslant 0$, we may conclude from this representation that

3.11) $$\frac{d}{dx}\left[x\,\frac{r(x)}{s(x)}\right] \geqslant 0.$$

Next from (3.9) we may conclude that

3.12) $$\frac{d}{dx\cdot}\left[x\,\frac{r(x)}{s(x)}\right] = -2\left(\frac{2}{3} - 8c'\right)x^{-3}(1+o(1)), \qquad x \to \infty.$$

Comparing (3.11) and (3.12) we deduce that

3.13) $$\frac{2}{3} - 8c' \leqslant 0.$$

If $p \geqslant 3$, $c' = 0$ and (3.13) is impossible. This demonstrates the first assertion of the theorem.

If $p=2$, $\frac{2}{3} - 8c' \leqslant 0$ or $c^* \geqslant 1/12$. For the trapezoidal formula, $\rho(\omega) = \omega - 1$, $\sigma(\omega) = \frac{1}{2}(\omega+1)$, $r(z) = \sqrt{2}$, $s(z) = z/\sqrt{2}$ so that

$$\frac{r(z)}{s(z)} = \frac{2}{z}.$$

Comparing this with (3.9), we deduce that $\frac{2}{3} - 8c^* = 0$ or $c^* = \frac{1}{12}$. This demonstrates the second assertion of the theorem and completes its proof.

3.5 A Sufficient Condition for A-stability

Condition (ii) of Lemma 3.1 requires the verification of a property of $q(\omega)$ for all ω in W. A less stringent requirement furnishes the following sufficient condition for A-stability.

Theorem 3.3: If

i) the roots σ_i of $\sigma(\omega)$ satisfy $|\sigma_i| < 1$, $i=1,\dots,k$

and

ii) $u(\omega) \equiv Re\ q(\omega) \geqslant 0$ on the unit circle,

then the linear multistep method is A-stable.

Proof i) implies that $q(\omega)$ is analytic in $\overline{W}$ and in particular at $\omega = \infty$. Then $u(\omega)$ is harmonic in $\overline{W}$ and from the minimum principle

$$u(\omega) \geqslant \min_{|\omega|=1} u(\omega)$$

for all $\omega \in \overline{W}$. Then (ii) implies that $u(\omega) \geqslant 0$ for all $\omega \in \overline{W}$. Then (3.6) (c) implies that the method is A-stable, completing the proof of the theorem.

3.6 Applications

As an application of theorem 3.3 consider the formula

3.14) $$x_{n+1} - x_n - h[(1-a)\dot{x}_{n+1} + a\dot{x}_n] = 0,$$

for which $p \geqslant 1$ for all real values of the parameter a. For $a = 1,\ 1/2,\ 0$ respectively, this formula becomes the Euler formula, the trapezoidal formula and the backward Euler formula, respectively. In any case we have

$$\sigma(\omega) = (1-a)\omega + a.$$

The root $\sigma_1 = -a(1-a)^{-1}$ of $\sigma(\omega)$ is less than unity in magnitude if and only if $a < \frac{1}{2}$. A calculation shows that

$$u(e^{i\theta}) = |\sigma(e^{i\theta})|^{-2} P(e^{i\theta})$$

where

$$P(e^{i\theta}) = (1-2a)(1-\cos\theta).$$

$P(e^{i\theta}) \geqslant 0$ if and only if $a \leqslant \frac{1}{2}$. Thus (3.14) is A-stable if $a < \frac{1}{2}$.

Note that the trapezoidal formula (which is A-stable) fails to satisfy the sufficient condition of theorem 3.3.

A second application is the following formula

$$(-1-a+b)x_n + 2(a-b)x_{n+1} + (1-a+b)x_{n+2} - h[a\dot{x}_n + (2-a-b)\dot{x}_{n+1} + b\dot{x}_{n+2}] = 0.$$

For this formula $p \geqslant 2$ for all real values of the parameters a and b. One may show that for this formula, hypotheses i) and

of theorem 3.3
ii) are equivalent to the following two inequalities:

$$b - a > 0$$

$$-1 + a + b > 0.$$

REFERENCES

[3.1] Achieser, N.I. and Glassman, I.M., "Theorie der Linearen Operatoren in Hilbert-Raum" Akademie-Verlag Berlin (1954)

[3.2] Dahlquist, G.G., "A special Stability Problem for Linear Multistep Methods", BIT 3 (1963) pp.27-43.

§ 4. NOTIONS OF DIMINISHED ABSOLUTE STABILITY

The family of linear multistep methods is so desirable because of its simple form for computation and analysis that the limitations imposed on this family by A-stability made a great impact. In order to attempt to save the family for the solution of stiff differential equations a sequence of weakened forms of absolute stability were invented in order.

We will look at one of these, $A(\alpha)$-stability and see by just how much it improves things. We start with the following definition.

4.1 $A(\alpha)$-stability

Def. 4.1: A linear multistep method is $A(\alpha)$-stable, $0 < \alpha < \frac{\pi}{2}$, if all solutions of the difference equation arising through the application of this method to the test equation, tend to zero as $n \to \infty$ for each fixed mesh increment $h > 0$ and for all $\lambda \neq 0$ where

$$q \equiv \lambda h \in S_\alpha \equiv \{q \mid |arg(-q)| < \alpha,\ q \neq 0\}.$$

We may note the following remarks:

Remark 4.1: Let ω_i, $i=1,\ldots,k$ be the roots of the characteristic equation, $X=0$ corresponding to the difference equation arising from the application of the test equation (c. f. Def. 3.2) to the linear multistep method. Then the corresponding linear multistep method is $A(\alpha)$ stable if $q \in S_\alpha$ implies that the $|\omega_i| < 1$, $i=1,\dots,k$.

Remark 4.2: a) $A(\alpha)$-stability $\Longrightarrow A(\beta)$-stability for $0<\beta<\alpha$.

b) A stability is equivalent to $A\left(\frac{\pi}{2}\right)$-stability.

The case $\alpha=0$ is described in the following definition.

Def. 4.2: A linear multistep method is $A(0)$-stable if it is $A(\alpha)$-stable for all sufficiently small $\alpha>0$.

The following lemma is the analogue of lemma 3.1.

Lemma 4.1: The linear multistep method $\sum_{j=0}^{k} (\alpha_j - q\beta_j)x_{n+j}$ is $A(\alpha)$-stable, $\alpha>0$, if and only if

i) The roots s_i of $s(z)$ satisfy $Re\ s_i \leqslant 0$, $i=1,\dots,k$

ii) $r(z)/s(z)$ is in the compliment of S_α for all z with $Re\ z>0$. (c.f. Theorem 3.2).

For the case of $A(0)$-stability we have the following necessary condition.

Lemma 4.2: If a linear multistep method is $A(0)$-stable then $\{a_\nu \geqslant 0$ or $a_\nu \leqslant 0\}$ and $\{b_\nu \geqslant 0$ or $b_\nu \leqslant 0\}$, $\nu=1,\dots,k$.

4.2 Properties of $A(\alpha)$-stable Methods

As usual we will suppose that $(\rho|\sigma)=1$ and that $p \geqslant 1$ (so that the methods are consistent).

The first result which shows that we do not recover the explicit methods is the subject of the following theorem.

Theorem 4.1: An explicit linear multistep method can not be $A(0)$-stable.

The order restriction is weakened at least somewhat as the following two theorems show.

Theorem 4.2: The trapezoidal formula is the only $A(0)$-stable linear multistep method with $p \geqslant k+1$.

Theorem 4.3: There exist $A(\alpha)$-stable linear multistep methods, $0 \leqslant \alpha < \frac{\pi}{2}$ for $k=p=3$ and $k=p=4$.

We forego developing the proofs of Lemmas 4.1 and 4.2 and of Theorems 4.1, 4.2, and 4.3 since the proofs are generally analogous to the proofs in § 3.

REFERENCES

[4.1] Widlund, O. "A Note on Unconditionally Stable Linear Multistep Methods", BIT 7 (1967) 65-70.

§ 5. THE METHOD OF JAIN

With the limited success of removing the restrictions on the class of linear multistep methods imposed by the various notions of absolute stability, there remains the possibility of retaining the strongest of these notions, but to leave the class of linear multistep methods. An example of such a method which we will now describe is due to Jain.

5.1 Description of the Method

We start with the initial value problem

5.1) $$y'(t) = f(t,y), \qquad t \in (a,b];$$
$$y(a) = s.$$

Here y and f are m-vectors.

We consider the function

$$y'(t) + Py(t)$$

where P is an $m \times m$ matrix to be specified, and we perform the following three steps.

i) Approximate $y'(t)+Py(t)$ by a polynomial of interpolation, $Q(t)$, which uses Hermite interpolatory data at the points t_{n-i}, $i=0,1,\ldots,n-1$.

ii) Integrate the differential equations $y'+Py=Q$ from t_n to t_{n+1}.

iii) Choose P as an approximation to $\left(\frac{\partial f}{\partial y}\right)_n \equiv - \frac{\partial f(t_n,y(t_n))}{\partial y}$.

Step (i) results in

$$5.2) \qquad y'(t)+Py(t) = \sum_{i=1}^{n} h_i(t)(f_i+Py_i) + \sum_{i=1}^{n} \overline{h}_i(t)(f_i'+Pf_i)+TE.$$

Here h_i and $\overline{h}_i$ are the fundamental Hermite interpolation polynomials of the first and second kind, respectively, corresponding to the points t_{n-i}; $i=0,1,\ldots,n-1$. Also

$$f_i = f(t_i,y_i), \quad f_i' = f'(t_i,y_i), \qquad i=0,1,\ldots,n-1$$

and

$$TE = \frac{1}{(2n)!} F^{(2n)}(\xi)\pi^2(t), \qquad a<\xi<b,$$

where

$$F(t) = f(t) + Py(t).$$

Now we apply step ii) (i.e. integrate (5.2)). We find

$$5.3) \qquad y_{n+1} = e^{-Ph}y_n + e^{-Pt_{n+1}}\left[\sum_{i=1}^{n} H_iF_i+\overline{H}_iF_i'\right] + R_n \ ,$$

where

$$H_i = \int_{t_n}^{t_{n+1}} e^{Pt}h_i(t)dt \ ,$$

$$\overline{H}_i = \int_{t_n}^{t_{n+1}} e^{Pt}\overline{h}_i(t)dt$$

and

$$R_n = \frac{e^{-Pt_{n+1}}}{2n!}\int_{t_n}^{t_{n+1}} e^{Pt}F^{(2n)}(\xi)\pi^2(t)d\xi \ .$$

As far as step (iii) is concerned and in the case where $m=1$, a natural choice for P is

$$P = -\frac{f_n-f(t_n,y_{n-1})}{y_n-y_{n-1}}.$$

In the case $m>1$, the choices for P depend upon the relative difficulty in evaluating e^{Ph} A simple choice is the diagonal matrix whose ii-th entry is

$$P_{ii} = -\frac{f_n^i - f^i(t_n, y_n^1, \dots, y_n^{i-1}, y_{n-1}^i, y_n^{i+1}, \dots, y_n^m)}{y_n^i - y_{n-1}^i}$$

As we see in (5.3), the method is far from being a linear multistep method.

5.2 Properties of the Method

The properties of this method are given by the following theorem.

Theorem 5.1: The method of Jain is A-stable and of order $2n$.

Proof: Let $f(t,y)=\lambda y$ where λ is a complex constant with $Re\ \lambda < 0$ (i.e., the case of the test equation). Then $P=-\lambda$ and for each i,

$$F_i = f_i + Py_i = \lambda y_i - \lambda y_i = 0$$

and

$$F_i' = f_i' + Pf_i = \lambda y_i' - \lambda y_i' = 0.$$

Then (5.3) becomes

$$y_{n+1} = e^{\lambda h} y_n$$

Then since $Re\ \lambda < 0$, $\lim_{n\to\infty} y_n = 0$ for each fixed $h > 0$.

This demonstrates the A-stability of the method.

Now insert $s=(t-t_n)/h$ into (5.2). It becomes

$$y_{n+1} = e^{-Ph} y_n + he^{-Ph}\left[\sum_{i=1}^{n} (H_iF_i + \bar{H}_iF_i')\right] + R_n . \tag*{5.3)}$$

Here

$$H_i = \int_0^1 e^{Phs} k_i(s)ds, \qquad k_i(s) = h_i(hs+t_i)$$

$$\bar{H}_i = \int_0^1 e^{Phs} \bar{k}_i(s)ds, \qquad \bar{k}_i(s) = \bar{h}_i(hs+t_i), \qquad i=1,\dots,n$$

and

$$R_n = \frac{h^{2n+1}}{2n!} e^{-Ph} \int_0^1 e^{Phs} F^{(2n)}(\xi)\pi^2(s)ds$$

$$= \frac{h^{2n+1}}{2n!} e^{-Ph} \int_0^1 F^{(2n)}(\xi)\pi^2(s)ds + O(h^{2n+2})$$

$$= \frac{h^{2n+1}}{2n!} e^{Ph} F^{(2n)}(\bar{\xi}) \int_0^1 \pi^2(s)ds + O(h^{2n+2})$$

by the second mean value theorem. Then

$$R_n = h^{2n+1} e^{-Ph} F^{(2n)}(\bar{\xi})\Lambda_n + O(h^{2n+2})$$

where

$$\Lambda_n = \frac{1}{2n!}\int_0^1 \pi^2(s)ds.$$

Thus the method is of order $2n$ and the theorem is proved.

5.3 Some Special Cases

The integrals for the determination of the $H_i, \bar{H}_i$ and R_n are of the form

$$I_n = \int_0^1 e^{Phs}\left(\sum_{i=1}^{N} A_i s^i\right)ds$$

where $N=N(n)$ is an integer. In addition

$$H_i = \sum_{r=1}^{2n} a_r (Ph)^{-r} e^{Ph} + h \sum_{r=1}^{2n} b_r (Ph)^{-r}$$

$$\bar{H}_i = \sum_{i=1}^{2n} \alpha_r (Ph)^{-r} e^{Ph} + h \sum_{i=1}^{2n} \beta_r (Ph)^{-r}.$$

In the sample case, $n=1$, we find

$$h_1(t) = 1 \qquad \bar{h}_1(t) = t - t_1$$

$$k_1(s) = 1 \qquad \bar{k}_1(s) = s \qquad \pi(s) = s$$

$$H_1 = \int_0^1 e^{Phs}\,ds = (Ph)^{-1}(e^{Ph} - 1)$$

$$\bar{H}_1 = \int_0^1 s e^{Phs}\,ds = [(Ph)^{-1} - (Ph)^{-2}]\,e^{Ph} - (Ph)^{-2}$$

$$\Lambda_1 = \frac{1}{2}\int_0^1 s^2 ds = \frac{1}{6}$$

$$a_1 = 1, \qquad a_2 = 0, \qquad b_1 = -1, \qquad b_2 = 0$$

$$\alpha_1 = 1, \qquad \alpha_2 = -1, \qquad \beta_1 = 0, \qquad \beta_2 = 1.$$

5.4 Criticism

While Jain's method is A-stable and of higher accuracy, it is costly to use.

REFERENCES

[5.1] Jain, R.K., "Some *A*-stable Methods for Stiff Ordinary Differential Equations", Math Comp *26* (1972) pp.71-78.

[5.2] See also references [12.1] and [12.2].

§ 6. METHODS OF THE IMPLICIT RUNGE-KUTTA TYPE

By leaving the class of linear multistep methods we found in § 5 in the method of Jain, an *A*-stable method of order $2n$, $n=1, 2, \dots$. In this section we will discuss the well known class of Runge-Kutta methods and show that in this class of methods we may also find *A*-stable methods of higher order.

6.1 Runge-Kutta Methods with ν-levels

We start with the differential equation

$$\dot{x} = f(x) \tag{6.1}$$

where x and f are m-vectors.

A Runge-Kutta process with ν-levels is defined by the following relations

$$\begin{aligned} &a) \quad x^{+} = x + h \sum_{i=1}^{\nu} b_i k_i \\ &b) \quad k_i = f(x+h \sum_{j=1}^{\nu} a_{ij} k_j), \qquad i=1,2,\dots,\nu. \end{aligned} \tag{6.2}$$

These relations are used to define an approximation, x^{+}, to $x(t_{n+1})$ in terms of an approximation to $x(t_n)$, denoted simply by x in (6.2). The coefficients b_i, a_{ij}, $i,j=1,2,\dots,\nu$ are to be determined by a procedure which we will now describe.

6.2 Determination of the Coefficients

By using (6.1) we may write the following list of formal relations

$$\begin{aligned} x^{(1)} &= f \\ x^{(2)} &= f_1 f \\ x^{(3)} &= f_2 f^2 + f_1^2 f \\ x^{(4)} &= f_3 f^3 + 3(f_2 f)(f_1 f) \\ &\dots \\ x^{(n)} &= \sum_{i=1}^{p_r} \alpha_{rs} F_{rs} \\ &\dots \end{aligned} \tag{6.3}$$

Here $f_1 = f_x$, the Jacobian, an array of order 2, $f_2 = f_{xx}$, the Hessian, an array of order 3, ...

The F_{rs}, $r=1,2,\dots,s=1,\dots,p_r$ are called the elementary differentials. For each index r, there are p_r such differentials. For example, $p_1=1$, $p_2=1$, $p_3=2$, $p_4=4$, ... and

$$F_{11} = f, \quad F_{21} = f_1 f, \quad F_{31} = f_2 f^2, \quad F_{32} = f_1^2 f.$$

Now let x^+ and x denote the exact value of x at t_{n+1} and t_n respectively. Next substituting the relations in (6.3) into the formal statement $x^+ - x = \sum_{n=1}^{\infty} h^n x^{(n)}/n!$, of Taylor's theorem gives

6.4) $$x^+ - x = \sum_{n=1}^{\infty} \frac{1}{r!} h^r \left(\sum_{s=1}^{p_r} \alpha_{rs} F_{rs} \right).$$

Now if we formally develop each k_i, $i=1,\ldots,\nu$ in a series, we may write the first relation in (6.2) as

6.5) $$h \sum_{i=1}^{\nu} b_i k_i = \sum_{n=1}^{\infty} \frac{1}{(r-1)!} h^r \left(\sum_{s=1}^{p_r} \beta_{rs} \phi_{rs} F_{rs} \right).$$

Here the β_{rs} are numerical coefficients while the ϕ_{rs} are functions of the b_i and the a_{ij}.

For a Runge-Kutta process to be of order of precision, p, it is necessary that the formal series in (6.4) and (6.5) agree to p terms. Thus we find

6.6) $$\phi_{rs} = \alpha_{rs}/(r\beta_{rs}), \qquad r=1,\ldots,p, \quad s=1,\ldots,p_r,$$

as a set of $M = \sum_{n=1}^{p} p_r$ equations for the determination of the $\nu(\nu+1)$ coefficients a_i, b_{ij} $i,j=1,\ldots,\nu$.

One distinguishes three classes of Runge-Kutta processes as follows:

Def. 6.1: A Runge-Kutta process is said to be explicit if $a_{ij}=0$, $j \geqslant i$, is said to be semi-explicit if $a_{ij}=0$, $j > 1$, and is said to be implicit otherwise. The number of available coefficients in these three cases are N_e, N_s, and N_i, respectively where

$$N_e = \nu(\nu+1)/2, \qquad N_s = \nu(\nu+3)/2, \qquad N_i = \nu(\nu+1).$$

The relation between the quantities ν, N_e, N_s, N_i; p and M is expressed in the following Table 6.1.

The M equations in (6.6) are not independent and so it is usually possible to satisfy them with a number N of coefficients considerably smaller than M.

ν	N_e	N_s	N_i	p	M
1	1	2	2	1	1
2	3	5	6	2	2
3	6	9	12	3	4
4	10	14	20	4	8
5	15	20	30	5	17
6	21	27	42	6	37
7	28	35	56	7	85

Table 6.1

6.3 An Example

Let us illustrate the last point by means of the case $p=\nu=3$. In this case an explicit calculation using (6.2) gives

$$6.7)\qquad h\sum_{i=1}^{3} b_i k_i = h\left(\sum_{i=1}^{3} b_i\right)F_{11} + h^2\left(\sum_{i=1}^{3} b_i c_i\right)F_{21}$$

$$+\frac{h^3}{2}\left[\left(\sum_{i=1}^{3} b_i c_i^2\right)F_{31} + 2\left(\sum_{i=1}^{3}\sum_{j=1}^{3} b_i a_{ij} c_j\right)F_{32}\right] + O(h^4)$$

where

$$c_i = \sum_{j=1}^{3} a_{ij}.$$

This must be set equal to the right member of (6.5) which is

$$6.8)\qquad h(\beta_{11}\phi_{11})F_{11} + h^2(\beta_{21}\phi_{21})F_{21} + \frac{h^3}{2}(\beta_{31}\phi_{31}F_{31} + \beta_{32}\phi_{32}F_{32}) + O(h^4).$$

Comparing coefficients of the elementary differentials in (6.7) and (6.8) allows us to determine $\beta_{rs}\phi_{rs}$ as functions of

$\beta\varphi = \alpha/r$	r	α	p_r
$\beta_{11}\sum b_i = 1$	1	1	1
$\beta_{21}\sum b_i c_i = 1/2$	2	1	1
$\beta_{31}\sum b_i c_i^2 = 1/3$	3	1	2
$\beta_{32}\sum_i \sum_j b_i a_{ij} c_i = 1/3$	3	1	

Table 6.2

the a_{ij} and the b_i. These are

$$\begin{aligned}
\beta_{11}\phi_{11} &= \sum_{i=1}^{3} b_i \\
\beta_{21}\phi_{21} &= \sum_{i=1}^{3} b_i c_i \\
\beta_{31}\phi_{31} &= \sum_{i=1}^{3} b_i c_i^2 \\
\beta_{32}\phi_{32} &= 2 \sum_{i=1}^{3} \sum_{j=1}^{3} b_i a_{ij} c_j
\end{aligned} \tag{6.9}$$

Next the expression in (6.4) must be developed so that the α_{rs} may be obtained. This reveals that $\alpha_{11}=1$, $\alpha_{21}=1$, $\alpha_{31}=1$ and $\alpha_{32}=1$. (Recall that we have already noted that $p_1=p_2=1$ and $p_3=2$).

Now we may assemble the information developed for this example in the Table 6.2.

One associates the tableau of coefficients in Table 6.3 with the process

$$\begin{array}{ccc|c}
a_{11} & \dots & a_{1\nu} & c_1 \\
a_{\nu 1} & \dots & a_{\nu\nu} & c_\nu \\
\hline
b_1 & \dots & b_\nu &
\end{array} \qquad \text{where } c_i = \sum_{j=1}^{\nu} a_{ij}$$

Table 6.3

A particular solution of the equations displayed in Table 6.2 is displayed in the version of Table 6.3 corresponding to $\nu=3$ as follows:

$$\begin{array}{ccc|c}
0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 1/2 \\
-1 & 2 & 0 & 1 \\
\hline
1/6 & 2/3 & 1/6 &
\end{array}$$

Table 6.4

with $\beta_{11}=\beta_{21}=\beta_{31}=1$ and $\beta_{32}=2$.

This particular solution is due to Kutta.

6.4 Semi-explicit Processes and the Method of Rosenbrock

Among the implicit and semi-explicit Runge-Kutta processes are A-stable methods. The implicit processes lead to methods which are difficult to apply in general because at each step of the integration the k_i, $i=1,\dots,\nu$ must be determined as the solution of the system of ν nonlinear equations (6.2b).

In the semi-explicit case the nonlinear system is triangular in the sense that the j-th equation in this system contains only the unknowns k_i, $i=1,\dots,j$. Thus each equation in turn need only be solved for one unknown, i.e., the i-th equation for k_i, $i=1,\dots,\nu$.

Let us consider the semi-explicit case and replace the solution procedure for the k_i, $i=1,\dots,\nu$, by a single step of a Newton-Raphson iteration. The resulting method is

6.10)
$$x^{+} = x + h\sum_{i=1}^{\nu} b_i k_i$$

$$k_i = \left[I - ha_{ii} f_x\left(x + h\sum_{j=1}^{i-1} a_{ij}k_j\right)\right]^{-1} f\left(x + h\sum_{j=1}^{i-1} c_{ij}k_j\ ,\right)\quad i=1,\dots,\nu,$$

where I is the $m \times m$ identity matrix. This is an example of a method which may be called a linearized semi-explicit Runge-Kutta process of the Rosenbrock type, or simply a Rosenbrock method.

The case $p=3$, $\nu=2$ becomes, using Rosenbrock's notation,

6.11)
$$\begin{aligned} x^{+} &= x + h(R_1k_1 + R_2k_2) \\ k_1 &= [I - ha_1f_1]^{-1} f \\ k_2 &= [I - ha_2f_1(x + hc_1k_1)]^{-1} f(x + hb_1k_1). \end{aligned}$$

The are six undetermined coefficients. The set of equations analogous to (6.6) for the determination of the six unknowns are four in number and are

6.12)
$$\begin{aligned} R_1 + R_2 &= 1 \\ R_1a_1 + R_2(a_2 + b_1) &= \frac{1}{2} \\ R_1a_1^2 + R_2[a_2^2 + (a_1 + a_2)b_1] &= \frac{1}{6} \\ R_2\left(a_2c_1 + \frac{1}{2}\,b_1^2\right) &= \frac{1}{6}\,. \end{aligned}$$

A particular solution of (6.12) due to Rosenbrock is

$$
\begin{aligned}
a_1 &= 1 + 1/\sqrt{6}\\
a_2 &= 1 - 1/\sqrt{6}\\
b_1 &= c_1 = [-6-\sqrt{6} + (58+20\sqrt{6})^{1/2}]/(6+2\sqrt{6})\\
R_1 &= -0.413154\\
R_2 &= -1.413154
\end{aligned}
$$

The two matrices in (6.11) which must be inverted become identical under the constraints $a_1=a_2$ and $c_1=0$. This considerably reduces the computation per step. Under these constraints the equations (6.12) become

6.13)

$$
\begin{aligned}
R_1 + R_2 &= 1\\
a_1 + R_2 b_1 &= 1/2\\
a_1^2 + 2R_2 a_1 b_1 &= 1/6\\
R_2 b_1^2 &= 1/3.
\end{aligned}
$$

(6.13) has two solutions. Calahan studied the following one

$$R_1 = 3/4, \qquad R_2 = 1/4, \qquad a_1 = (1+\sqrt{3})/2, \qquad b_1 = -2/\sqrt{3}.$$

6.5 *A*-stability

To demonstrate the *A*-stability of these linearized methods requires their application to the scalar test equation (i.e., $f=\lambda x$, $f_x=\lambda$) and a study of the location of the roots of the characteristic equation corresponding to the difference equation which results. We forego these details.

REFERENCES

[6.1] Butcher, J.C., "Implicit Runge-Kutta Processes", Math. Comp. *18* (1964) 50-64.

[6.2] Calahan, D.A., "Numerical Solution of Linear Systems with Widely Separated Time Constants" Proc. IEEE 55 (1967) 2016-2017.

[6.3] Rosenbrock, H.H., "Some General Implicit Processes for the Numerical Solution of Differential Equations", Comp. J. 5 (1962) 329-330.

§ 7. EXPONENTIAL FITTING: LINEAR MULTISTEP METHODS

7.1 Exponential Fitting

We have now completed a review of some of the ideas and methods for approximating the solution of stiff equations which use a technique coupling small mesh increments during a transistory stage with a property of absolute stability during a permanent stage.

We now turn to a second class of methods which employ a different idea. Namely those which employ exponential fitting.

In the context of a simple example we have seen in § 1 that the control of the error, $e_n = u_n - y_n$ (c.f. (1.6)), depends on the stability of the amplification operator, $K(hA)$, and the closeness of $K(hA)$ to the solution operator, $S(hA)$. We saw in that example that $K(hA)$ is made close to $S(hA)$ by making $K(hz)$ close to $S(hz)$ for z in the spectrum $\sigma(A)$. This in turn is accomplished by making K close to S in a neighborhood of the origin and then shrinking $h\sigma(A)$ into this neighborhood by taking h small enough.

The methods of exponential fitting replace the single point at the origin by a set of points which we may call the fitting points in the complex plane. Then $K(z)$ is made close to $S(z)$ at all points in this set. Then by taking h small, the collection of points $h\sigma(A)$ tend to one or another of the fitting points.

This idea becomes interesting for stiff systems when we note that fitting points may be very large in magnitude, so that h is not required to scale the entire spectrum of A into a neighborhood of the origin. Of course in addition to being fitted, a method must be stable and convergent in some sense. Otherwise it is of no computational value. We discuss these latter points in § 7.4 and in § 8.

7.2 Some Examples of Exponential Fitting for Linear Multistep Methods

We may see how this idea works through use of several examples. Consider the following linear multistep formulas (7.1) - (7.4)

7.1) $$F_1: \quad x_{n+1} - x_n - h[(1-a)\dot{x}_{n+1} + a\dot{x}_n] = 0.$$

The order of the method is $p=2$ if $a=1/2$ and $p=1$ otherwise.

7.2) $$F_2 \quad x_{n+1} - x_n - \frac{1}{2} h[(1+a)\dot{x}_{n+1} + (1-a)\dot{x}_n)] + \frac{1}{4} h^2[(b+a)\ddot{x}_{n+1} - (b-a)\ddot{x}_n] = 0.$$

Here $p=4$ if $b=\frac{1}{3}$ and $a=0$, $p=3$ if $b=\frac{1}{3}$, $a\neq 0$ and $p=2$ if $b\neq\frac{1}{3}$. In particular for $b=\frac{1}{3}$, (7.2) becomes

7.3) $$F_3: \quad x_{n+1} - x_n - \frac{h}{2}[(1+a)\dot{x}_{n+1} + (1-a)\dot{x}_n] + \frac{h^2}{12}[(1+3a)\ddot{x}_{n+1} - (1-3a)\ddot{x}_n] = 0.$$

In turn, when $a=0$, (7.3) becomes

7.4) $$F_4: \quad x_{n+1} - x_n - \frac{h}{2}(\dot{x}_{n+1} + \dot{x}_n) + \frac{h^2}{12}(\ddot{x}_{n+1} - \ddot{x}_n) = 0.$$

(7.2) and (7.4) are not the usual linear multistep methods since they employ second derivatives of x.

The exact solution of the test equation (c.f. (3.2)) satisfies the following recurrence relation

7.5) $$x(t_{n+1}) = e^q x(t_n), \qquad q = \lambda h.$$

The amplification factor of F_ν is $K_\nu(q)$, $\nu=1,2,3,4$ where

7.6) $$\begin{aligned} K_1(q) &= (1+aq)/[1-(1-a)q] \\ K_2(q) &= [4+2(1-a)q+(b-a)q^2]/[4-2(1+a)q+(b+a)q^2] \\ K_3(q) &= [12+6(1-a)q+(1-3a)q^2]/[12-6(1+a)q+(1+3a)q^2] \\ K_4(q) &= [12+6q+q^2]/[12-6q+q^2]. \end{aligned}$$

It is a simple matter to verify that

7.7) $$T_\nu(q) = K_\nu(q) - e^q = O(q^{\nu+1})$$

as $q \to 0$, since p has the various values 2, 3, or 4 as we have noted as the case may be.

We introduce the following definition of exponential fitting.

Def. 7.1: A method with truncation operator $T(q)$ is exponentially fitted to order r at a point c if $\frac{d^j}{dq^j} T(q) \big|_{q=c} = 0, \; j=0,1,\dots,r.$

We note that the formulas F_ν are exponentially fitted to order $r \geqslant \nu$ at the origin. The remaining parameters may be chosen so that fitting occurs elsewhere as well. If we can adjust F_ν so that $T_\nu(h\gamma)=0$, where the magnitude of γ is very large, then it is reasonable to use F_ν to solve stiff systems whose spectrum is divided into two clusters. The first cluster lying near $q=0$ corresponds to slowly varying modes; the second cluster, lying near $q=h\gamma=c$, corresponds to rapidly varying (stiff) modes.

Let us now consider some fittings of the F_ν.

For $a=0$, F_1 is fitted to order $r=0$ at $c=-\infty$.

For $a=\frac{1}{2}$, the trapezoidal formula, the fitting is maximal at $q=0$ $(p=r=2)$, but there is no fitting at $c=-\infty$, since $\lim\limits_{q \to -\infty} T_1(q) = -1$.

For $\nu=1$ or 3, $T_\nu(c)=0$ defines the parameter a as a function $a=a_\nu(c)$ where

7.8) $$a_1(q) = -q^{-1} - (e^{-q} - 1)^{-1}$$

and

$$a_3(q) = \frac{1}{3}[12+6q+q^2-(12-6+q^2)e^q]/[2q+q^2-(2q-q^2)e^q],$$

respectively.

$T_2(c)=T_2(c')=0$ define a and b as functions of both c and c'. These two functions are respectively:

$$a_2(q,q') = 2[f(q)-f(q')]/[q'f(q)-qf(q')]$$

and

$$b_2(q,q') = 2(q'-q)[q'f(q)-qf(q')].$$

Here

$$f(q) = q^2(e^q-1)/[2+q+(q-2)e^q].$$

7.3 Minimax Fitting

As an alternate use of free parameters, we may attempt to minimize $T(q)$ in some global sense. We illustrate this by means of the following example dealing with F_1.

Let

$$\overline{T}(a) = \max_{-\infty<q\leqslant 0} |T(q)|.$$

From (7.7) the following lemma results from a direct calculation.

Lemma 7.1: $a=a_1(c)$ defines a one-one mapping of $(-\infty,0]$ into $[0,1/2]$.

Now let a_0 be defined by

$$\overline{T}(a_0) = \min_{0\leqslant a\leqslant \frac{1}{2}} \overline{T}(a) = \min_{-\infty<c\leqslant 0} \overline{T}(a_1(c)).$$

Then

$$a_0 = 0.122 \ldots, \qquad \overline{T}(a_0) = 0.139 \ldots$$

and the corresponding fitting point $c_0=-8.19 \ldots$. Notice that for the backward Euler formula $\overline{T}(0)=0.204 \ldots$, while $\overline{T}(1/2)=1$ for the trapezoidal formula.

7.4 An Error Analysis for an Exponentially Fitted F_1

In the classical case fitting at the origin is a form of control of the local error, i.e., is tantamount to what we call local error analysis. Then we see that exponential fitting is a somewhat complicated variant of local error analysis. Just as in the classical procedure wherein a local error analysis by no means assures the control of the global error, we also lack this assurance in the case of exponential fitting. We must supplement the local analysis with a stability analysis and them combine the two to demonstrate the value of the method by constructing a global error analysis.

We will illustrate such a global error analysis with F_1 (c.f. (7.1)). In § 8, we will consider a more general framework.

When F_1 is applied to the linear system (1.1), viz.,

7.9) $$\dot{y} = Ay\ ,$$

we find the following recurrence relation for the global error, e_n;

7.10) $$e_{n+1} = K_1(hA)e_n + T_1(hA)y_n .$$

From this in turn we get

7.11) $$e_n = \sum_{j=0}^{n-1} K_1^j(hA)T_1(hA)y_{n-j-1}\ ,$$

where we have assumed that the initial error, $e_0=0$.

The following lemma follows from a direct calculation.

Lemma 7.2: $|K_1(z)| < 1$ for $z \in (0,-\infty)$ and $a \in [0,1/2]$.

This lemma asserts that F_1 is A-stable for $a \in [0,1/2]$. We now consider a to be restricted to this interval.

Now let us suppose that A is negative definite and has distinct eigenvalues $0 > \lambda_1, > \ldots, > \lambda_m$. Let the resolution of the identity, relative to A be given by

7.12) $$I = \sum_{i=1}^{m} P_i(A)$$

where the P_i, $i=1,\ldots,m$ are appropriate polynomials.

Then

7.13) $$||K_1^j(hA)|| = ||\sum_{i=1}^{m} K_1^j(h\lambda_i)P_i(A)|| \leqslant const \sum_{i=1}^{m} |K_1(h\lambda_i)| \leqslant const$$

The first equality in (7.13) follows from (7.12) while the last inequality follows from Lemma 7.2, since the λ_i are negative Using (7.13), (7.11) becomes

7.14) $$||e_n|| \leqslant const\ n||T_1(hA)|| .$$

Now from the properties of $T_1(z)$ for z near zero, we may conclude that

7.15) $$|T_1(z)| < const\ min(1,z^2), \qquad z \leqslant 0.$$

On the other hand given $c > 0$ and if $a=a_1(c)$ (c.f. (7.8)), then from Taylor's theorem, we conclude that

$$T_1(z) = (c-z)(K_1'(\tilde{z})+e^{\tilde{z}}).$$

From this in turn we have that

7.16) $$|T_1(z)| \leqslant const|c-z|, \qquad c < 0, \qquad z \leqslant 0.$$

Now let (I_1, I_2) be a partition, II, of the integers $I=\{1,\ldots,m\}$. Then combining (7.14)-(7.16) and utilizing the resolution of the identity, we get the following estimate for $||e_n||$.

7.17) $$||e_n|| \leqslant n\ const \min_{II} \ [\max_{i\in I_1} |h^2\lambda_i^2| + \max_{i\in I_2} h|\gamma-\lambda_i|].$$

$$\leqslant \max_{i\in I} \ [min(|h\lambda_i|^2, |\gamma-\lambda_i|)].$$

(Recall that $c=h\gamma$).

The property of Lemma 7.1 (i.e. the fitting) was observed by R.A. Willoughby, while that of Lemma 7.2 (i.e., the A-stability) was observed by W.Liniger. The global error analysis was made by W.L. Miranker. Thus, the simple scheme F_1 used in an exponential fitting mode for approximating the solution of stiff equations is called the Willoughby-Liniger-Miranker method.

REFERENCES

[7.1] Liniger, W., and Willoughby, R., "Efficient Integration Methods for Stiff Systems of Ordinary Differential Equations", SIAM J. Numer. Anal. 7 (1970) pp.47-66.

[7.2] See also the appendix of reference [8.1].

§ 8. FITTING IN THE MATRICIAL CASE

In this chapter we will study the process of exponential fitting in a setting which is more general than that of § 7. In particular, we consider a class of linear multistep methods with matricial coefficients.

8.1 The Matricial Multistep Method

We consider the initial value problem for the following system

8.1) $$\dot{x} = Ax\,, \qquad t > 0$$

Here x is an m-vector and A is an $m\times m$-matrix of constants. Evidently

8.2) $$x_n = e^{Ah} x_{n-1}\,.$$

Now consider the three functions $L(z)$, $R(z)$ and $C(z)$ given as follows:

8.3) $$L(z) = \sum_{j=0}^{r} (\alpha_j + z\beta_j) e^{(r-j)z}$$
$$R(z) = \sum_{j=0}^{r} (\gamma_j + z\delta_j) e^{(r-j)z}$$
$$C(z) = L(z)[R(z)]^{-1}\,.$$

Here the $\alpha_j, \beta_j, \gamma_j$ and δ_j, $j=0,\ldots,r$ are each $m\times m$-matrices. Note that

8.4) $$L(hA) - C(hA)\, R(hA) \equiv 0\,.$$

Let $P(z)$ be an approximation to $C(z)$ and consider the following formula, which is an approximation to (8.4), as a numerical method for determining u_n as an approximation to x_n, $n=r, r+1, \ldots$.

8.5) $$\sum_{j=0}^{r} \alpha_j u_{n-j} + h \sum_{j=0}^{r} \beta_j A u_{n-j} - P(hA)\left[\sum_{j=0}^{r} \alpha_j u_{n-j} + h \sum_{j=0}^{r} \sigma_j A u_{n-j}\right] = 0.$$

If $P(z)$ were equal to $C(z)$, this expression would be an identity for solutions of (8.1), (c.f. (8.8)), that is (8.5) would be fitted (exponentially) at all points in the spectrum $\sigma(A)$. However, $C(hA)$ is too difficult to calculate, especially if we use (8.5) on systems of the form (8.1) where A changes at each step. Thus we will choose $P(z)$ as a function for which $P(hA)$ is easy to calculate and such that $P(z)$ is an approximation to $C(z)$ in a sense to be made precise.

8.2 The Error Equation

To determine the quality of (8.5) as a numerical method we proceed to derive an equation for the global error $e_n = u_n - x_n$. To do this we introduce the shift operator, H where

8.6) $$Hf(t) = f(t+h),$$

and we introduce two operations $\mathcal{L}(H)$ and $\mathcal{R}(H)$ associated respectively with L and R as follows:

8.7)
$$\mathcal{L}(H) = \sum_{j=0}^{r} (\alpha_j + hA\beta_j)H^{r-j}$$
$$\mathcal{R}(H) = \sum_{j=0}^{r} (\gamma_j + hA\delta_j)H^{r-j}$$

(Except for the sign change, $\beta_j \to -\beta_j$, the $\mathcal{L}$ here is the same as the one used in § 2).

Now

8.8) $$Hx = e^{hA}x,$$

where x is a solution of (8.1).

Thus

8.9) $$(HA-AH)x = 0.$$

From this we may deduce that

8.10)
$$\mathcal{R}(H)x = R(hA)x,$$
$$\mathcal{L}(H)x = L(hA)x$$

and

8.11) $$[\mathcal{L}(H) - C(hA)\mathcal{R}(H)]x_{n-r} = 0, \qquad n=r, r+1, \dots .$$

On the other hand, we may write (8.5) as

8.12) $$[\mathcal{L}(H) - P(hA)\mathcal{R}(H)]u_{n-r} = 0, \qquad n=r, r+1, \dots .$$

Then by subtracting (8.11) and (8.12), we find the following error equation

8.13) $$[\mathcal{L}(H)-P(hA)\mathcal{R}(H)]e_{n-r} = [P(hA)-C(hA)]\mathcal{R}(H)x_{n-r}\,, \qquad n=r,\,r+1,\,\ldots\,.$$

8.3 Solution of the Error Equation

To solve (8.13) we introduce the operator $\mathcal{S}(H)$ as follows:

8.14) $$\mathcal{S}(H) = \mathcal{L}(H) - P(hA)\mathcal{R}(H).$$

We may write $\mathcal{S}(H)$ as a polynomial in H as follows:

8.15) $$\mathcal{S}(H) = \sum_{j=0}^{r} s_j H^{n-j}$$

where

8.16) $$s_j \equiv s_j(A) \equiv \alpha_j + hA\beta_j - P(hA)(\gamma_j + hA\delta_j), \qquad j=0,\ldots,r.$$

Thus (8.13) may be written in the following form

8.17) $$\mathcal{S}(H)e_{n-r} = [P(hA)-C(hA)]\mathcal{R}(hA)x_{n-r}, \qquad n=r,\,r+1,\,\ldots\,.$$

Now let

8.18) $$S(z) = \sum_{j=0}^{r} s_j z^{r-j}$$

be a polynomial with the matricial coefficients s_j, $j=0,\ldots,r$. Suppose that $[z^r S(z^{-1})]^{-1}$ is an analytic function of z in a neighborhood of $z=0$ and let its power series be given by

8.19) $$[z^r S(z^{-1})]^{-1} = \sum_{j=0}^{\infty} \sigma_j z^j$$

where the σ_j are matrices. (c.f. Lemma 8.2 below)

Multiply (8.17) by σ_{N-n}, and sum the result over n from r to N. For the left member of this operation we have

$$8.20)\quad \sum_{n=r}^{N} \sigma_{N-n}\mathcal{f}(H)e_{n-r} = \sum_{n=r}^{N} \sigma_{N-n} \sum_{j=0}^{r} s_j H^{r-j} e_{n-r}$$

$$= \sum_{n=r}^{N} \sigma_{N-n} \sum_{j=0}^{r} s_j e_{n-j}$$

$$= \sigma_0 s_0 e_N + (\sigma_1 \mathbf{R}_0 + \sigma_0 \mathbf{R}_1) e_{N-1} + \ldots$$

$$+ (\sigma_{N-r} \mathbf{R}_0 + \ldots + \sigma_N s_n) e_r$$

$$+ \text{linear combination of } e_0, e_1, \ldots, e_{r-1}.$$

From the defining property (8.19) of the σ_j, $j=0,\ldots,$ we may deduce the following:

$$8.21)\qquad \sum_{j=0}^{r} \sigma_{N-j} s_j = \delta_{N0} I_m$$

where I_m is the $m \times m$ identity matrix. Using (8.21) in (8.20) and assuming that the initial errors $e_0=e_1=\ldots=e_{r-1}=0$, we find that the right member of (8.20) becomes simply e_N. Thus we are led to the solution of (8.17), viz.,

$$8.22)\qquad e_N = \sum_{n=r}^{N} \sigma_{N-n} [P(hA)-C(hA)] R(hA) x_{n-r}.$$

8.4 Estimate of Global Error

To estimate e_N we require a usual stability statement and an accuracy statement. Stability is the subject of the following two lemmas.

Lemma 8.1: If $\sum_{j=0}^{r} s_j(\lambda) z^{r-j}$ satisfies the root condition for each eigenvalue $\lambda \in \sigma(A)$, then the determinant $|S(z)|$ also satisfies the root condition.

Proof: Let $f(A) = \sum_{j=0}^{r} s_j(A) z^{r-j}$. Suppose that the determinant $|f(A)|$ vanishes for a value of z, then $|f(A)+\mu I_m - \mu I_m|$ vanishes. Then $\mu=\mu+f(\lambda)$, for each $\lambda \in \sigma(A)$ or $f(\lambda)=0$ for that value of z. This completes the proof of the lemma.

Lemma 8.2: Let the determinant $|S(z)|$ of $S(z)$ obey the root condition. If the determinant of s_0 is not zero, then the matrix $[z^r S(z^{-1})]^{-1}$ is analytic in a neighborhood of $z=0$. Furthermore, the matrices σ_j, $j=0,1,\ldots$, given by (8.19), have uniformly bounded norms.

Proof: Since $z^r S(z^{-1}) = \sum_{j=0}^{r} s_j z^j$ and $|s_0| \neq 0$, it is clear that $[z^r S(z^{-1})]^{-1}$ is analytic in a neighborhood of the origin. Since $|z^r S(z^{-1})| = z^{mr}|S(z^{-1})|$, the root condition locates the roots of the polynomial $|zS(z^{-1})|$ outside the open unit disc and those roots on the boundary of the unit disc are simple. Since

$$[z^r S(z^{-1})]^{-1} = [\text{matrix of polynomials}]/|z^r S(z^{-1})|,$$

it suffices to show that the power series for the reciprocal polynomial, $|z^r S(z^{-1})|^{-1}$ has bounded coefficients, given that its roots are outside the open unit disc, with those on the boundary being simple. Let $mr=q$ and let

$$|z^r S(z^{-1})|^{-1} = \left[\sum_{j=0}^{q} t_j z^j\right]^{-1} = \sum_{j=0}^{\infty} u_j z^j .$$

Then

$$u_n = \frac{1}{2\pi i}\oint\left(\zeta^{n+1} \sum_{j=0}^{q} t_j \zeta^j\right)^{-1} d\zeta ,$$

where the contour of integration lies inside the unit disc and encircles the origin. If we move the contour through the unit disc and out to infinity in all directions, the integral will vanish if $q \geqslant 1$ and we are left with a sum of residues. If there is a pole ζ_0 on the unit disc, it is simple. Let the residue from it be τ_0. Then

$$|\tau_0| = \left|\left(\zeta_0^{n+1} \sum_{j=0}^{q} j\, t_j \zeta_0^{j-1}\right)^{-1}\right| = \left|\sum_{j=0}^{q} j\, t_j \zeta_0^{j-1}\right|^{-1}$$

which is independent of n.

If there is a pole at ζ_1 of order $\rho+1$ outside the unit disc, let the residue from it be τ_1. Then

$$\tau_1 = \frac{d\rho}{d\zeta^\rho}\left[(\zeta-\zeta_1)^{\rho+1}\left(\zeta^{n+1} \sum_{j=0}^{q} t_j \zeta^j\right)^{-1}\right]_{\zeta=\zeta_1} .$$

Let $Q(\zeta)$ be the polynomial given by

$$Q(\zeta)=\left(\sum_{j=0}^{q} t_j\zeta^j\right)/(\zeta-\zeta_1)^{\rho+1}.$$

Then since $Q(\zeta)$ is independent of ζ_1

$$\tau_1=\frac{d\rho}{d\zeta_1^{\rho}}[(\zeta_1^{n+1}Q(\zeta))^{-1}].$$

Then performing the differentiation we get

$$\tau_1=\sum_{j=0}^{\rho}\binom{\rho}{j}\left(\frac{d^j}{d\zeta_1^j}\zeta_1^{-n}\right)\frac{d^{\rho-j}}{d\zeta_1^{\rho-j}}(\zeta_1Q(\zeta_1))^{-1}$$

$$=\sum_{j=0}^{\rho}(-1)^j\binom{\rho}{j}n(n+1)\ \ldots\ (n+j-1)\zeta_1^{-(n+j)}\ \frac{d^{\rho-j}}{d\zeta_1^{\rho-j}}(\zeta_1Q(\zeta_1))^{-1}.$$

Thus

$$|\tau_1|\leqslant(n+\rho)^{\rho}F/|\zeta_1|^{n}$$

where F is a constant independent of n. This estimate shows that $|\tau_1|$ tends to zero when n tends to infinity since $|\zeta_1|>1$. Since there are at most a finite number of residues to be accounted for, the coefficients u_n, $n=0,1,\ldots,$ are bounded uniformly in n and the lemma is proved.

If $S(z)$ satisfies the hypothesis of Lemma 8.2 then that lemma and (8.22) may be combined to yield

8.23) $$||e_N||\leqslant const||[P(hA)-C(hA)]R(hA)||\sum_{n=r}^{N}||x_{n-r}||.$$

If $Nh=1$, (8.23) becomes

8.24) $$||e_N||\leqslant const\ h^{-1}||[P(hA)-C(hA)]R(hA)||.$$

To complete the error analysis the local error, which here is $||[P(hA)-C(hA)]R(hA)||$ must be made $o(h)$. To accomplish this we have at our disposal the specification of P, L and R to which we now turn.

8.5 Specification of P

Let $P(z)$ be a polynomial which has contact of order τ_i+1 with $C(z)$ at a set of points in the complex plane which we denote by hz_i, $i=1,\ldots,p$. That is

$$8.25)\qquad P^{(m)}(hz_i)-C^{(m)}(hz_i)=0, \qquad m=0,1,\ldots,\tau_i\,.$$

We suppose that $z_i \neq 0$, $i=1,\ldots,p$ and we set $z_0=0$.

Now divide the eigenvalues of A into $p+1$ disjoint clusters called $k_0,\ldots,k_p$, respectively, where

$$k_i=\{\lambda_j \in \sigma(A) \,|\, |\lambda_j-z_i| \leqslant \min_{0\leqslant l\leqslant p} |\lambda_j-z_l|\}.$$

Ties are decided randomly.

Let

$$d_i=\max_{\lambda_j\in k_i} |\lambda_j-z_i|, \qquad i=0,\ldots,p.$$

Now we resolve the identity by writing

$$8.26)\qquad I_m=\sum_{i=0}^{p}\sum_{\lambda_j\in k_i} Z_{ij}(hA),$$

where the Z_{ij} are appropriate polynomials and where for simplicity we have supposed that the eigenvalues of A are distinct. Using (8.26) we may obtain

8.27)
$$[P(hA)-C(hA)]R(hA)=\sum_{i=0}^{p}\sum_{\lambda_j\in k_i}[P(h\lambda_j)-C(h\lambda_j)]R(h\lambda_j)Z_{ij}(hA).$$

Using Taylor's theorem with remainder and (8.25), (8.27) becomes

$$8.28)\qquad [P(hA)-C(hA)]R(hA)=\sum_{\lambda_j\in k_0}[P(h\lambda_j)R(h\lambda_j)-L(h\lambda_j)]Z_{0j}(hA)$$
$$+\sum_{i=1}^{p}\sum_{\lambda_j\in k_i}\frac{1}{\tau_i!}[h(\lambda_j-z_i)]^{\tau_i}[P^{(\tau_i)}(h\tilde{\lambda}_{ij})-C^{(\tau_i)}(h\tilde{\tilde{\lambda}}_{ij})]$$
$$R(h\lambda_j)Z_{ij}(hA).$$

The $\tilde{\lambda}_{ij}$ and the $\tilde{\tilde{\lambda}}_{ij}$ are values of λ arising in the remainder term.

8.6 Specification of L and R

To specify L and R we make the hypothesis

$$8.29)\qquad \begin{aligned} L(z) &= O(z^{\mu+1}) \\ R(z) &= O(z^{\nu+1}) \end{aligned}$$

This hypothesis says that the classical (matricial) linear multistep methods $\mathcal{L}(H)u_{n-r}=0$ and $\mathcal{R}(H)u_{n-r}=0$ have order of accuracy μ and ν, respectively.

Using (8.29) in (8.28) gives

$$8.30)\qquad ||[P(hA)-C(hA)]R(hA)|| \leqslant C_1\, max\left(|hd_0|^{\nu+1}|hd_0|^{\mu+1}\right) + C_2 \sum_{i=1}^{p} \frac{1}{\tau_i!}|hd_i|^{\tau_i+1}$$

Here C_1 and C_2 are appropriate constants. (8.30) is the local error (estimate) for the numerical method, (8.5) which we are studying. Combining (8.30) with (8.24) gives finally the global error estimate

$$8.31)\qquad ||e_N|| \leqslant const\left[max\left(|hd_0|^{\nu}, |hd_0|^{\mu}\right) + \sum_{i=1}^{p}|hd_i|^{\tau_i}\right].$$

Remark 8.1: The classical theory of linear multistep methods corresponds to the case $P\equiv 0$.

8.7 An Example

A simple example of the method (8.5) corresponds to the case $r=1$, $\alpha_0=1$, $\alpha_1=-1$ and $\delta_1=1$. All other coefficients are zero. We select one cluster, i.e., $p=1$ and $P(z)$ is taken to be the constant, $C(hz_1)$. The numerical method is

$$8.32)\qquad u_n - u_{n-1} = \frac{e^{hz_1}-1}{hz_1}\, h\dot{u}_{n-1}.$$

For this method $\mu=\nu=\tau_1=0$. Thus the method has accuracy of order zero at the origin and at z_1. This low accuracy method may be viewed as the forward Euler method with a mesh increment scaled by $(e^{hz_1}-1)/(hz_1)$.

For this method $S(z)=Iz-(I+((e^{hz_1}-1)/z_1)A)$. By Lemma 8.1, $|S(z)|$ obeys the root condition if $z-1-((e^{hz_1}-1)/z_1)\lambda$ does for every eigenvalue λ of A. This latter requirement is seen to be satisfied for any choice of z_1 in an interval which itself is contained in the interval $(-\infty,\lambda)$. (We are assuming that $\lambda<0$). Thus if z_1 is chosen as any lower estimate for the spectrum of A, (8.32) will be stable.

Let us choose $z_1=\min\limits_{\lambda\in\sigma(A)}\{\lambda-d\}$ for some $d\geqslant 0$. To simplify things, let us consider the special case corresponding to $m=2$ and to say $\lambda_2=-1$ and λ_1 some very large negative number. The difference scheme then becomes

$$8.33)\qquad u_n-u_{n-1}=\frac{e^{h(\lambda_1-d)}-1}{\lambda_1-d}Au_{n-1}\approx\frac{1}{d-\lambda_1}Au_{n-1},$$

since $\lambda<<-1$.

Now since $x_n=e^{Ah}x_{n-1}$ and $u_n=\left[I+\frac{e^{h(\lambda_1-d)}-1}{\lambda_1-d}A\right]u_{n-1}$, we have

$$8.34)\qquad e_n=T(hA)e_{n-1}$$

where

$$T(hA)=I+\frac{e^{h(\lambda_1-d)}-1}{\lambda_1-d}A-e^{Ah}.$$

$T(h\lambda)$ is then the difference between the exponential $e^{\lambda h}$ and the straight line $1-(e^{h(\lambda_1-d)}-1)\lambda/(\lambda_1-d)$. At the eigenvalue λ_1, we have

$$T(h\lambda_1)=\frac{d}{\lambda_1}+e^{h\lambda_1}\left[1+d\left(\frac{1}{\lambda_1}-h\right)+O\left(d^2\left(\frac{1}{\lambda_1}-h\right)^2\right]+O\left(\frac{d^2}{\lambda_1^2}\right).$$

The following figure indicates how a forward Euler-type formula may be used to stably integrate a stiff system.

From the figure we see that we scale the z-axis so that we use the method (the straight line) in a region where it is stable, but where its value (of the straight line) is equal to the value of the exponential (the transfer function of the solution) at the large eigenvalue.

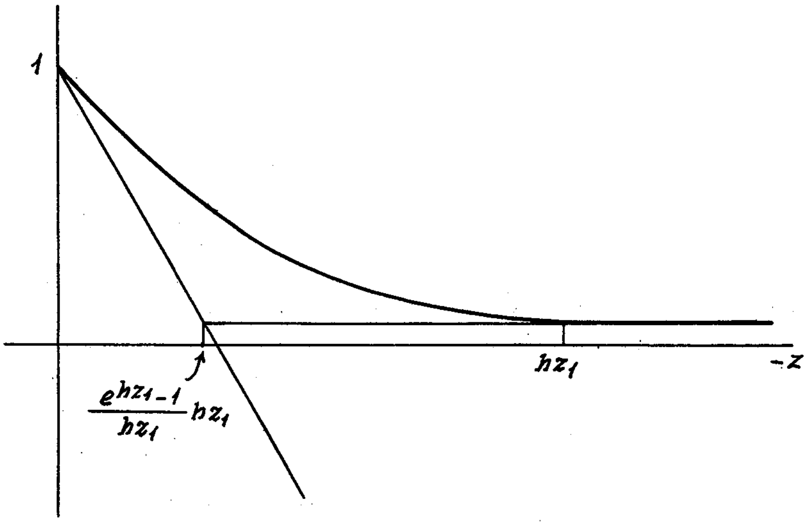

Figure 8.1

We remark that the matricial class of methods being discussed here is very wide and the operative qualities of the class are by no means restricted to the scaling concept of the example.

REFERENCES

[8.1] Miranker, W.L., "Matricial Difference Schemes for Integrating Stiff Systems of Ordinary Differential Equations", Math. Comp. 25 (1971) pp. 717-728.

§ 9. FITTING IN THE CASE OF PARTIAL DIFFERENTIAL EQUATIONS

Partial differential equations of evolutionary type along with their numerical treatment are subject to being ill conditioned. In some cases this ill conditioning resembles the state of affairs for stiff ordinary differential equations. The remedy of exponential fitting for the latter has a counter part for partial differential equations and we will review this counterpart in this chapter. As we might expect in the partial differential equations case, the idea of exponential fitting is susceptible to a much wider scope of possibilities and results than in the ordinary differential equations case.

We begin with a review of a simple problem and an elementary error analysis to motivate our discussion.

9.1 The Problem Treated

Let D be the domain of points, $D=\{(x,t) | t\in[0,T], |x|<\infty\}$ and consider the initial value problem

9.1)
$$u_t = \lambda u_x, \qquad (x,t)\in D, \quad t\neq 0,$$
$$u(x,0) = f(x), \qquad t = 0.$$

Here λ is a scalar and u and f are real valued scalar functions. This elementary problem has the solution

9.2)
$$u(x,t) = f(x+\lambda t).$$

In the half plane, $t\geqslant 0$, we set down a mesh, M, with increments Δt and Δx, i.e. $M=\{(x_j,t_n)=(j\Delta x, n\Delta t)\ j=0,\pm 1,\ldots;\ n=0,1,\ldots\}$. We may suppose without loss of generality that $\Delta t=\Delta x=h$.

Let $u_n(x)\equiv u_n\equiv u(x,nh)$. Then letting S denote the solution operator of (9.1), we find

9.3)
$$u_{n+1} = S\left(h\frac{\partial}{\partial x}\right)u_n, \qquad n=0,1,\ldots,$$
$$S(z) = e^{\lambda z},$$

as (9.2) shows.

As a numerical approximation to u_n we take $v_n\equiv v_n(x)$ where

9.3)
$$v_{n+1} = \sum_{|j|\leqslant l} a_j H^j v_n, \qquad n=0,1,\ldots$$
$$v_0 = f(x)$$

Here $l\geqslant 0$ is an integer and H is the shift operator, $Hf(x)=f(x+h)$.

(9.3) is commonly called a two level explicit difference scheme. If $|a_l|+|a_{-l}|\neq 0$, we will say that this scheme has width l. We write (9.3) as

9.4) $$v_{n+1} = Kv_n$$

where K is the amplification operator of the scheme.

If the powers $||K||^j$, $j=1,2,\ldots$ are bounded, then the numerical scheme is stable and we may obtain the following bound for the global error, $e_n = v_n - u_n$.

9.5) $$||e_n|| \leqslant const\ n \max_{0\leqslant p\leqslant n} ||Tu_p||.$$

Here $T=K-S$ is the truncation operator and we are using $||\cdot||$ to denote the norm in $L^2[-\infty,\infty]$.

Using Taylor's theorem and the consistency relations

9.6) $$1 - \sum_{|j|\leqslant l} a_j = 0$$

$$\lambda - \sum_{|j|\leqslant l} ja_j = 0,$$

(9.5) becomes

9.7) $$||e_n|| \leqslant const\ nh^2 \max_{0\leqslant p\leqslant n} ||u_p''(\eta)||,\quad \eta\in(x-lh,x+lh).$$

If u_p'' exists and is bounded by a constant M uniformly in the domain D, the bound (9.7) becomes

9.8) $$||e_n|| \leqslant const\ Mh\ ,$$

provided that $nh \leqslant T$.

As the data, $f(x)$ or the solution, $u_p(x)$ becomes less smooth, the bound, (9.8) becomes less satisfactory and convergence of the pointwise error to zero with h becomes slower and slower. Indeed, when the data or solution becomes discontinuous, there is no bound, M, at all and the convergence of the pointwise error is a delicate question. This difficulty in turn is reflected in an inadequate state of affairs in actual computations for such problems. The problems are ill-conditioned. Indeed, as the data becomes less smooth; the values of its Fourier transform at larger frequencies tend to grow. Since the spectrum of $\lambda\partial/\partial x$ is continuous, we see then that as $S(\lambda\partial/\partial x)$ develops, the solution, it receives increasing input at greater frequencies as the data degrades.

We see then that the situation is quite analogous to the case of stiff systems of ordinary differential equations.

What we will do is to return to the bound (9.5) for $||e_n||$ and make $||Tu_p||$ as small as possible. That is we will minimize $||Tu_p||$ over the set of real coefficient vectors $a\equiv(a_{-l},\ldots,a_l)$. An alternative approach would be to minimize the $\max\limits_{u_p}||Tu_p||$, a procedure which resembles the minimax fitting discussed in §7.3. We will not discuss this possibility here, but refer to [9.2] and [9.3] for details. Instead we will consider a set of special cases in which we replace this maximation over u_p by an appropriate choice of u_p itself. The principle being that if we wish to derive a numerical method with desirable properties relative to a given type of problem (or data), we cause the properties which are wanted, to be taken on by constraining the minimization or fixing the weight function u_p. We will henceforth drop this subscript p.

9.2 The Minimization Problem

To formulate the minimization problem to be considered we introduce the Fourier transform $\hat{f}$ of f where

$$\hat{f} = \hat{f}(\omega) = \int e^{-i\omega x} f(x)dx.$$

Then the minimization problem becomes

$$\min_a ||Tu|| = \min_a ||(\hat{K}-\hat{S})\hat{u}||. \tag{9.9}$$

Here

$$\hat{K}(z) = \sum_{|j|\leqslant l} a_j e^{ijz} \tag{9.10}$$

$$\hat{S}(z) = e^{i\lambda z}.$$

Then the function to be minimized is

9.11) $$J = ||(\hat{K}-\hat{S})\hat{u}||^2 = \int_{-\infty}^{\infty} |\hat{K}(h\omega)-\hat{S}(h\omega)|^2 |u(\omega)|^2 d\omega.$$

Finally consider the following definition and the ensuing remark.

Def. 9.1: We call the schemes which use the vector of coefficients, a, determined by the minimization problem (9.9), schemes with best possible (local-) truncation error, or simply best possible schemes.

Remarks 9.1: Schemes for which the $2l+1$ degrees of freedom represented by a are chosen so as to achieve the relation

9.12) $$\hat{K}(h\omega) = \hat{S}(h\omega) + O((h\omega)^p), \qquad p = 2l$$

are the classical schemes. These schemes are schemes of maximal order or of maximal (local-) accuracy. They have been named the most accurate schemes by G. Strang.

The relation (9.12) for any $p \leqslant 2l$ is equivalent to the following p moment conditions

9.13) $$\sum_{|j| \leqslant l} j^r a_j = \lambda^r, \qquad r=0,1,\dots,p, \qquad p \leqslant 2l.$$

9.3 Highly Oscillatory Data
Derivation of the Quadratic Form

For problems with highly oscillatory data, a good choice of $u(x)$ is one such that

9.14) $$|\hat{u}(\omega)|^2 = \begin{cases} 1, & |\omega| < c/h \\ 0, & |\omega| > c/h, \quad c \text{ constant.} \end{cases}$$

In this case we denote J (c.f. (9.11)) by J_c. Evidently

9.15) $$J_c = \frac{2h}{c}\int_{-c/h}^{c/h} |\hat{K}(h\omega)-\hat{S}(h\omega)|^2 d\omega.$$

For $c=\pi/\lambda$, we find

9.16)
$$J_{\pi/\lambda} = 1 + \sum_{|j|\leqslant l} a_j^2 - \frac{2\lambda}{\pi} a_j \frac{\sin j\frac{\pi}{\lambda}}{\lambda - j}$$
$$+ \frac{2\lambda}{\pi} \sum_{k=1}^{2l} \left(\sum_{j_1 - j_2 = k} a_{j_1} a_{j_2} \right) \frac{\sin\frac{k\pi}{\lambda}}{k}.$$

For $c=\pi$, we find

9.17)
$$J_\pi = 1 + \sum_{|j|\leqslant p} a_j^2 - 2\frac{\sin\lambda\pi}{\pi} \Sigma(-1)^j \frac{a_j}{\lambda}$$

For $c=p\pi$, p an integer, we find

9.18)
$$J_{p\pi} = 1 + \sum_{|j|\leqslant l} a_j^2 .$$

Consistent Formulas

Let us minimize J_π with respect to a and subject to the constraints of consistency (9.6). We may expect the resulting finite difference scheme to be good uniformly over all frequencies. The minimizing a_j is

9.19)
$$a_j = \frac{1}{2l+1}\left[1-\rho\sum_{|j|\leqslant l}\frac{(-1)^k}{\lambda-k}\right] + \frac{1}{2S_2}\left[\lambda-\rho\sum_{|j|\leqslant l}\frac{(-1)^k k}{\lambda-k}\right] + \frac{(-1)^j}{\lambda-j}\rho ,$$
$$\rho = \frac{\sin\lambda\pi}{\pi}, \qquad S_2 = \frac{1}{6}\, l(l+1)(2l+1).$$

If in (9.19) we set $\lambda=m$, an integer, we get

9.20)
$$a_j = \delta_{jm} .$$

In this case the difference scheme propagates information precisely along the characteristic of the partial differential equation, i.e., the numerical solution is exact.

In the case $l=1$, (9.19) becomes

9.21)
$$a_0 = \frac{1}{3}\left[1 + \frac{\lambda^2+1}{\lambda(\lambda^2-1)}\,\rho\right] + \frac{\rho}{\lambda},$$
$$a_{\pm 1} = \frac{1}{3}\left[1 + \frac{\lambda^2+1}{\lambda(\lambda^2-1)}\,\rho\right] \pm \frac{1}{2}\left[\lambda + \frac{2}{\lambda^2-1}\,\rho\right] - \frac{1}{\lambda \mp 1}.$$

We also find that the minimum of $J_{p\pi}$ is taken on when

9.22)
$$a_j = \frac{1}{2l+1} + \frac{j\lambda}{2S_2}.$$

This scheme is always stable. To see this note that

$$\min_{|j|\leqslant l} a_j = a_{-l} = \frac{p(l+1)-3}{p(l+1)(2l+1)} > 0,$$

since $p \geqslant 2$, $l \geqslant 1$, and appeal to the following lemma.

Lemma 9.1: Difference schemes of the type (9.3) for which $\sum_{|j|\leqslant l} a_j = 1$ and $a_j \geqslant 0$, $j = 0, \pm 1, \ldots, \pm l$ are stable.

Consistent Formulas Which Are Fitted at High Frequency

If the data has large frequency components, the constraints

9.23)
$$\hat{T}(z)\Big|_{z=\pm\frac{c}{h}} = 0$$

suggest themselves. The minimum of $J_{p\pi}$ subject to the four constraints

9.24)
$$\hat{T}(0) = \hat{T}'(0) = \hat{T}\left(p\frac{\pi}{h}\right) = T\left(-p\frac{\pi}{h}\right) = 0$$

occurs at

9.25)
$$a_j = \begin{cases} \dfrac{1}{2l+1} + \dfrac{j}{2pS_2}, & p \text{ even} \\[2ex] [1-(-1)^j]\left[\dfrac{(-1)^l}{l(1+2l)} + \dfrac{1}{2l}\right] + \dfrac{j}{2p\,S_2}, & p \text{ odd}. \end{cases}$$

In the case of even p,

$$\min_{|j|\leqslant l} a_j = a_{-l} = \frac{p(l+1)-3}{p(l+1)(2l+1)} > 0,$$

since p, $l \geqslant 2$. Thus in the case of even p, the schemes given by (9.25) are always stable.

9.4 Systems

Derivation of the Quadratic Form

This approach to the determination of difference schemes may be carried over directly to the case of systems of first order partial differential equations.

Let u and v be q-vectors and let A be a $q \times q$ matrix. We consider the initial value problem for

9.26) $$u_t = Au_x, \qquad (x,t)\epsilon D, \qquad t\neq 0.$$

The difference scheme is

9.27) $$v_{n+1} = \sum_{|j|\leqslant l} B_j H^j v_n ,$$

where the B_j, $|j| \leqslant l$ are $q \times q$ matrices.

Proceeding as before by taking Fourier transforms of (9.26) and (9.27), we are led to the problem of minimizing the following functional

9.28) $$J = ||(\hat{K}(h\omega)-\hat{S}(h\omega))u(\omega)||^2.$$

Here $\hat{K}(z)$ and $\hat{S}(z)$ are the $q \times q$ matrices given by

9.29) $$\hat{K}(z) = \sum_{|j|\leqslant l} B_j e^{ijz}$$
$$\hat{S}(z) = e^{izA}$$

respectively.

For the weight function we choose $\hat{u}(\omega)$ to be

9.30) $$\hat{u}(\omega) = \eta \cdot \begin{cases} 1, & |\omega| \leqslant \pi/h, \\ 0, & \text{otherwise} \end{cases}$$

Here η is the q-vector all of whose components are unity. This choice of $\hat{u}(\omega)$ makes J correspond to the functional J_π in (9.17).

Now using $(\cdot,\cdot)$ to denote the inner product in Euclidean q-space, we may rewrite J as follows:

9.31) $$J = \frac{1}{2\pi}\int_{-\pi}^{\pi}((\hat{S}(z)-\hat{K}(z))\eta\ ,(\hat{S}(z) - \hat{K}(z))\eta)dz\ .$$

Now suppose that A is a symmetric matrix with eigenvalues, λ_i, $i=1,\dots,q$. Let U be the unitary matrix which diagonalizes A, viz.,

9.32) $$UAU^{-1} = \Lambda$$

where Λ is the diagonal matrix whose ii-th entry is λ_i, $i=1,\dots,q$. Let $UB_jU^{-1}=C_j$ and let $U\eta=\mu$. Then (9.31) becomes

9.33) $$J=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\left(e^{i\Lambda z} - \sum_{|j|\leqslant l} C_j e^{ijz}\right)\mu,\ \left(e^{i\Lambda z} - \sum_{|j|\leqslant l} C_j e^{ijz}\right)\mu\right)dz.$$

Now let $C_j=(c^j_{mn})$, $m,n=1,\dots,q$ and let $\mu=(\mu,\dots,\mu_q)^T$. Also let

9.34) $$\gamma^j_m = \sum_{n=1}^{q} c^j_{mn}\mu_n .$$

Then J becomes

9.35) $$J = \sum_{m=1}^{q}\left[\mu_m^2 + \sum_{|j|\leqslant l}(\gamma^j_m)^2 - 2\frac{\sin\lambda_m\pi}{\pi}\sum_{|j|\leqslant l}(-1)^j\frac{\gamma^j_m\mu_m}{\lambda-j}\right].$$

The constraints of consistency are

9.36) $$\sum_{|j|\leqslant l} j^k B_j = A^k, \qquad k=0,1$$

or

9.37) $$\sum_{|j|\leqslant l} j^k C_j = \Lambda^k, \qquad k=0,1.$$

An Example

Let us consider the case corresponding to the wave equation. Here $q=2$ and $A=c\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$. Then we find

$$9.38)\quad U=\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},\ \mu=\begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix},\ \lambda_1=-\lambda_2=c \quad\text{and}\quad \gamma_m^j=\sqrt{2}\,c_{m1}^j .$$

J becomes

$$9.39)\quad J=2\left[1+\sum_{j=-1}^{1}[(c_{11}^j)^2+(c_{21}^j)^2]-2\frac{\sin c\pi}{\pi}\sum_{j=-1}^{1}(-1)^j\frac{c_{11}^j}{c-j}\right].$$

The solution to the constrained minimization problem is

$$9.40)\quad \begin{aligned} B_1 &= \frac{1}{2}\gamma_{-1}M_1+\frac{c}{4}M_2+\frac{1}{4}M_3 , \\ B_0 &= \frac{1}{2}\gamma_0 M_1-\frac{1}{4}M_3 , \\ B_1 &= \frac{1}{2}\gamma_1 M_1-\frac{c}{4}M_2-\frac{1}{4}M_3 . \end{aligned}$$

Here

$$9.41)\quad M_1=\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\quad M_2=\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},\quad M_3=\begin{pmatrix} \alpha-\beta & -\alpha+\beta \\ \alpha+\beta & -\alpha-\beta \end{pmatrix}.$$

α and β are arbitrary parameters and

$$9.42)\quad \begin{aligned} \gamma_{-1} &= \frac{1}{3}-\frac{c}{2}-\frac{1}{6}\,\frac{\rho}{c+1}\,\frac{2c^2-1}{c(c-1)} , \\ \gamma_0 &= \frac{1}{3}+\frac{1}{2}\,\frac{\rho}{c}\left(1+\frac{1}{3}\,\frac{1+c^2}{c^2-1}\right), \\ \gamma_1 &= \frac{1}{3}+\frac{c}{2}\,\frac{\rho}{6(c-1)}\,\frac{2c^2-1}{c(c+1)} \end{aligned}$$

with $\rho=2\dfrac{\sin c\pi}{\pi}$.

α and β may be chosen so that the resulting difference is more like the usual scalar scheme. This may be accomplished by demanding that

9.43) $$B_k = P_k(A), \qquad k=0,\pm 1,$$

where the $P_k(A)$ are polynomials in A. We find that $\alpha=\beta=0$ and

$$P_{-1}(A) = \left(\frac{1}{2}\gamma_{-1} + \frac{c}{4} I\right) + \left(\frac{1}{4} - \frac{\gamma_{-1}}{2c}\right) A,$$

9.44) $$P_0(A) = \frac{1}{2}\gamma_0 I + \frac{1}{2c}\gamma_0 A\ ,$$

$$P_1(A) = \left(\frac{1}{2}\gamma_1 - \frac{c}{4}\right) I - \left(\frac{1}{4} + \frac{\gamma_1}{2c}\right) A.$$

9.5 Discontinuous Data

The Scalar Case

We return to the scalar case and consider the problem of optimizing the difference scheme when the data is discontinuous. Thus we are interested in minimizing $||Tu||$ when

9.45) $$u(x) = \begin{cases} 1, & x \geqslant 0 \\ 0, & x < 0. \end{cases}$$

In this case

9.46) $$\hat{u}(\omega) = \lim_{d\to\infty} \int_{-d}^{d} e^{i\omega x} u_n(x)dx = \lim_{d\to\infty} \frac{i}{\omega} [e^{i\omega d} - 1].$$

With this choice of $\hat{u}(\omega)$ and the associated weight function $|\hat{u}(\omega)|^2$, we denote the corresponding value of $||Tu||^2$ by J_D. Then

9.47) $$J_D = \lim_{d\to\infty} \int_{-\infty}^{\infty} |T(h\omega)|^2 \left| \frac{e^{-i\omega d} - 1}{\omega} \right|^2 d\omega$$

$$= 2\int_{-\infty}^{\infty} |T(h\omega)|^2 \frac{d\omega}{\omega^2} - 2 \lim_{d\to\infty} \int_{-\infty}^{\infty} |T(h\omega)|^2 \frac{\cos \omega d}{\omega^2} d\omega .$$

Both integrals exist if $T(0)=0$, which we will always assume.

The last integral here tends to zero as $d \to \infty$ as an integration by parts shows.

A straightforward calculation now gives the value of J_D which is

$$9.48) \qquad J_D = 4h\pi \left[\sum_{|j| \leqslant l} |\lambda - j| \, a_j - \sum_{k=1}^{2l} k \left(\sum_{j_1 - j_2 = k} a_{j_1} a_{j_2} \right) \right].$$

We now minimize J_D over the vectors a and subject to the constraints of consistency (9.6). The minimum occurs at the following value of a.

$$9.49) \qquad a_j = \begin{cases} \lambda + 1 - j \,, & j - 1 \leqslant \lambda \leqslant j \,, \\ -\lambda + 1 - j \,, & j \leqslant \lambda \leqslant j + 1 \\ 0 \,, & \text{otherwise}, \quad |j| \leqslant l \end{cases}$$

These $a_j(\lambda)$ are the translates of the cardinal spline of order unity. From (9.49) we see that only those coefficients corresponding to mesh points which immediately surround the characteristic of the differential equation, (9.1), which passes through the forward time point, $(x, (n+1)\Delta t)$ are non-zero. Notice also that the a_j given in (9.49) are non-negative. Thus this most accurate scheme is always stable.

Systems

The procedure of § 9.4 for a system may be carried over to the case of discontinuous data at hand. The details are quite similar and we merely display the following analogue of (9.40).

$$\begin{aligned} B_{-1} &= \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{c}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} - \frac{1}{4}\begin{pmatrix} \alpha & \alpha \\ \beta & \beta \end{pmatrix}, \\ 9.50) \qquad B_0 &= \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{c}{2}\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \alpha & \alpha \\ \beta & \beta \end{pmatrix}, \\ B_1 &= \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{c}{4}\begin{pmatrix} 1 & -3 \\ -3 & 1 \end{pmatrix} - \begin{pmatrix} \alpha & \alpha \\ \beta & \beta \end{pmatrix}. \end{aligned}$$

Casting B_j into the form $P_j = a_j I + b_j A$, $j = 0, \pm 1$ where a_j and b_j are scalars and I is the 2×2 identity matrix, we find that $\alpha = \beta$ and

$$P_{-1}(A) = \frac{1+c-\alpha}{4} I - \frac{1+c-\alpha}{4c} A$$

9.51)
$$P_0(A) = \frac{1-c+\alpha}{2} I + \frac{1-c-\alpha}{2c} A$$

$$P_1(A) = \frac{1+c-\alpha}{4} I - \frac{1-3c-\alpha}{4c} A.$$

A simple and interesting special case of (9.51) corresponds to setting $\alpha = c$.

REFERENCES

[9.1] Miranker, W.L., "Difference Schemes with Best Possible Truncation Error", Numer. Math. *17* (1971) pp.124-142.

[9.2] Micchelli, C.A., and Miranker, W.L., "Optimal Difference Schemes for Linear Initial Value Problems" SIAM J.Numer. Anal. *10* (1973) pp. 983.1009.

[9.3] Micchelli, C.A., and Miranker, W.L., "Asymptotically Optimal Approximation in Fractional Sobolev Spaces and the Numerical Solution of Differential Equations", Numer. Math. 22 (1974) pp.75-87.

[9.4] Strang, G., "Trigonometric Polynomials and Difference Methods of Maximum Accuracy", J. Math. and Phys. *41* (1962) pp.147-154.

§ 10. METHODS OF BOUNDARY LAYER TYPE

10.1 The Idea of the Method

The generic initial value problem for a singularly perturbed system of differential equations may be written in the following form:

10.1)
$$\frac{dx}{dt} = f(t,x,y,\varepsilon), \qquad x(0) = \xi,$$
$$\varepsilon\frac{dy}{dt} = g(t,x,y,\varepsilon), \qquad y(0) = \eta,$$

where $x(t)$, $f \in R^m$ and $y(t)$, $g \in R^n$. f and g depend regularly on ε and $g(t,x,y,0) \neq 0$.

We observe that this class of systems are stiff. For example, in the case that $f=y$ and $g=x+y$, the eigenvalues of the system are $\varepsilon^{-1}+O(1)$ and $-1+O(\varepsilon)$. In a sense the smaller is ε, the stiffer is the system. Thus the large collection of analytic methods, commonly called boundary layer methods, used to characterize solutions of singularly perturbed systems, should be exploited to generate numerical methods for stiff systems. Since the approximations produced by these analytic methods improve with decreasing ε, we may expect that the numerical methods will likewise improve with increasing stiffness in the system.

We will refer to numerical methods developed according to this idea as numerical methods of boundary layer type.

10.2 The Boundary Layer Formation

We begin with a review of the formalism of boundary layers. The solutions $x(t)$ and $y(t)$ of (10.1) have expansions of the type

10.2) $$x(t) \sim \sum_{r=0}^{\infty} x_r(t)\frac{\varepsilon^r}{r!} + \sum_{r=0}^{\infty} X_r(\tau)\frac{\varepsilon^r}{r!}$$

10.3) $$y(t) \sim \sum_{r=0}^{\infty} y_r(t)\frac{\varepsilon^r}{r!} + \sum_{r=0}^{\infty} Y_r(\tau)\frac{\varepsilon^r}{r!} ,$$

where

10.4) $$\tau = t/\varepsilon .$$

The symbol $\sim$, is used to denote the fact that the series in (10.2) and (10.3) are asymptotic expansions. The first and second sums in (10.2) and (10.3) are called the outer solution and the boundary layer respectively.

Following well known procedures (c.f. [10.2] and [10.4]) we find that the coefficients $\{x_r, y_r\}$ of the outer solutions are determined from

10.5)$_0$ $$\begin{aligned} \dot{x}_0 &= f(t,x_0,y_0,0) \\ 0 &= g(t,x_0,y_0,0) \end{aligned}$$

10.5)$_r$ $$\begin{aligned} \dot{x}_r &= f_x(t,x_0,y_0,0)x_r + f_y(t,x_0,y_0,0)y_r + Q_r \\ \dot{y}_{r-1} &= g_x(t,x_0,y_0,0)x_r + g_y(t,x_0,y_0,0)y_r + R_r \\ r &= 1,2,\ldots \end{aligned}$$

The dot represents d/dt. f_x denotes the $m \times m$ matrix whose ij-th component is the derivative of the i-th component of f with respect to the j-th component of x. f_y, g_x and g_y are similarly defined. Q_r and R_r depend on $t, x_0, y_0, \ldots, x_{r-1}, y_{r-1}$, $r=1,2,\ldots$. In particular

10.6)
$$Q_1 = f_\varepsilon(t, x_0, y_0, 0)$$
$$R_1 = g_\varepsilon(t, x_0, y_0, 0).$$

The subscript ε denotes $\partial/\partial\varepsilon$.

Notice that for each $r=0,1,2,\ldots$, the first equation in $(10.5)_r$ represents a system of differential equations, while the second represents a system of finite equations.

Continuing to follow well known procedures, we find the following equations:

$10.7)_0$
$$X_0' = 0$$
$$Y_0' = g(0, x_0(0)+X_0, y_0(0)+Y_0, 0)$$

$10.7)_r$
$$X_r' = p_r$$
$$Y_r' = g_x(0, x_0(0)+X_0, y_0(0)+Y_0, 0)X_r + $$
$$+ g_y(0, x_0(0), y_0(0)+Y_0, 0)Y_r + q_r$$
$$r = 1, 2, \ldots,$$

from which the coefficients $\{X_r, Y_r\}$ of the boundary layer are determined. The prime represents $d/d\tau$. p_r and q_r depend only on $\tau, x_0(0), y_0(0), \ldots, x_{r-1}(0), y_{r-1}(0), X_0, Y_0, \ldots, X_{r-1}, Y_{r-1}$, $r=1,2,\ldots$ In particular

10.8)
$$p_1(\tau) = f(0, \xi, y_0(0)+Y_0, 0) - f(0, \xi, y_0(0), 0).$$

Supplementing the equations $(10.5)_r$ and $(10.7)_r$ for the x_r, y_r, X_r and Y_r is the set of initial conditions:

10.9)
$$x_r(0) + X_r(0) = \xi\delta_{r0},$$
$$y_r(0) + Y_r(0) = \eta\delta_{r0}, \qquad r = 0, 1, \ldots,$$

where δ_{r0} is the Kronecker-δ. The determination of the expansion is still not complete, requiring yet the following procedure for distributing the underdetermined initial conditions (10.9).

(i.e., there is one condition for each pair of variables).

We require that the X_r, Y_r be boundary layers; namely that

10.10)
$$\lim_{\tau\to\infty} X_r(\tau) = \lim_{\tau\to\infty} Y_r(\tau) = 0.$$

Now the specification of the coefficients in the expansions is complete, and we determine them in ordered groups of four; $\{X_r, x_r, y_r, Y_r\}$, $r=0,1,\dots,$ as follows:

From (10.5), (10.7), (10.9) and (10.10) we have for $r=0$

10.11)
$$\begin{array}{lll} \text{a)} & X_0' = 0, & \lim_{\tau\to\infty} X_0 = 0 \\ \text{b)} & \dot{x}_0 = f(t, x_0, y_0, 0), & x_0(0) = \xi \\ \text{c)} & 0 = g(t, x_0, y_0, 0), & \\ \text{d)} & Y_0' = g(0, \xi, y_0(0)+Y_0, 0), & Y_0(0) = \eta - y_0(0). \end{array}$$

(10.11a) has the solution $X_0=0$, and the succeeding equations uniquely determine x_0, y_0 and Y_0. The condition (10.10) for Y_0 is satisfied if the eigenvalues of g_y, denoted $\lambda(g_y)$, satisfy

10.12)
$$\lambda(g_y) < 0.$$

Note: This condition (10.12) characterizes the class of stiff systems to which the methods which we are now discussing are designed to be applied.

We henceforth assume that (10.12) holds.

Similarly, for $r=1$, we have

10.13)
$$\begin{array}{lll} \text{a)} & X_1' = p_1(\tau), & \lim_{\tau\to\infty} X_1(\tau) = 0 \\ \text{b)} & \dot{x}_1 = f_x x_1 + f_y y_1 + f_\varepsilon, & x_1(0) = -X_1(0) \\ \text{c)} & \dot{y}_0 = g_x x_1 + g_y y_1 + g_\varepsilon, & \\ \text{d)} & Y_1' = g_x X_1 + g_y Y_1 + q_1 & Y_1(0) = -y_1(0). \end{array}$$

To solve (10.13) we proceed as follows. From (10.13a) we get

$$X_1(0) = -\int_0^\infty p_1(\sigma)\,d\sigma.$$

This and (10.9) determine $x_1(0) = -X_1(0)$ so that (10.13b) and

(10.13c) may be solved simultaneously for x_1 and y_1. Then (10.13d) may be solved for Y_1. This procedure may now be repeated for each $r=2,3,\ldots$.

10.3 The Boundary Layer Numerical Method

We describe a numerical method which consists of constructing the formal boundary layer expansion by solving the equations determining its terms numerically.

Let $h>0$ be a mesh increment. Let $z=(x,y)^T$ and $Z=(X,Y)^T$ be $N=m+n$ vectors. Then from (10.2) and (10.3)

10.14) $z(h)=z_0(h)+\varepsilon z_1(h)+Z_0(h/\varepsilon)+\varepsilon Z_1(h/\varepsilon)+O(\varepsilon^2).$

Since the equations are stiff we are interested in the case

10.15) $$h \gg \varepsilon.$$

This and condition (10.10) imply that $Z_0(h/\varepsilon)$ and $Z_1(h/\varepsilon)$ will be near zero. In fact these terms will in general be exponentially small in h/ε. Thus we approximate $z(h)$ by $z_0(h)+\varepsilon z_1(h)$, the approximation being $O(\varepsilon^2)$ (i.e. it improves with increasing stiffness). The numerical method consists of calculating $z_0(h)$ and $z_1(h)$. We must still compute Z_0 in order to obtain the initial condition $x_1(0)$, required for the determination of $z_1(h)$. (Of course more terms in the expansion may be calculated if they are wanted).

The numerical method consists of the following steps (i)-(iv):

(i) Solve

10.16) $$\begin{aligned} &\text{a)}\quad \dot{x}_0=f(t,x_0,y_0,0), \qquad x_0(0)=\xi \\ &\text{b)}\quad 0=g(t,x_0,y_0,0) \end{aligned}$$

for $x_0(h)$, $y_0(0)$ and $y_0(h)$. The numerical method for solving (10.16a) should be of the self starting type.

(ii) Having determined $y_0(0)$ in step (i), solve

10.17) $$Y_0'=g(0,\xi,y_0(0)+Y_0,0), \qquad Y_0(0)=\eta-y_0(0)$$

for $Y_0(\tau)$, $\tau \geqslant 0$. This must be done for a net of τ-values, say $\{0,k,2k,\ldots,Mk\}$, so that

10.18) $$x_1(0) = -X_1(0) = \int_0^\infty p_1(\sigma)d\sigma$$

can be approximated to some prescribed degree of accuracy by a quadrature rule:

(iii)

10.19)

$$\xi_1 = \sum_{j=0}^{M} a_j p_1(jk) = \sum_{j=0}^{M} a_j [f(0,\xi,y_0(0)+Y_0(jk),0)-f(0,\xi,y_0(0),0)]$$

(iv) Having determined ξ_1, the approximation to $x_1(0)$, in step (iii), solve

10.20)

a) $\dot{x}_1 = f_x(t,x_0,y_0,0)x_1+f_y(t,x_0,y_0,0)y_1+f_\varepsilon(t,x_0,y_0,0), x_1(0)=\xi_1$

b) $y_1 = -g_y^{-1}(t,x_0,y_0,0)[g_x(t,x_0,y_0,0)x_1-\dot{y}_0+g_\varepsilon(t,x_0,y_0,0)]$

for $x_1(h)$ and $y_1(h)$.

Comment: Steps (i) and (iv) determine $z_0(h)$ and $z_1(h)$ respectively. Steps (ii) and (iii) deal with Z_0 and are used to determine the initial condition, ξ_1, for x_1. The method seems to step across the rapidly varying modes (the boundary layers) as they change over the comparatively great interval $(0,h)$. This is not quite true, nor is it accomplished without cost. Steps (ii) and (iii) perform a mesh calculation with increment k in τ. Since $\tau=t/\varepsilon$, k will be $hO(\varepsilon)$. Thus, in order to calculate Z_0 and $x_1(0)$, a fine mesh calculation must be performed. The critique of this boundary layer method is:

a) the parts or aspects of the given initial value problem upon which to perform the fine mesh calculation are a well defined subpart of the original system.

b) this fine subpart may be calculated with less precision than the coarse part (step (i)). To see that this is so, note that $z_1(h)$ depends on the fine part of the calculation through $x_1(0)$. Thus an error in determining the fine part leads to a proportional error in $z_1(h)$. But the approximation to the solution is $z_0(h)+\varepsilon z_1(h)$. Thus the effect of such an error is reduced in order by the stiffness. Thus here again the stiffer the system, the less precision needed in the fine part calculation, *Remark 10.1:* In Section 10.5 we will show how to eliminate this fine mesh calculation.

10.4 An Example

We will now consider an example for which the steps of the numerical method may be carried out analytically, (i.e., to infinite precision).

The example consists of the following initial value problem:

10.21)
$$\dot{x} = y - x, \qquad x(0) = \xi$$
$$\dot{y} = -100y + 1, \qquad y(0) = \eta.$$

The exact solution of this problem is

10.22)
$$x = \frac{1}{100} + \left(\xi + \frac{\eta - \frac{1}{100}}{99} - \frac{1}{100}\right)e^{-t} - \frac{\eta - \frac{1}{100}}{99}e^{-100t}$$
$$y = \frac{1}{100} + \left(\eta - \frac{1}{100}\right)e^{-100t}$$

The steps of the numerical method are the following ones:

(i) Solve

10.23)
$$\dot{x}_0 = y_0 - x_0, \qquad x(0) = \xi$$
$$0 = y_0$$

for $x_0(h)$, $y_0(0)$, and $y_0(h)$. We use Euler's method with increment h in t to solve (10.23). We find

10.24)
$$x_0(h) = (1-h)\xi$$
$$y_0(0) = y_0(h) = 0$$

(ii) and (iii) Solve

10:25)
$$Y_0'(\tau) = -Y_0(\tau), \qquad Y_0(0) = \eta - y_0(0) = \eta$$

on the mesh $\tau_i = ik$, $i = 0, \ldots, M$. Then evaluate

10.26)
$$x_1(0) = \int_0^{\infty} Y_0(\sigma)d\sigma.$$

Using Euler's method with increment k in τ on (10.25) and using the rectangle rule on (10.26), with the upper limit of integration replaced by kM. We find

10.27)
$$x_1(0) = \eta(1-k^{M+1}).$$

(iv) Solve

$$\dot{x}_1 = 1-x_1, \qquad x_1(0) = \eta(1-k^{M+1})$$
$$y_1 = 1.$$

Again using Euler's method with increment h, we find

10.28)
$$\begin{aligned} x_1(h) &= h+(1-h)(1-k^{M+1})\eta \\ y_1(h) &= 1. \end{aligned}$$

Combining (10.24) and (10.28) we find

10.29)
$$\begin{aligned} x(h) &= (1-h)\xi + \varepsilon(h+(1-h)(1-k^{M+1})\eta) \\ y(h) &= \varepsilon. \end{aligned}$$

Identifying ε with 1/100, (10.29) becomes

10.30)
$$\begin{aligned} x(h) &= \frac{1}{100} + (1-h)\left(-\frac{1}{100} + \xi + \frac{1}{100}(1-k^{M+1})\eta\right) \\ y(h) &= \frac{1}{100} \end{aligned}$$

which approximates (10.22) to the claimed accuracy.

10.5 The ε-independent Method

A criticism of the boundary layer method which we have just discussed is that it depends on the stiff system being given in a form in which there is an identifiable small parameter which characterizes the system as one of singular perturbation type.

To deal with this *criticism* we will now consider how boundary layer methods may be developed even though there is no identifiable small parameter. Then the boundary layer numerical method will be capable of being applied to wider classes of stiff systems.

We proceed by writing $k=(f,g)^T$, $z=(x,y)^T$, and $\xi=(\xi,\eta)^T$. The initial value problem (10.1) is supposedly given in the following form.

10.31) $$\dot{z}=k(t,z;\varepsilon), \qquad z(0)=\zeta.$$

ε here, although displayed, is regarded as unidentifiable. We solve the system (10.31) numerically along the mesh with increment h, proceeding as if the system were not stiff. In terms of the notation in §10.3 we start with m regarded as equal to the number of dimensions, N, in z and with n as equal to zero. Our method then produces $z_0(h)$ by a standard self-starting method. Now we compare $z_0(h)$ and ζ component wise, i.e., we test the following inequality:

10.32) $$\frac{|z_{0,j}(h)-\zeta_j|}{1+|\zeta_j|} > \theta, \qquad j=1,\dots,N.$$

Here θ is a prescribed tolerance. If the tolerance is not exceeded by any component of $z_0(h)$, we accept the value of $z_0(h)$ produced. If the tolerance is exceeded by a set $J(j_1,\dots,j_n)$ of $n>0$ components of $z_0(h)$, we reject the integration step and redo it as follows.

Set

10.33) $$x_i=z_i \qquad \xi_i=\zeta_i \qquad f_i=k_i, \qquad i=1,\dots,N, i\notin J,$$

and set

10.34) $$y_j=z_j \qquad \eta_j=\zeta_j \qquad g_j=k_j, \qquad j=1,\dots,N, j\in J.$$

Now the system has the form

10.35) $$\dot{x} = f(t,z;\varepsilon), \qquad x(0) = \xi$$
$$\dot{y} = g(t,z;\varepsilon), \qquad y(0) = \eta$$

The parameter ε is still unidentifiable, but we make the following assumption:

Assumption 10.1: $f(t,z;\varepsilon)$ and $g(t,z;\varepsilon)$ are analytic in ε in a neighborhood of $\varepsilon=0$ except that $g(t,z;\varepsilon)$ has a simple pole at $\varepsilon=0$. We also maintain the requirement of (10.12) assuming the boundary layer nature of the solution.

We look for a solution of (10.35) in the form

10.36) $$x(t) = x_0(t) + \varepsilon x_1(t) + X_0(\tau) + \varepsilon X_1(\tau) + \ldots$$

10.37) $$y(t) = y_0(t) + \varepsilon y_1(t) + Y_0(\tau) + \varepsilon Y_1(\tau) + \ldots$$

For the outer solution we have

10.38) $$\dot{x}_0 + \varepsilon\dot{x}_1 = f(t,x_0,t_0;\varepsilon) + \varepsilon f_x(t,x_0,y_0;\varepsilon)x_1 + \varepsilon f_y(t,x_0,y_0;\varepsilon)y_1 + \ldots$$

10.39) $$\dot{y}_0 + \varepsilon\dot{y}_1 = g(t,x_0,y_0;\varepsilon) + \varepsilon g_x(t,x_0,y_0;\varepsilon)x_1 + \varepsilon g_y(t,x_0,y_0;\varepsilon)y_1 + \ldots$$

By our assumption, the terms g, g_x and g_y have simple poles at $\varepsilon=0$. Thus from (10.38) and (10.39) we deduce the following equations (10.40) and (10.41) for x_0, y_0, and for εx_1 and εy_1, respectively:

10.40) $$\dot{x}_0 = f(t,x_0,y_0;\varepsilon), \qquad x_0(0) = \xi$$
$$0 = g(t,x_0,y_0;\varepsilon).$$

Notice that we do not set $\varepsilon=0$.

For convenience we will hereafter suppress the arguments $(t, x_0,y_0;\varepsilon)$ of f and g. The equations for εx_1 and εy_1 are

10.41) $$\varepsilon\dot{x}_1 = \varepsilon f_x x_1 + \varepsilon f_y y_1$$
$$\dot{y}_0 = \varepsilon g_x x_1 + \varepsilon g_y y_1.$$

We solve the last equation here for εy_1 as follows:

10.42) $$\varepsilon y_1 = g_y^{-1}[\dot{y}_0 - g_x\varepsilon x_1] = g_y^{-1}[-g_y^{-1}(g_t + g_x f) - g_x\varepsilon x_1].$$

Here we replace $\dot{y}_0$ by its value obtained by differentiating the second equation in (10.40) with respect to t.

Combining (10.41) and (10.42), the equations determining εx_1 and εy_1 are respectively

10.43)
$$\begin{aligned}\varepsilon\dot{x}_1 &= (f_x - f_y g_y^{-1} g_x)\varepsilon x_1 - f_y g_y^{-2}(g_t+g_x f)\\ \varepsilon y_1 &= -g_y^{-1} g_x \varepsilon x_1 - g_y^{-2}(g_t+g_x f).\end{aligned}$$

Notice that ε is still unspecified, but the quantities εx_1 and εy_1 which are sought are, except for the initial condition, $\varepsilon x_1(0)$, well defined. Moreover examining the right members of (10.43) we see by Assumption 10.1 that the large quantities g, g_x and g_y are neutralized, in the sense that they occur as quotients, one of the other.

To determine the initial condition, $\varepsilon x_1(0)$, we obtain an ε-independent determination of the boundary layers. Inserting (10.36) and (10.37) into (10.35), we find

10.44)
$$\begin{aligned}&\varepsilon x_0'(\varepsilon\tau)+\varepsilon^2 x_1'(\varepsilon\tau)+X_0'(\tau)+\varepsilon X_1'(\tau)+\dots\\ &=\varepsilon f(\varepsilon\tau, x_0(\varepsilon\tau)+\varepsilon x_1(\varepsilon\tau)+X_0(\tau)+\varepsilon X_1(\tau)+\dots, y_0(\varepsilon\tau)+\dots;\varepsilon)\\ &\varepsilon y_0'(\varepsilon\tau)+\varepsilon^2 y_1'(\varepsilon\tau)+Y_0'(\tau)+\varepsilon Y_1'(\tau)+\dots\\ &=\varepsilon g(\varepsilon\tau, x_0(\varepsilon\tau)+\varepsilon x_1(\varepsilon\tau)+X_0(\tau)+\varepsilon X_1(\tau)+\dots, y_0(\varepsilon\tau)+\dots;\varepsilon).\end{aligned}$$

Here and hereafter we use the prime to denote differentiation with respect to argument.

Using Assumption 10.1, we deduce the following equations for X_0, X_1, and Y_0 from (10.44).

First

10.45)
$$X_0'(\tau) = 0.$$

As before $X_0(0)=0$, since $X_0(0)+x_0(0)=\xi$, so that $X_0(\tau)\equiv 0$. Next from (10.44) we deduce the following equations for X_1 and Y_0:

10.46)
$$X_1'(\tau) = f(0,\xi,y_0(0)+Y_0(\tau);\varepsilon) - f(0,\xi,y_0(0);\varepsilon),$$

and

10.47)
$$Y_0'(\tau) = \varepsilon g(0,\xi,y_0(0)+Y_0(\tau);\varepsilon).$$

We integrate (10.46) from zero to infinity, using the boundary layer property, $\lim_{\tau\to\infty} X(\tau)=0$. Also using, $x_1(0)+X_1(0)=0$, we get

10.48) $$\varepsilon x_1(0) = \varepsilon \int_0^\infty [f(0,\xi,y_0(0)+Y_0(\tau);\varepsilon) - f(0,\xi,y_0(0);\varepsilon)]\,d\tau.$$

Now since $Y_0(\tau)$ vanishes exponentially fast as τ increases from zero, the bulk of the value of the integral in (10.48) comes from the neighborhood of $\tau=0$. Thus we may expect a good approximation to the integral by replacing the integrand by an interpolant using data at $\tau=0$. This data is first,

10.49) $$Y_0(0) = \eta - y_0(0).$$

from the initial condition, $y_0(0)+Y_0(0) = \eta$, while from (10.47) itself we have

10.50) $$Y_0'(0) = \varepsilon g(0,\xi,\eta;\varepsilon)$$

While we can obtain more data by differentiating (10.47), let us approximate (10.48) using just (10.49) and (10.50). The simplest approximation comes from replacing the integrand in (10.48) by its tangent at $\tau=0$ and integrating this tangent from zero to its positive root. In this manner we obtain from (10.48):

10.51) $$\varepsilon x_1(0) = \frac{1}{2}\,\frac{[f(0,\xi,\eta;\varepsilon)-f(0,\xi,y_0(0);\varepsilon)]^2}{f_y(0,\xi,\eta;\varepsilon)g(0,\xi,\eta;\varepsilon)}.$$

In (10.51) all arithmetic is componentwise except the matrix vector product $f_y g$ in the denominator. Notice that as far as ε is concerned, the dimensions of both sides of (10.51) are in agreement.

A second choice in approximating (10.48) is to use the data (10.49) and (10.51) to fit an exponential to the integrand, and then to integrate the exponential from zero to infinity. In this manner we obtain from (10.48):

10.52) $$\varepsilon x_1(0) = \frac{f(0,\xi,y_0(0);\varepsilon)-f(0,\xi,\eta;\varepsilon)}{f_y(0,\xi,\eta;\varepsilon)g(0,\xi,\eta;\varepsilon)}.$$

The arithmetic here is to be performed exactly as in the previous case.

With either (10.51) or (10.52), (10.43) determines εx_1 and εy_1 completely.

We now solve (10.40) for $y_0(0)$, $y_0(h)$ and $x_0(h)$, by a numerical method as described earlier in 10.3. Then (10.43) and (10.51) or (10.52) are used to solve for $\varepsilon x_1(h)$ and $\varepsilon y_1(h)$ by a numerical method also described earlier. Finally we take

10.53) $$z(h)=\begin{pmatrix} x_0(h)+\varepsilon x_1(h) \\ y_0(h)+\varepsilon y_1(h) \end{pmatrix}$$

We now repeat the procedure on the interval $(h,2h)$. This time we start with the system already divided into a regular and singular part as in (10.35). We then make a tolerance test on $z(2h)$ compared with $z(h)$ analogous to (10.32). If the tolerance is not exceeded by any component of $z(2h)$,we accept the integration step. Otherwise we reject it and redivide the system according to the scheme described above. We then redo this integration step. Once a component is placed into the singular part of the system, we do not remove it, even though its solution settles down and passes the tolerance test. Thus the flow of components of z from x status to y status is one way. If this policy is not followed, the component in question usually regenerates a stiff mode (becomes unstable) at once and it is then pushed back into the singular part anyway. This aspect of the ε-independent numerical method concerning the tolerance test is an algorithmic aspect and should be adjusted to the particular problem being considered. It is likely that for nonlinear systems where the stiffness comes and goes as the solution progresses, a two-directional flow components of z between the regular and singular parts may be called for.

REFERENCES

[10.1] Dahlquist, G., "A Numerical Method for Some Ordinary Differential Equations with Large Lipschitz Constants", IFIP Congress (1968) Supplement pp.132-136.

[10.2] Hoppensteadt, F., "Properties of Solutions of Ordinary Differential Equations with Small Parameters",Comm. Pure and Appl. Math. *XXIV* (1971) pp.807-840.

[10.3] AIKEN, R.C., and Lapidus, L., "An Effective Numerical Integration Method for Typical Stiff Systems", AICHE J. 20 (1974) pp. 368-374.

[10.4] Levin, J. and Levinson, N., "Singular Perturbations of Nonlinear Systems of Differential Equations and Associated Boundary Layer Equation", J. Rational Mech. Anal. *3* (1954) pp.247-270.

[10.5] Miranker, W.L., "Numerical Methods of Boundary Layer Type for Stiff Systems of Differential Equations", Computing *11* (1973) pp.221-234.

§ 11. BOUNDARY LAYER ELEMENTS

11.1 The Model Stiff Boundary Value Problem

The difficulties associated with stiffness for the numerical solution of initial value problems are also present for boundary value problems. In fact, the computational aspects of the latter class of problems may be richer than those of the former (c f. [11.7] and [11.9]). The techniques of boundary layer analysis allow us to take a brief look at some of these aspects and for a model problem, and so we are at a natural place in this course for this small detour away from the initial value problem.

The model problem is

11.1) $$\nu u'' + u' = 0, \qquad x \in (0,1),$$

11.2) $$u(0) = 0, \quad u(1) = 1.$$

Here u is a scalar, ν is a parameter, considered small and the prime denotes differentiation with respect to x.

The exact solution of the model problem is

11.3) $$u(x) = \frac{1-\exp(-x/\nu)}{1-\exp(-1/\nu)},$$

exhibiting a boundary layer near $x=0$.

To discretize (11.1) we introduce a mesh increment, h, $Nh=1$, N a prescribed integer and the following difference operators

$$f_x(x) = \frac{f(x+h)-f(x)}{h},$$

11.4) $$f_{\bar{x}}(x) = f_x(x-h),$$

$$f_{\tilde{x}}(x) = \frac{1}{2}(f_x+f_{\bar{x}}).$$

Then $u^h(x)$ is a numerical approximation to $u(x)$ and is determined as a solution of the following difference boundary value problem

11.5) $$\nu u^h_{x\bar{x}} + u^h_{\underset{\sim}{x}} = 0$$

11.6) $$u^h(0) = 0, \quad u^h(1) = 1.$$

This well known and canonical numerical approach yields the exact solution

11.7) $$u^h(kh) = \frac{1 - \left(\frac{2\nu - h}{2\nu + h}\right)^k}{1 - \left(\frac{2\nu - h}{2\nu + h}\right)^N}, \qquad \nu \neq \frac{h}{2}, \quad k = 0, 1, \ldots, N.$$

As is well known $\lim\limits_{k\to 0} u^h(x) = u(x)$. However, the limit is not taken on uniformly in ν. Indeed if $\nu = h$,

$$u^h(h) = 2^{-1}(1 - 2^{-N})^{-1}$$
$$u\;(h) = (1 - e^{-1})(1 - e^{-N})^{-1},$$

and

$$\lim_{\substack{h\to 0\\ \nu = h}} [u(h) - u^h(h)] = 2^{-1} - e^{-1}.$$

This lack of uniformity is a characterization of the stiffness or ill conditioning of the boundary value problem.

11.2 Methods for Obtaining Uniform Convergence

A first idea for obtaining uniform convergence is due to Ilin (c.f. [11.4]). We note that the differential equation,

$$\nu u_{xx} + a u_x = 0,$$

has the particular solution, $exp(-ax/\nu)$. Then in place of the difference equation (11.5) take

11.8) $$L^h u^h \equiv \gamma u^h_{x\bar{x}} + a u^h_{\underset{\sim}{x}} = 0,$$

where the coefficient γ is determined by the condition that L^h annihilates the particular solution in question, i.e.,

$$L^h \exp(-ax/\nu) = 0.$$

This gives

11.9)
$$\gamma = \frac{ah}{2} \, cth \, \frac{ah}{2\gamma}.$$

Notice that

11.10)
$$\lim_{h\to 0} \gamma = \nu \quad ; \quad \lim_{\nu\to 0} \gamma = |a| \cdot \frac{h}{2}.$$

In [11.4], the following theorem is proved.

Theorem 11.1: Let $u(x)$ be a solution of the boundary value problem

$$\nu u'' + a(x)u' = f(x), \qquad 0 < \nu \leqslant 1$$

$$u(0) = u_0, \quad u(1) = u_1,$$

and let $u^h(x)$ satisfy the same boundary conditions and be a solution of the following difference equation

$$L^h u^h \equiv \frac{a(x)h}{2} \, cth \, \frac{a(x)h}{2\nu} \, u^h_{x\bar{x}} + a(x) u^h_{\tilde{x}} = f(x)$$

on the mesh $\{0, h, 2h, \ldots, Nh=1\}$. Let there exist positive constants α and m such that

$$a(x) \geqslant \alpha, \quad ||a(x)||_{C_2} \leqslant m, \quad ||f(x)||_{C_2} \leqslant m, \quad |u_0|, \quad |u_1| \leqslant m.$$

Then there exists a constant $M = M(m,\alpha)$ independent of ν such that

$$|u^h(x) - u(x)| < Mh,$$

at each point of the mesh.

Another technique for obtaining this uniform convergence is due to Abrahamson, Keller and Kreiss (c.f. [11.5]). They produce the same result as that of Theorem 11.1 and moreover in the case of a second order system. They use a device resembling upstream differencing (c.f. [11.3]) and choose γ in (11.8) to be $\nu+ch$, i.e., they stretch the boundary layer. They show that a best stretching is achieved for $\gamma^*=\nu+|a|h/2$. Notice that γ^* is a linear interpolant of the γ given in (11.9) which uses the limiting values, (11.10), as interpolatory data.

11.3 Finite Elements

The finite element method may be used to produce uniformity of convergence as well, provided that the right elements are used. This approach has the advantage of suggesting a systematic procedure for stiff boundary value problems (c.f. [11.8]).

The finite element approach determines an approximation $v(x)$ to $u(x)$ where

$$v(x) = \sum_{j=0}^{N} \gamma_j \phi_j(x) + \beta b(x). \tag*{11.11)}$$

$\phi_j(x)$ is an element which is associated with each mesh point $x_j=jh$, $j=0,\dots,N$ and $b(x)$ is a so called boundary layer element

$$b(x) = e^{-x/\nu}. \tag*{11.12)}$$

Let $(\cdot,\cdot)$ denote the inner product in $L^2(0,1)$. Then the conditions for determining β and the γ_j, $j=0,\dots,N$ are

$$\begin{aligned}
&\text{i)}\quad 0=(\nu v''+v',\phi_i)=\nu v'\phi_i\Big|_{0}^{1}-\int_0^1 v'(\nu\phi_i'-\phi_i)\,dx, \qquad i=1,\dots,N-1,\\
&\text{ii)}\quad 0=(\nu v''+v',b)=\nu v'b\Big|_0^1+2\int_0^1 v'b\,dx,\\
&\text{iii)}\quad 0=\gamma_0\phi_0(0)+\beta,\\
&\text{iv)}\quad 1=\gamma_N\phi_N(1)+\beta e^{-\frac{1}{\nu}}.
\end{aligned} \tag*{11.13)}$$

(11.13) (i) and (ii) form the statement that v is the weak solution of (11.1) in the span of b and the ϕ_i, $i=1,\dots,N-1$. (11.13) (iii) and (iv) assert that v obeys the boundary conditions (11.2).

Now we specify $\phi_j(x)$ to be the most primitive finite element $\phi_j(x)=\phi(x-jh)$, $j=0,\dots,N$ where

$$\text{11.14)}\qquad \phi(x)=\begin{cases} -\dfrac{x}{h}+1, & x\in(0,1),\\[2ex] \dfrac{x}{h}-1, & x\in(-1,0),\\[2ex] 0, & \text{otherwise.}\end{cases}$$

In this case (11.13) (i) and (ii) become respectively

$$\text{11.15)}\qquad 0=h\left[\nu\frac{\gamma_{j+1}-2\gamma_j+\gamma_{j-1}}{h^2}+\frac{\gamma_{j+1}-\gamma_{j-1}}{2h}\right.$$
$$\left.-\beta\nu^2\frac{e^{-j\frac{h}{\nu}}\,e^{-\frac{h}{\nu}}-2+e^{\frac{h}{\nu}}}{h^2}\right],\qquad j=1,\dots,N-1$$

and

11.16)

$$0=\beta\left[e^{-\frac{2h}{\nu}}-e^{-\frac{1}{\nu}}+e^{-\frac{2}{\nu}}-e^{-2\frac{(1-h)}{\nu}}-2\cosh\frac{2h}{\nu}\sum_{j=1}^{N-1}e^{-\frac{2jh}{\nu}}\right]$$
$$+\frac{2\nu}{h}\left[(\gamma_0-\gamma_1)\left(e^{-\frac{h}{\nu}}-\frac{1}{2}\right)+\sum_{j=1}^{N-1}\left\{(\gamma_{j+1}-2\gamma_j+\gamma_{j-1})e^{-\frac{2h}{\nu}}\right.\right.$$
$$\left.\left.+(\gamma_j-\gamma_{j-1})e^{-\frac{h}{\nu}(j-1)}-(\gamma_{j+1}-\gamma_j)e^{-\frac{h}{\nu}(j+1)}\right\}+(\gamma_N-\gamma_{N-1})e^{-\frac{1}{\nu}}\left(e^{-\frac{h}{\nu}}-\frac{1}{2}\right)\right].$$

Now

$$c=(\beta,\gamma_0,\dots,\gamma_N)\quad\text{and}\quad f=(0,\dots,0,1)$$

be $(N+2)$-vectors and write (11.15), (11.16) and (11.13) (iii) and (iv) in the form

$$\text{11.17)}\qquad Sc=f$$

where S is an $(N+2)\times(N+2)$ matrix. Let $\mu=exp(-h/\nu)$. Then

11.18) $$S = S_0(1+O(\mu^2)),$$

where

11.19) $$S_0 = \begin{bmatrix}
0 & -\frac{3}{2}+2\mu & \frac{3}{2}-4\mu & 2\mu & & & & \\
1 & 1 & 0 & 0 & & & & \\
\nu^2(2\mu-1) & \nu-h/2 & -2\nu & \nu+h/2 & & & & \\
-\nu^2\mu & 0 & \nu-h/2 & -2\nu & \nu+h/2 & & & \\
 & & & \nu-h/2 & -2\nu & \nu+h/2 & & \\
 & & & & \ddots & \ddots & \ddots & \\
 & & & & & \nu-h/2 & -2\nu & \nu+h/2 \\
 & & & & & & & 1
\end{bmatrix}$$

All missing entries in S_0 are zero.

In the limit as $\nu \to 0$, the system $Sc=f$ becomes

11.20) $$\begin{aligned} 0 &= \gamma_1 - \gamma_0 \\ 0 &= \beta + \gamma_0 \\ 0 &= \gamma_{j+1} - \gamma_{j-1} \\ 1 &= \gamma_N \end{aligned}$$

The solution of (11.20) is

11.21) $$-\beta = \gamma_j = 1, \qquad j=0,1,\ldots,N.$$

This is the desired limiting form of the numerical solution as an inspection of (11.3) shows.

11.4 A Numerical Experiment

A computation with $S_0c = f$ was performed with the results displayed in the following table.

N / ε	3	5	7
1	6.9	$3.3\ E^{-1}$	$2.3\ E^{-1}$
.1	$2.9\ E^{-2}$	$4.0\ E^{-2}$	$4.8\ E^{-2}$
.01	$5.1\ E^{-4}$	$8.9\ E^{-4}$	$1.1\ E^{-3}$
.001	$5.9\ E^{-6}$	$1.3\ E^{-5}$	$2.1\ E^{-5}$

l^2-norm of error

The numerical results displayed in this table show remarkable accuracy for very few elements; (the analogue of a course grid). They do reveal a disturbing feature. There is an improvement at first as N increases which is then followed by degradation. An examination of S_0 shows that the finite equations which generate the approximation are not of positive type. Thus as N gets large we may expect some instability. It seems we are not disappointed.

11.5 Some Remarks

We conclude § 11 with several remarks.

Remark 11.1: The finite element procedure is a variant of a well known technique of finding Galerkin approximations to solutions of problems with singularities. For those problems one adjoins to the set of functions in terms of which the Galerkin approximation is sought, special functions having the same singular form as the solution.

Remark 11.2: The special elements needed to be adjoined in the case of a singular perturbation are well known. In fact, the body of literature dealing with these problems has the characterization of the boundary layers as one of its themes (c.f. [11.1] and [11.2]). These elements are known even in the case of systems of equations, nonlinear equations and equations in more than one independent variable. Thus the numerical method outlined here is applicable to all of these classes of problems.

Remark 11.3: The boundary layer element does not have compact support which causes the total loss of sparseness in the stiffness matrix, S. Nevertheless, as we have seen in § 11.3, the exponential decrease of the boundary layer element gives us a stiffness matrix which is sparse up to an error which is exponentially small.

Remark 11.4: A proof of the uniform convergence of the Galerkin approximation, which includes boundary layer elements, to the solution of the singular perturbation problem follows along familiar lines:

a) We choose a sequence of manifolds which contain the right kind of functions to secure a uniform approximation to the solution.

b) The results of the analytic theory of singular perturbations supply us with the functions needed for this uniform approximation.

c) The Galerkin approximation to the solution in each of the manifolds in (a) being itself the best approximation to the solution in each of these manifolds respectively, will have the property of uniform approximation,

d) The fact that S_0 is not of positive type must be dealt with.

Remark 11.5: The projection procedure giving the numerical results supplies us quite simply with a matrix S which is in fact a discrete version of a matching matrix. The latter is a key element in the analytic theory of singular perturbations and is in general very difficult to construct.

REFERENCES

[11.1] R.E. O'Malley, "Introduction to Singular Perturbations", Acad. Press (1974).

[11.2] G.H. Handelman, J.B. Keller and R.E. O'Malley, Loss of boundary conditions in the asymptotic solution of linear ordinary differential equations, *Comm. Pure Appl. Math,* *21* (1968) pp.243-261.

[11.3] F.W. Dorr, An example of ill-conditioning in the numerical solution of singular perturbation problems, *Math Comp.,* 25 (1971).

[11.4] A.M. Il'in, Differencing scheme for a differential equation with a small parameter affecting the highest derivative, *Math Zametki,* 6 (1969), pp.237-248.

[11.5] L.R. Abrahamson, H.B. Keller and H.O. Kreiss, Difference approximations for singular perturbations of ordinary differential equations, *Math. Comp.* (to appear).

[11.6] C.E. Pearson, On a differential equation of boundary layer type, *J. Math and Phys.*, *47* (1968), pp.134-154.

[11.7] G. Strang and G.J. Fix, "An Analysis of the Finite Element Method", Prentice-Hall (1973).

[11.8] Miranker, W.L., and Yun, D.Y., "A Note on Boundary Layer Elements", IBM Research Center Report RC 4990, August, 1974.

[11.9] Miranker, W.L., and Morreeuw, J.O., "Semianalytic Numerical Studies of Turning Points Arising in Stiff Boundary Value Problems", Math. Comp. *28* (1974) pp.1017-1034.

§ 12. EXPONENTIAL FITTING IN THE OSCILLATORY CASE

12.1 Failure of Previous Methods

The numerical methods which we have discussed thus far have used the fact that the rapid changes in the solution are transitory, although possibly recurrent on a time scale which is long compared to that of the rapid changes. When the stiff system has solutions of a highly oscillatory character, the methods which we have looked at do not work at all. For example, the key idea behind the introduction of notions of absolute stability was based on the existence of slowly varying stages in the development of the solution. In this section we begin our considerations of this oscillatory problem with a discussion of a method which employs a form of exponential fitting based on a process called aliasing (c.f. [12.3]).

12.2 Aliasing

Let $f(t)$ be periodic with period 2π. For a fixed integer $N>0$, let the following values of $f(t)$ be given:

$$12.1) \qquad f(t_j), \quad t_j = \left(\frac{j}{2N}\right)2\pi, \qquad j=0,1,\dots,2N.$$

We will call these points t_j, the data points.

In terms of these values, the discrete Fourier series,

$C_N(t)$, of $f(t)$ is

12.2) $$C_N(t) = \frac{A_0}{2} + \sum_{r=1}^{N} (A_r \cos rt + B_r \sin rt) + \frac{A_N}{2} \cos Nt.$$

The coefficients of this series are

12.3) $$A_r = \frac{1}{N} \sum_{j=0}^{2N-1} f(t_j) \cos rt_j$$
$$B_r = \frac{1}{N} \sum_{j=0}^{2N-1} f(t_j) \sin rt_j, \qquad j=0,1,\ldots,N.$$

If $f(t)$ is highly oscillatory, then for a good representation of $f(t)$ by $C_n(t)$ we require N to be quite large. In fact we would need $2N$ values of $f(t)$ (c f. (12.1)) and $2N$ terms in the series (12.2), a large number of values and terms respectively.

Now suppose that $f(t)$ has a special form so that its frequencies fall into clusters. In particular suppose that

12.4) $$f(t) = h(t) + \sum_{m=1}^{p} c_m \cos R_m t + d_m \sin R_m t.$$

We suppose that $h(t)$ is a smooth function. That is

12.5) $$h(t) = \frac{a_0'}{2} + \sum_{r=1}^{\infty} (a_r' \cos rt + b_r' \sin rt)$$

and that there exists an integer $L > 0$ such that the quantities $|a_r'|$ and $|b_r'|$ are negligible for $r > L$. Furthermore we suppose that the frequencies $R_p > R_{p-1} > \ldots > R_1 > L$ are known (and are large).

The objective is to estimate the coefficients c_m and d_m, $m=1,\ldots,p$ and the coefficients a_r' and b_r', $r=1,\ldots,L$. This may be efficiently accomplished through aliasing.

Note that at each of the data points, the functions $\cos R_m t$ and $\sin R_m t$ can be replaced by $\cos r_m t$ and $\sin r_m t$, respectively, where $R_m > N > r_m$. This is accomplished by use of the following identities:

$$\cos[(2q)N+r]t_j = \cos rt_j$$

$$\cos[(2q+1)N+r]t_j = \cos(N-r)t_j$$

$$\sin[(2q)N+r]t_j = \sin rt_j$$

$$\sin[(2q+1)N+r]t_j = -\sin(N-r)t_j .$$

One may view the first of these identities, for example, as the statement that $\cos[(2q)N+r]t$ takes on the same values as $\cos rt$ at the data points but oscillates faster in between. Thus if we use a coarse mesh composed of $2N+1$ mesh points where $N < R_1$, each of the high frequencies R_m will be replaceable by a harmonic with the lower frequency $r_m < N$.

The relation between the Fourier coefficients (a_r, b_r) of $f(t)$ and the coefficients (A_r, B_r) of its finite Fourier series (c f. (12.2)) is

$$A_r = a_r + \sum_{m=1}^{\infty} (a_{2mN+r} - a_{2mN-r})$$

12.6)

$$B_r = b_r + \sum_{m=1}^{\infty} (b_{2mN+r} - b_{2mN-r}).$$

Thus the replacement of higher frequencies by lower ones will not confuse components if N is chosen in such a way that each of the frequencies $\omega = 0, 1, 2, \ldots, L-1$, $R_1, R_2, \ldots, R_p$ occurs in a separate sum in the right member of (12.6). Clearly $N \geqslant L+p$ but usually N is smaller than R_p making the process reasonably efficient.

12.3 An Example of Aliasing

These ideas are clearly illustrated with the following example. Suppose that $f(t)$ is the sum of a slowly varying function plus three harmonics of frequencies 177, 589 and 1000, respectively. Using $N=52$ or 105 data points we have

$$\begin{aligned}
\cos 1000\, t_j &= \cos 40\, t_j \\
\sin 1000\, t_j &= -\sin 40\, t_j \\
\cos 589\, t_j &= \cos 35\, t_j \\
\sin 589\, t_j &= -\sin 35\, t_j \\
\cos 177\, t_j &= \cos 31\, t_j \\
\sin 177\, t_j &= -\sin 31\, t_j
\end{aligned}$$

where $t_j = j\pi/52$, $j=0,1,\ldots,104$.

Thus if we find the discrete Fourier series for $f(t)$ at these data points, viz.,

$$f(t_j) = \frac{A_0}{2} + \sum_{r=1}^{51} (A_r \cos rt_j + B_r \sin rt_j) + \frac{A_{52}}{2} \cos 52\, t_j,$$

we can say that at the data points

$$\begin{aligned}
f(t) = \frac{A_0}{2} &+ \sum_{r=1}^{30} (A_r \cos rt + B_r \sin rt) \\
&+ A_{31} \cos 177\, t - B_{31} \sin 177\, t \\
&+ A_{35} \cos 589\, t - B_{35} \sin 589\, t \\
&+ A_{40} \cos 1000\, t - B_{40} \sin 1000\, t
\end{aligned}$$

within an error depending on the size of the Fourier coefficients of the slowly varying part of $f(t)$. For a precise error analysis of this procedure we refer to [12.4].

12.4 Application to Highly Oscillatory Systems

We begin by describing the method of Certaine (c.f. [12.3] and [12.4]) which is a simpler variant of that of Jain treated in § 5.

The system of differential equations is written in the form

12.7) $$y'(x) = -Dy(x) + g(y(x), x).$$

Here y and g are m-vectors and D is an m × m constant matrix with at least one large eigenvalue. We integrate (12.7) to obtain

$$12.8)\ y(x_{m+1}) = e^{-Dh}y_n + \int_{x_n}^{x_{n+1}} e^{D(x-x_{n+1})}g(y,x)dx, \quad h = x_{n+1}-x_n.$$

Certaine's method consists of the following two steps.

i) Approximate $g(y,x)$ by an interpolation polynomial, $g_k(x)$, of degree k at the points $x_{n-k}, x_{n-k+1}, \dots, x_n$. Replace g in (12.8) by g_k and use the resulting expression for $y(x_{n+1})$ as a predictor.

ii) Using the predicted value of $y(x_{n+1})$ repeat step (i) using the points $x_{n-k+1}, \dots, x_{n+1}$ to determine the correction.

Thus Certaine's method is given by two utilizations of the following expression

$$12.9) \qquad y_{n+1} = e^{-Dh}y_n + e^{-Dx_{n+1}} \int_{x_n}^{x_{n+1}} e^{Dx} g_k(x)dx.$$

We now make several observations about Certaine's method.

Remark 12.1: The integral in (12.9) may be evaluated explicitly. If the exponential matrix e^{-D} is difficult to evaluate one may take $D=D_1+D_2$ where e^{-D_1} is easy to evaluate and e^{D_2x} is adjoined to g.

Remark 12.2: If g is a polynomial of order less than $k+1$ the method is exact. Thus the method is A-stable.

In the oscillating case the polynomial g_k is replaced by a trigonometric polynomial. In this case as well, the integral in (12.9) may be explicitly evaluated. However, we will have an inefficient procedure unless we use aliasing. That is we must know the higher frequencies in the problem (i.e., the large imaginary eigenvalues of D) and then we must alias these higher frequencies so that g_k is a trigonometric polynomial of low degree.

A criticism of this method arises in the case of a nonlinear system. For in such a system, even though the frequencies are known to start with, we may find the introduction of sum and difference frequencies into the solution as it develops. Of course the determination of N depending on L and the R_j, $j=1,\dots,p$ requires a computation also.

REFERENCES

[12.1] Certaine, J., "The Solution of Ordinary Differential Equations with Large Time Constants", in Mathematical Methods for Digital Computers, A. Ralston and H.S. Wilf (Editors), Wiley, New York (1960) pp.128-132.

[12.2] Guderly, K.G. and Hsu, C.C., "A Predictor-corrector Method for a Certain Class of Stiff Differential Equations", Math. Comp. *26* (1972) pp.51-69.

[12.3] Snider, A.D. and Fleming, G.C.,"Approximation by Aliasing with Applications to 'Certaine' Stiff Differential Equations", Math. Comp. *28* (1974) pp.465-473.

[12.4] Snider, A.D, "An Improved Estimate of the Accuracy of Trigonometric Interpolation", SIAM J. Numer. Anal. *9* (1972) pp.505-508.

§ 13. A TWO-TIME METHOD FOR THE OSCILLATORY PROBLEM

13.1 The Model Problem

We continue our study of the highly oscillatory problem through use of the two time technique of singular perturbation theory. To illustrate this approach we consider the following model problem:

13.1)
$$\varepsilon \frac{du}{dt} = (A_0+\varepsilon A_1)u \ , \qquad t\in(0,T] ,$$
$$u(0) = u_0 \ ,$$

where u is an n-vector and A_0 and A_1, are $n \times n$ matrices.

In terms of the matrizant $\Psi(t,\varepsilon)$, we may write the solution of the model problem as

13.2)
$$u = \Psi(t,\varepsilon)u_0,$$
$$\Psi(t,\varepsilon) = exp[(A_0+\varepsilon A_1)t/\varepsilon] .$$

The numerical evaluation of this matrix at $t=T$, is difficult when ε is near zero.

If we introduce a new time scale

13.3) $$\tau = t/\varepsilon ,$$

the solution becomes

13.4) $$u = e^{A_0\tau + A_1 t}u_0, \qquad 0 \leqslant \tau \leqslant T/\varepsilon .$$

This indicates that the solution developes on two different time scales, t called the slow time and τ called the fast time. If A_0 and A_1 commute, (13.4) becomes

13.5) $$u = e^{A_0\tau}e^{A_1 t}u_0 .$$

In this case the dependence on the two scales separates and in principle we could determine each of the factors in (13.5) separately.

However in general A_0 and A_1 don't commute and moreover it is not necessarily the case that the development of the solution on the τ-scale is even useful to approximate numerically.

13.2 Numerical Solution Concept

Consider the example corresponding to

$$A_0 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \qquad A_1 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} .$$

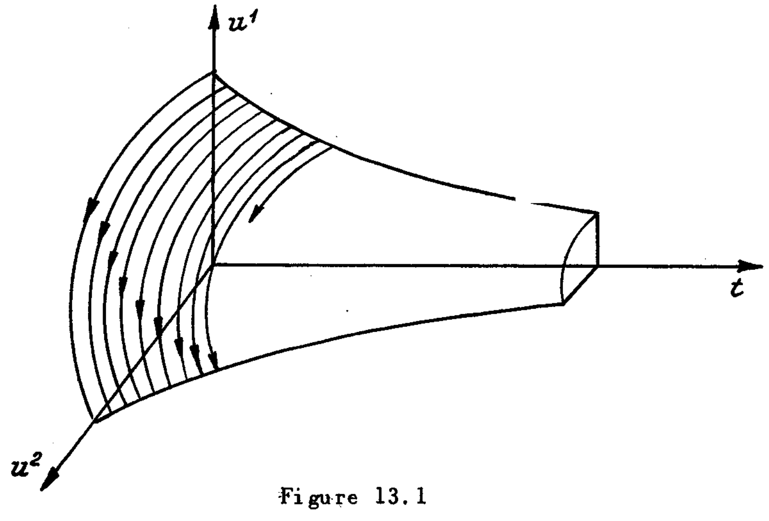

Figure 13.1

With these matrices the motion described by (13.1) corresponds to a slowly damped (t-scale) extremely rapid (τ-scale) harmonic motion. The solution, schematized in figure 13.1 for the case $n=2$, is practically a space filling curve.

As $\varepsilon \to 0$ the solution converges (in an approximate sense) to the cone obtained by rotating the curve $||u_0|| e^{-t}$ about the t-axis. Thus the meaningfulness of describing a trajectory by a set of its values on the points of a mesh is lost (i.e. is an ill conditioned process).

A variety of alternate numerical solution concepts may be formulated. Consider the following one:

Solution concept: Given $\varepsilon' > 0$ and $\delta > 0$, we say that $U(t)$ is an (ε', δ') (numerical) approximation to $u(t)$ if there exists τ with $|\tau| \leqslant \delta$ such that

$$|U(t)-u(t+\tau)| \leqslant \varepsilon'.$$

Of course $\delta=0$ for the usual concept of (numerical) approximation. In figure 13.2 an example of this approximation is given.

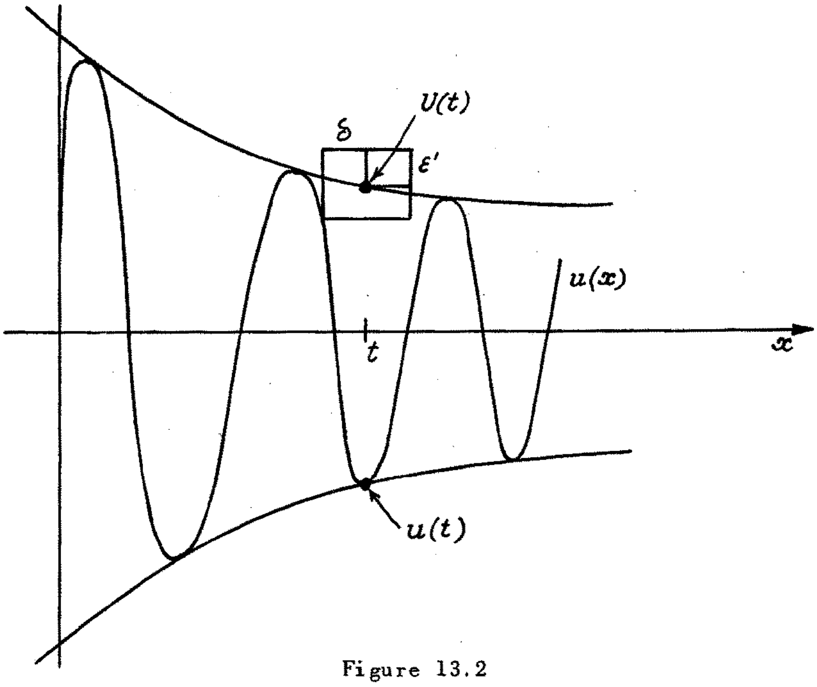

Figure 13.2

In terms of the model problem, we may accept by means of this solution concept a numerical approximation to the slow time part of the solution as a numerical approximation to the solution itself. The problem is to extract this part out of the whole solution and to do this we employ the method of two times. Of course nothing prevents us from computing the fast time part, as we shall see, locally. That is to remove the ill-conditioning of the highly oscillatory problem we must abandon some aspect of the solution and in particular we will abandon the determination of its precise (fast-time) phase.

13.3 The two time expansion

We seek approximations to the solution of (13.1) in the form of a general two-time expansion

$$u = \sum_{r=0}^{\infty} u_r(t,\tau)\varepsilon^r. \tag{13.6)}$$

This will be a useful series for purposes of approximation, if we have

$$u_r(t,t/\varepsilon)\varepsilon^r = o(\varepsilon^{r-1}), \qquad r=1,2,\ldots, \tag{13.7)}$$

as $\varepsilon \to 0$, uniformly for $0 \leqslant t \leqslant T$. With (13.7) valid we say that (13.6) is an asymptotic expansion with asymptotic scale ε^r. A sufficient condition for (13.7) is that

$$u_r(t,\tau) = o(\tau) \tag{13.8)}$$

as $\tau \to \infty$ for $r=1,2,\ldots$.

The expansion resulting from this prescription of the form (13.6)-(13.8) of the solution will be derived below. It is sometimes possible to obtain more information from the expansion by placing a stronger condition on the coefficients than (13.8). In particular we will determine conditions on A_0 and A_1 so that the requirement

$$u_r(t,\tau) = o(\tau e^{A_0 \tau}) \tag{13.9)}$$

as $\tau \to \infty$ for $r=1,2,\ldots$, can be used to obtain a valid expansion.

If A_0 is an oscillatory matrix (all eigenvalues have zero real part), then conditions (13.8) and (13.9) are equivalent. If A_0 is a stable matrix (all eigenvalues have negative real parts), the condition (13.9) is more restrictive than (13.8). In the stable case it may not be possible to obtain an expansion of the solution of (13.1) in the form (13.6) whose coefficients satisfy either (13.8) or (13.9). However, we will describe another restriction on the problem which when used with (13.9) guarantees that the solution of (13.1) can be approximately solved in the form (13.6). This approximation technique proceeds via the twotime approach. This result is valid when the eigenvalues of A_0 lie in the stable half plane; therefore, it contains both the stable and oscillatory cases. In the stable case, the expansion found by this method reduces to the one which would be obtained by the method of matched asymptotic expansions. In the oscillatory case, this result reduces to an expansion equivalent to the one obtained by the Bogoliubov method of averaging.

13.4 Formal Expansion Procedure

We consider the initial value problem for the system (13.1) and we write the initial conditions in the form

13.10) $$u(0) = \sum_{r=0}^{\infty} a_r \varepsilon^r$$

To simplify computation let

13.11) $$v(t,\tau) = e^{-A_0 \tau} u(t,\tau).$$

Since v is considered as a function of the two variables τ and $t = \varepsilon\tau$,

13.12) $$\frac{dv(\varepsilon t,\tau)}{dt} = \varepsilon \frac{\partial v(t,\tau)}{\partial t} + \frac{\partial v(t,\tau)}{\partial \tau}.$$

Then (13.1) becomes the following equation for v:

13.13) $$\varepsilon \frac{\partial v}{\partial t} + \frac{\partial v}{\partial \tau} = \varepsilon B(\tau) v, \qquad v(0) = \sum_{r=0}^{\infty} a_r \varepsilon^r$$

where

13.14) $$B(\tau) = e^{-A_0 \tau} A_1 e^{A_0 \tau}.$$

We seek a solution in the form (13.6) which becomes

13.15) $$v = \sum_{r=0}^{\infty} v_r(t,\tau)\varepsilon^r$$

subject to the condition (13.9) on the u_r. In terms of the v_r, the latter becomes

13.16) $$v_r(t,\tau) = o(\tau) \quad \text{as} \quad \tau \to \infty, \qquad r = 0,1,\ldots .$$

Substituting (13.15) into (13.13) and equating coefficients of the like powers of ε gives

13.17) $$\frac{\partial v_r}{\partial \tau} = B(\tau)v_{r-1} - \frac{\partial v_{r-1}}{\partial \tau}, \qquad v_r(0,0) = a_r, \qquad r = 0,1,\ldots .$$

Here $v_{-1} \equiv 0$.

The problem (13.17) is underdetermined. The equation (13.17) for v_r can be integrated to give

13.18) $$v_r(t,\tau) = \tilde{v}_r(t) + \int_0^{\tau} \left[B(\sigma)v_{r-1}(t,\sigma) - \frac{\partial v_{r-1}(t,\sigma)}{\partial t} \right] d\sigma, \qquad r = 0,1,\ldots,$$

where

13.19) $$\tilde{v}_r(0) = a_r.$$

Except for (13.19), $\tilde{v}_r(t)$ is arbitrary. Differentiating (13.18) with respect to t gives

13.20) $$\frac{\partial v_r}{\partial t} = \frac{\partial \tilde{v}_r}{\partial t} + \int_0^{\tau} \left[B(\sigma)\frac{\partial v_{r-1}}{\partial t} - \frac{\partial^2 v_{r-1}}{\partial t^2} \right] d\sigma .$$

Combining this with (13.18) gives

13.21) $$v_r(t,\tau) = \tilde{v}_r(t) + \tilde{v}_{r-1}(t) \int_0^{\tau} B(\sigma)d\sigma - \tau \frac{d\tilde{v}_{r-1}}{dt} + \int_0^{\tau} R_{r-1}(t,\sigma)d\sigma$$

where

$$R_r(t,\sigma) = -\int_0^{\sigma}\left[B(\sigma)\,\frac{\partial v_{r-1}(t,\sigma')}{\partial t} - \frac{\partial^2 v_{r-1}(t,\sigma')}{\partial t^2}\right]d\sigma'$$

13.22)

$$+B(\sigma)\int_0^{\sigma}\left[B(\sigma')v_{r-1}(t,\sigma') - \frac{\partial v_{r-1}(t,\sigma')}{\partial t}\right]d\sigma'.$$

(13.21) and (13.22) hold for $r=0,1,\ldots,$ where $\tilde{v}_{-1}\equiv R_0\equiv 0$. Let us impose the growth condition (13.16) in (13.21). To do this divide (13.21) by τ and take the limit as $\tau\to\infty$. This results in the following condition for $\tilde{v}_{r-1}$.

13.23) $$\frac{d\tilde{v}_{r-1}}{dt} = \left(\lim_{\tau\to\infty}\frac{1}{\tau}\int_0^{\tau}B(\sigma)d\sigma\right)\tilde{v}_{r-1} + \lim_{\tau\to\infty}\left(\frac{1}{\tau}\int_0^{\tau}R_{r-1}(t,\sigma)d\sigma\right),$$
$$r = 0,1\ldots .$$

When these limits exist, (13.23) along with (13.19) determine $\tilde{v}_r,\ r=0,1,\ldots.$

This approach dependes critically on the existence of the limits in (13.23). The development will be simplified by using the notation

13.24) $$\bar{f} = \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^{\tau}f(x)dx.$$

If $\bar{f}$ exists we will call it the average of f. In terms of this notation (13.23) becomes

13.25) $$\frac{d\tilde{v}_r}{dt} = \bar{B}\tilde{v}_r + \bar{R}_r(t),\quad \tilde{v}_r(0) = a_r,\qquad r = 0,1,\ldots,$$

provided the averages exist.

13.5 Comments on the existence of the average and estimates of the remainder

In the case that A_0 is an oscillatory matrix, B is an almost periodic function (cf. (13.14)) and so the existence of the average, $\bar{B}$ is assured. The existence of $\bar{R}_1$ is implied by this existence of $\bar{B}$. These statements are proved in [13.1].The existence of these two averages provides us with the approximation $v_0+\varepsilon v_1$ to v. This approximation is adequate for our computational purposes. In [13.2], Hoppensteadt and Miranker develope a more complete treatment of (13.1) by the two time method in the general case where the eigenvalues of A_0 may be anywhere in the complex plane and where nonlinear forcing terms are adjoined to the system as well. However we restrict our descriptions tq the setting of the earlier paper of these two authors since that description of the results, being less technical, is easier to present as is the ensuing numerical development.

Under the hypotheses that the matrix A_0 has simple elementary divisors and in the case that the eigenvalues λ_i, $i=1,\dots,n$ of A_0 are such that $Re\ \lambda_i \leqslant 0$, we find the following three results in that earlier paper.

Theorem 13.1: $\bar{B}$ exists if and only if the elements a^1_{ij} of A_1 vanish whenever $Re(\lambda_j-\lambda_i) > 0$.

Theorem 13.2: $\bar{R}_1$ exists whenever $\bar{B}$ exists.

Theorem 13.3:

$$\max_{0\leqslant\tau\leqslant T/\varepsilon} |v(\varepsilon\tau,\tau)-v_0(\varepsilon\tau)-\varepsilon v_1(\varepsilon\tau,\tau)| \leqslant const\ \varepsilon^2.$$

13.6 The Numerical algorithm

We take the leading term, $u_0(t,\tau)$ of the expansion (13.6) as approximation to the solution of the initial value problem (13.1) with the initial condition given by (13.10).

Then from (13.11) and (13.18)

13.26) $$u_0(t,\tau) = \Phi(\tau)\tilde{v}_0(t).$$

$\Phi(\tau)$ is the fundamental matrix given by

13.27) $$\Phi_\tau = A_0\Phi, \qquad \Phi(0) = I,$$

while from (13.25)

13.28) $$\frac{d\tilde{v}_0}{dt} = \bar{B}\,\tilde{v}_0, \qquad \tilde{v}_0(0) = a_0.$$

From (13.14)

13.29) $$\bar{B} = \lim_{T\to\infty}\frac{1}{T}\int_0^T \Phi^{-1}(\sigma)A_1\,\Phi(\sigma)d\sigma.$$

We describe the algorithm for replacing a_0 the approximation to $u(0)$ by $U(h)$ the approximation to $u(h)$ (in the sense of the solution concept in section 13.2 above). The algorithm is to be repeated approximating $u(t)$ at $u(2h),\dots,u(nh)$ successively.

Algorithm

i) Solve (13.27) on a mesh of increment k in the τ scale by some self starting numerical method, obtaining the sequence $\Phi(jk)$, $j=0,\dots,N$.

ii) Using the values $\Phi(jk)$ obtained in (i), approximate $\overline{B}$ by truncating the limit of integration τ and replacing the integral in (13.29) by a quadrature formula, say

$$\overline{B} = \frac{1}{N} \sum_{j=0}^{N} C_k \Phi^{-1}(jk) A_1 \Phi(jk).$$

The integer N is determined by a numerical criterion which assures that the elements of the matrix $\overline{B}$ are calculated to some desired accuracy.

iii) With $\overline{B}$ (approximately) determined in ii), solve (13.28) for $\tilde{v}_0(h)$ by some self starting numerical method.

iv) Compute $u_0(h,Nk)=\Phi(Nk)\tilde{v}_0(h)$ and take this as the approximation to $u(h)$.

Refinement: The method may be refined by adding an approximation of $\varepsilon v_1(h,h/\varepsilon)$ to $\tilde{v}_0(h)$ prior to multiplication by $\Phi(Nk)$ (step (iv)). This approximation in turn is determined from a numerical solution of the equations defining $v_1(t,\tau)$; viz.

$$v_1(t,\tau) = \tilde{v}_1(t) - \overline{B}\sigma + \int_0^\tau B(\sigma)d\sigma$$

13.30) $$\frac{d\tilde{v}_1}{dt} = \overline{B}\tilde{v}_1 + \overline{R}_1(t)$$

$$R_1(t,\sigma)=\left[(\overline{B}^2-B(\sigma)\overline{B})\sigma-\int_0^\sigma B(\sigma')d\sigma'\overline{B}+B(\sigma)\int_0^\sigma B(\sigma')d\sigma'\right]\tilde{v}_0(t)$$

In figure 13.3 we schematize the computation. Of course in practice ε will be extremely small so that unlike the schematic an enormous number of oscillations of Φ will occur in the t interval $[0,h]$.Notice how far the computed answer $\Phi(Nk)\tilde{v}_0(h)$ may be from the usual approximation to the solution, $u_0(h,h/\varepsilon)$.

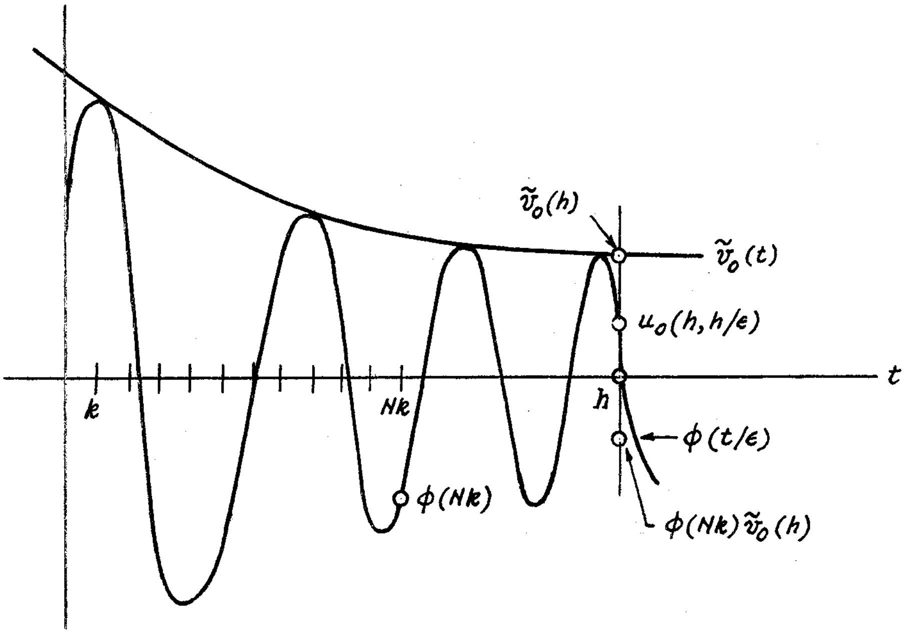

Figure 13.3

The fundamental matrix $\Phi(\tau)$ is composed of modes corresponding to the eigenvalues of A_0. Since the eigenvalues of A_0 lie in the closed left half plan, the profile for (a component) of Φ will after some moderate number of cycles settle down to an (almost) periodic function. Thus the set of mesh points $\{jk\,|\,j=0,\ldots,N\}$ may be expected to extend over just these cycles (approximately).

13.7 Numerical Results

In this section we tabulate the results of calculations with three sample problems, P_i, $i=1,2,3$. P_1 corresponds to a damped case (A_0 has real eigenvalues), P_2 to a purely oscillatory A_0 and P_3 to a mixed case.The numerical methods used were chosen to be the most elementary (e.g. Euler's method for differential equations and Simpson's rule for integrals) so that the results are accurate only to a few percent.Moreover $\varepsilon/h=.1$ or $.2$ so that the examples are not particularly stiff.

TABLES

Problem P_1 (damped case)

$$A_0 = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \qquad A_1 = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \qquad \begin{array}{l} \varepsilon = .01 \\ h = .05 \\ k = .05 \end{array}$$

t	$\tilde{u}_0$		$\Phi(kN)\tilde{u}_0$	
0	1.00	1.00	1.00	1.00
.05	.953	1.	.953	0.0
.10	.908	1.	.906	0.0
.15	.865	1.	.862	0.0
.20	.824	1.	.820	0.0
.25	.785	1.	.780	0.0

Problem P_2 (oscillatory case)

$$A_0 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \qquad A_1 = \begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix} \qquad \begin{array}{l} \varepsilon = .001 \\ h = .01 \\ k = .05 \end{array}$$

t	$\tilde{u}_0$		$\Phi(kN)\tilde{u}_0$	
0	0.5	0.5	0.5	0.5
.01	.495	.495	.325	.605
.02	.490	.490	.184	.669
.03	.485	.485	.007	.687
.04	.480	.481	-.167	.660
.05	.475	.476	-.327	.589

Problem P_3 (mixed case)

$$A_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \begin{array}{l} \varepsilon = .01 \\ h = .05 \\ k = .05 \end{array}$$

t	$\tilde{u}_0$				$\Phi(kN)\tilde{u}_0$			
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
.05	1.05	1.05	1.06	1.04	1.05	0.0	-1.45	.327
.10	1.11	1.11	1.12	1.09	1.10	0.0	.534	-1.46
.15	1.17	1.16	1.18	1.14	1.17	0.0	.997	1.31
.20	1.23	1.22	1.24	1.19	1.22	0.0	-1.72	.149
.25	1.29	1.28	1.31	1.24	1.28	0.0	.846	-1.60

REFERENCES

[13.1] Miranker, W.L. and Hoppensteadt, F,, "Numerical Methods for Stiff Systems of Differential Equations Related with Transistors, Tunnel Diodes, etc." Lecture Notes in Computer Science, *10*, Springer-Verlag (1973) pp.416-432.

[13.2] Hoppensteadt, F. and Miranker,W.L., "Differential Equations Having Rapidly Changing Solutions", to appear J. Differential Eqns.

[13.3] Amdursky, V. and Ziv, A., "On the Numerical Treatment of Stiff Highly-Oscillatory Systems", IBM Israel Scientific Center Report 015 (1974).

§ 14. A METHOD OF AVERAGING

14. 1 Stable functionals

Consider the following model problem

14.1) $$\ddot{x} + \lambda^2 x = \lambda^2 \sin t,$$

and the following family of solutions

14.2) $$x(t) = a \sin \lambda t + \frac{\sin t}{1-1/\lambda^2}.$$

For λ large, this solution family consists of a high frequency carrier wave, $a\ s\ n\lambda t$, modulated by a slow wave, $(\sin \lambda t)/(1-1/\lambda^2)$. The specification of the value at a point of such a function is an ill-conditioned problem.

We have seen that the linear multistep class of methods is highly desirable for numerical analysis since these methods are easy to calculate with and easy to analyze. However these methods consist of a linear combination of unstable functionals of the solution of (14.1), namely values and values of derivatives at points. In this section we will show how to replace these unstable functionals by stable ones, thereby producing a class of linear multistep methods suitable for the stiff problem.

We will not characterize the classes of functionals which are stable in an abstract way. Rather we select two special functionals which are an averaging functional and an appropriate evaluation functional which ought to be stable in the sense discussed. We construct the numerical methods out of these two functionals.

14.2 The problem treated

We develop our method in the context of the problem,

$$\ddot{x} + \lambda^2 x = f(x,t), \qquad t \in [0,T],$$

14.3)

$$x(0) = x_0 ,$$

where x and f are scalars.

The solution of this problem will be required to exist on the larger interval $I=[-\tau,T]$ where the quantity $\tau > 0$ will be specified in (14.9). Thus, we assume that $f(x,t)$ is continuous in t, $t \in I$ and Lipschitz continuous in x for all such t, with Lipschitz constant $\mathbb{L}$. In particular $f(x,t)$ is uniformly bounded for $t \in I$ and x restricted to any compact real set including in particular the set of values taken on by the solutions $x(t)$ for $t \in I$.

At first we restrict our attention to the linear problem in which $f(x,t)=f(t)$. Then in section 14.9 we make some comments about the nonlinear case and the case of second order systems.

14.3 Choice of functionals

Let $N > 0$ be an integer, let $h=T/N$ and let $t_i = ih$, $i=0,\pm 1,\ldots$ be the points of a mesh. We seek the functional $y(t)$ of x at the points of this mesh. Let $z(t)$ be a functional of x which can be calculated at each mesh point. Then we seek to determine $y_n = y(t_n)$, in terms of y_{n-i}, $i = 1,\ldots,r$ and $z_{n-i} = z(t_{n-i})$, $i = 0,1,\ldots,s$ by means of the linear multistep formula

14.4) $$\sum_{i=0}^{r} a_i y_{n-i} + \sum_{i=0}^{s} b_i z_{n-i} = 0, \qquad n = 0,1,\ldots,N.$$

The initial values y_i, $i=-1,\ldots,-r$ are assumed to be furnished by some independent means.

In the case (14.3) of interest and λ large we choose $y(t)$ to be

14.5) $$y(t) = \int_{-\infty}^{\infty} k(t-s)x(s)ds \ ,$$

where

14.6) $$k(z) = \frac{1}{\Delta}\begin{cases} 1, & -\Delta < z < 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus $y(t)$ represents the average of $x(t)$ over the interval $[t-\Delta, t]$.

The functional $z(t)$ is chosen to be $\left[\frac{d^2}{dt^2} + \lambda^2\right]x(t)$, i.e., $f(t)$, which can be calculated at each mesh point. Thus with a change in normalization (14.4) may be written as

14.7) $$y_n = \sum_{i=1}^{r} c_i y_{n-i} + h^2 \sum_{i=0}^{s} d_i f_{n-i}$$

14.4 Representers

We introduce the reproducing kernel space, $\mathcal{H} \equiv \mathcal{H}_m$ which is the Sobolev space $W_m^2[-\infty,\infty]$ with the inner product

14.8) $$<f, g> = \sum_{j=0}^{n} \binom{m}{j}\left(f^{(j)}, g*^{(j)}\right),$$

where

$$(f, g) = \int_{-\infty}^{\infty} f(t)g^*(t)dt.$$

An asterisk is used to denote the complex conjugate throughout. Since we are interested in solutions of (14.3) on the interval

14.9) $$I = [-h\Delta, T]$$

we may identify both a solution of 14.3 and $f(t)$ appearing in 14.3 with the unique functions of minimal norm in $\mathcal{H}$ with which they agree on I, respectively. Of course on I, f is required to have m-1 absolutely continuous derivatives and an m-th derivative a.e. which is square integrable.

We use a carot to denote the Fourier transform, viz.

14.10) $$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega t} \hat{f}(\omega) d\omega, \quad \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega t} f(t) dt.$$

Then the inner product in $\mathcal{H}$ may be written as

14.11) $$<f,g> = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) \hat{g}^*(\omega) |P_m(\omega)|^2 d\omega,$$

where

14.12) $$P_m(\omega) = (1-i\omega)^m.$$

The reproducing kernel in $\mathcal{H}$ is

14.13) $$R_t \equiv R_t^m(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{i(s-t)\omega}}{|P_m(\omega)|^2} d\omega.$$

A second Hilbert space, $\hat{\mathcal{H}}$ is introduced as follows:

14.14) $$\hat{\mathcal{H}} \equiv \hat{\mathcal{H}}_m = \{\hat{f} | \hat{f} P_m \in \mathcal{L}_2\}.$$

The inner product in $\hat{\mathcal{H}}$ is

14.15) $$<\hat{f},\hat{g}> = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f} \hat{g}^* |P_m|^2 d\omega.$$

(14.11) defines an isometric isomorphism between $\mathcal{H}$ and $\hat{\mathcal{H}}$. The symbol $\sim$, will denote this isomorphism. Then from (14.11) we see that the isomorphism between R_t and its image in $\hat{\mathcal{H}}$ is expressed by

14.16) $$R_t \sim \frac{e^{-i\omega t}}{|P_m(\omega)|^2}.$$

Then for the representer, η_t of $\dfrac{d^2}{dt^2}+\lambda^2$, we have

$$\eta_t \equiv R''_t + \lambda^2 R_t \sim (-\omega^2+\lambda^2)\,\frac{e^{-i\omega t}}{|P_m(\omega)|^2}\,. \tag{14.17}$$

For the representer k_t of $y(t)$ given by 14.5 and 14.6 we have

$$\begin{aligned} k_t \equiv k_t(s) &= \frac{1}{\Delta}\int_{t-\Delta}^{t} R_u(s)du \\ &\sim \frac{1}{\Delta}\int_{t-\Delta}^{t} \frac{e^{-i\omega u}}{|P_m(\omega)|^2}\,du \\ &= \frac{1}{|P_m(\omega)|^2}\, e^{-i\omega t}\left[\frac{1-e^{-i\omega\Delta}}{-i\omega\Delta}\right] \\ &= \frac{e^{-i\omega t}}{|P_m(\omega)|^2}\sqrt{2\pi}\,\hat{k}(\omega), \end{aligned} \tag{14.18}$$

where $\hat{k}(\omega)$ is the Fourier transform of $k(z)$ given in (14.6).

With these representors, the formula (14.7) leads us to introduce the following linear functional g_n.

$$g_n \equiv g_n[x] \equiv \langle k_{t_n} - \sum_{i=1}^{r} c_i k_{t_{n-i}} - h^2 \sum_{i=0}^{s} d_i \eta_{t_i},\ x\rangle. \tag{14.19}$$

g_n will be zero if x is the numerical solution. In general g_n is not zero and is the analogue of the local truncation error for classical linear multistep schemes.

14.5 Local Error and Generalized Moment Conditions

g_n is characterized in the following definition.

Def. 14.1 Using (14.15) as a definition, we call the linear functional, g_n appearing there, the local truncation error of the method (14.7).

To estimate the local truncation error we write

$$14.20)\qquad \|g_n\|^2 \leqslant \left|\left| k_{t_n} - \sum_{j=1}^{r} c_j k_{t_{n-j}} - h^2 \sum_{j=0}^{s} d_j \eta_{t_{n-j}} \right|\right|_{\wedge}^2$$

where as usual

$$14.21)\qquad ||x||^2 = <x,x> \quad \text{and} \quad ||x||_{\wedge}^2 = <x,x>_{\wedge}.$$

We will drop the subscript, $\wedge$, since no confusion should result.

Now using (14.15), (14.18) and (14.19), we find for the right member of (14.20) that

$$14.22)\quad \left|\left| k_{t_n} - \sum_{j=1}^{r} c_j k_{t_{n-j}} - h^2 \sum_{j=0}^{s} d_j \eta_{t_{n-j}} \right|\right|^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |t(\omega)|^2 \frac{d\omega}{|P_m(\omega)|^2}$$

where

$$14.23)\qquad t(\omega) = \sqrt{2\pi}\, \hat{k}(\omega) . \sum_{j=0}^{r} s_j e^{ij\omega h} - h^2(\lambda^2 - \omega^2) \sum_{j=1}^{s} d_j e^{ij\omega h}.$$

Here

$$14.24)\qquad s_0 = 1 \quad \text{and} \quad s_j = -c_j, \qquad j=1,\ldots,r.$$

Expanding $t(\omega)$ formally in a Taylor series with remainder gives

$$14.25)\qquad t(\omega) = \sum_{l=0}^{p-1} (ih\omega)^l m_l + R_p,$$

where from (14.23) and (14.25) we obtain

$$14.26)\qquad m_l = \frac{1}{(l+1)!} \sum_{k=1}^{l+1} \binom{l+1}{k} L^{k-1} \sum_{j=0}^{r} j^{1+l-k} s_j$$

$$- \frac{h^2 \lambda^2}{l!} \sum_{j=0}^{s} j^l d_j - \frac{1}{(l-2)!} \sum_{j=0}^{s} j^{l-2} d_j$$

and

$$R_p = \frac{(ih\omega)^p}{p!}\left[-\frac{1}{L(p+1)}\sum_{j=0}^{r} s_j\left(j^{p+1}e^{ijh\omega_{j,1}} - (j+L)^{p+1}e^{ijh(1+L)\omega_{j,2}}\right)\right.$$

14.27)

$$\left. - h^2\lambda^2 \sum_{j=0}^{s} j^p d_j e^{ij\omega_{j,3}} - p(p-1)\sum_{j=0}^{s} j^{p-2} d_j e^{ij\omega_{j,4}}\right].$$

In (14.26) and (14.27) we have used

14.28) $$L = \Delta/h.$$

That is in terms of the functional k of (14.5) and (14.6) the interval, Δ, over which the average is taken is a multiple, L, of the mesh increment h. In (14.25) the quantities $\omega_{j,1}$ and $\omega_{j,2}$, $j = 0,\dots,r$ and $\omega_{j,3}$ and $\omega_{j,4}$, $j = 0,\dots,s$ are values of ω which arise from the calculation of the remainder in Taylor's theorem.

The quantities m_l are characterized in the following definition.

Definition 14.2: We call the m_l, $l=0,1,\dots,$ the (generalized) moments (of the coefficients). Analogously $m_l=0$, $l=0,1,\dots,$ will be called the (generalized) moment conditions.

Consider the following remark.

Remark 14.1: View the equations $m_l=0$, $l = 0,\dots,r-1$ as r equations for the r unknowns s_j, $j=1,\dots,r$. The l-th row of the resulting coefficients matrix which has as its j-th term

14.29) $$\frac{1}{(l+1)!}\sum_{k=1}^{l}\binom{l+1}{r}L^{k-1}j^{1+l-k},$$

is a linear combination of the first l rows of the Vandermonde matrix. Thus the system of r equations has a solution in this case. Indeed by choosing the d_j, $j=0,\dots,s$ to be proportional to λ^{-2}, we obtain a solution for the s_j, $j=1,\dots,r$ which is $O(1)+O(\lambda^{-2})$.

From the form of $t(\omega)$ given in (14.23) we may make the following remark the assertion of which follows from a familiar arguement which proceeds by breaking up the range of integration in (14.22) appropriately.

Remark 14.2: If p is chosen less than m and the coefficients s_j, $j=1,\dots,r$ and d_j, $j=0,\dots,s$ are chosen as solutions of the generalized moment equations $m_l=0$, $l=0,1,\dots,p$, we may obtain an estimate of the local truncation error of the following form.

$$14.30)\qquad ||g_n|| \equiv \max_{\substack{x\in\mathcal{H}\\ ||x||\leqslant 1}} |g_n| \leqslant O(h^{P+1}),\quad p<m.$$

We collect these remarks into the following theorem.

Theorem 14.1. There exists a choice of coefficients s_j, $j=1,\dots,r$ and d_j, $j=0,\dots,s$ such that the local truncation error has a bound of the form (14.30). Moreover, this bound is uniform in λ for $|\lambda|\geqslant\lambda_0>0$.

14.6 Stability and Global Error Analyses

y_n, $n=0,1,\dots$ denotes the values obtained by the multistep formula, (14.7) from the initial values y_n, $n=-r,\dots,-1$. Let Y_n, $n=-r,-r+1,\dots$ denote the exact values of these functionals. Let

$$14.31)\qquad e_n = y_n - Y_n,\qquad n=-r,-r+1,\dots$$

denote the cumulative error. For convenience, assume that the initial functionals $e_n=0$, $n=-r,-r+1,\dots,-1$.

Subtract the following identity

$$14.32)\quad Y_n = \sum_{j=1}^{r} c_jY_{n-j} + h^2\sum_{j=0}^{s} d_jf_{n-j} + Y_n - \sum_{j=1}^{r} c_jY_{n-j} - h^2\sum_{j=0}^{s} d_jf_{n-j},$$

from (14.7). We get

$$14.33)\qquad e_n = \sum_{j=1}^{r} c_je_{n-j} + g_n.$$

Here

$$14.34)\qquad g_n = -Y_n + \sum_{j=1}^{r} c_jY_{n-j} + h^2\sum_{j=0}^{s} d_jf_{n-j},$$

is the value of the linear functional, g_n of (14.19) applied to x, the exact solution of the initial value problem (14.3). To solve (14.33) for e_n, we use the polynomial $S(z)$:

14.35) $$S(z) = \sum_{j=0}^{r} s_j z^{r-j}.$$

Since $s_0=1$, $[z^r S(z^{-1})]^{-1}$ is an analytic function of z in a neighborhood of $z=0$. Then let its power series be given by

14.36) $$[z^r S(z^{-1})]^{-1} = \sum_{j=0}^{\infty} \sigma_j z^j.$$

Now multiply (14.33) by σ_{N-n} and sum the result over n from r to N. The result is the solution of (14.33):

14.37) $$e_N = \sum_{n=r}^{N} \sigma_{N-n} g_n.$$

We use the following definition.

Definition 14.3 (Stability). If the sequence $\{\sigma_j, j=0,1,...\}$ is bounded, then the method is said to be stable.

We recall the following definition.

Definitions 14.4: $S(z)$ is said to obey the root condition if all of its roots lie in the closed unit disc while those of its roots which lie on the boundary of that disc are simple.

With this definition we may state the following lemma which characterizes the stability of the method.

Lemma 14.1 If the polynomial $S(z)$ obeys the root condition, then the sequence $\{\sigma_j, j=0,1,...\}$ is bounded. i.e. the method is stable. (cf. Lemma 8.2).

If this lemma is applicable (14.37) gives

14.38) $$|e_N| \leqslant const\ N \max_{r\leqslant n\leqslant N} ||g_n||\cdot||x||,$$

where x is the exact solution of (14.3)

Combining this with (14.30) gives the following theorem.

Theorem 14.2. If the choice of coefficients characterized in Theorem 14.1 give rise to a stable method, then for the method (14.7) with those coefficients,

14.39) $$||e_N|| \leqslant O(h^P), \qquad p < m,$$

uniformly in λ for $|\lambda| \geqslant \lambda_0 > 0$.

14.7. Examples

We now consider some examples of methods of the type (14.7) in which the coefficients are determined by the generalized moment conditions.

From 14.26 we have for $l=0,1$ and 2, respectively,

0. $$m_0 \equiv \sum_{j=0}^{r} s_j - h^2\lambda^2 \sum_{j=0}^{s} d_j$$

14.40) 1. $$m_1 \equiv \sum_{j=0}^{r} j s_j + \frac{L}{2} \sum_{j=0}^{r} s_j - h^2\lambda^2 \sum_{j=0}^{s} j d_j$$

2. $$m_2 \equiv \frac{1}{2} \sum_{j=0}^{r} j^2 s_j + \frac{L}{2} \sum_{j=0}^{r} j s_j + \frac{L^2}{6} \sum_{j=0}^{r} s_j - \frac{h^2\lambda^2}{2} \sum_{j=0}^{s} j^2 d_j - \sum_{j=0}^{s} d_j .$$

Consider the case

A. $m_0 = m_1 = 0$.

For $r=s=1$, we get

14.41)
$$\begin{aligned} c_1 &= 1 - \frac{2}{L} + \frac{2}{L} h^2\lambda^2 d_0 \\ d_1 &= \frac{2}{h^2\lambda^2 L} - \left(\frac{2}{L} + 1\right) d_0 . \end{aligned}$$

In the special case $d_0=0$, (14.41) becomes

14.42)
$$\mathrm{I} \begin{cases} c_1 = 1 - \dfrac{2}{L} \\ d_1 = \dfrac{2}{h^2\lambda^2 L} . \end{cases}$$

These coefficients (i.e. c_1) obey the root condition if and only if

14.43) $$L \geqslant 1.$$

In the special case $d_0 = d_1$, (14.41) becomes

14.44)
$$\text{II}\begin{cases} c_1 = 1 - \dfrac{2}{L+1} \\[2ex] d_0 = d_1 = \dfrac{1}{h^2\lambda^2}\ \dfrac{1}{L+1} \,. \end{cases}$$

Under the restriction $L \geqslant 0$, the root condition is equivalent to

14.45)
$$L \geqslant 0,$$

for the coefficients (14.44). For $r=s=2$,

14.46)
$$\begin{aligned} c_1 &= 1 - \frac{2}{L} - \left(1 + \frac{2}{L}\right)c_2 + \frac{2}{L}\lambda^2 h^2 (d_0 - d_1) \\ d_1 &= \frac{2}{L\lambda^2 h^2}(1+c_2) - \left(1 + \frac{2}{L}\right)d_0 - \left(1 - \frac{2}{L}\right)d_2 . \end{aligned}$$

In the special case $d_0=0$, $c_1=c_2$, $d_1=d_2$, (14.46) becomes

14.47)
$$\text{III}\begin{cases} c_1 = c_2 = \dfrac{L-3}{2L} \\[2ex] d_1 = d_2 = \dfrac{3}{2\lambda^2 h^2 L} \,. \end{cases}$$

In this case $S(z)=z^2 - \dfrac{L-3}{2L}\,z - \dfrac{L-3}{2L}$ and this polynomial, $S(z)$, obeys the root condition for a set of values of L which includes all $L \geqslant 1$.

In the special case $c_1=c_2$, $d_1=d_2=0$, (14.46) becomes

14.48)
$$\text{IV}\begin{cases} c_1 = c_2 = \dfrac{1}{2}\ \dfrac{L}{3+L} \\[2ex] d_0 = \dfrac{1}{\lambda^2 h^2}\ \dfrac{3}{3+L} \,. \end{cases}$$

Here $S(z)=z^2 - \dfrac{1}{2}\,\dfrac{L}{3+L}\,z - \dfrac{1}{2}\,\dfrac{L}{3+L}$. This polynomial obeys the root condition for a set of values of L which includes all $L > 0$.

In the special case $c_1=c_2$, $d_0=d_1=d_2$, (14.46) becomes

14.49) $$\text{V}\begin{cases} c_1 = c_2 = \dfrac{1}{2}\,\dfrac{L-1}{L+1} \\[2ex] d_0 = d_1 = d_2 = \dfrac{2}{3\lambda^2 h^2}\;\;\dfrac{1}{1+L}\,. \end{cases}$$

In this case the root conditions is obeyed for $L > 0$. Now we consider a case corresponding to three moment conditions:

B. $m_0 = m_1 = m_2 = 0$.

For $r=s=1$, we get

14.50) $$\text{VI}\begin{cases} c_1 = 1 - L\Big/\left(\dfrac{L^2}{3} + \dfrac{L}{2} - \dfrac{2}{h^2\lambda^2}\right) \\[2ex] d_0 = \dfrac{1}{\lambda^2 h^2}\left[1 - L + L^2\Big/\left(\dfrac{2}{3}L^2 + L - \dfrac{4}{h^2\lambda^2}\right)\right] \\[2ex] d_1 = \dfrac{1}{\lambda^2 h^2}\left[-1 + L + (2L - L^2)\Big/\left(\dfrac{2}{3}L^2 + L - \dfrac{4}{h^2\lambda^2}\right)\right]. \end{cases}$$

Notice that the root condition is obeyed for L large and positive but is violated for $h\lambda$ small compared to L.

Remark 14.3: In all of these example and in the general case, we see that the coefficients obtained as solutions of the moment conditions depend on λ^2. At first sight this seems to be more restrictive than the case of the classical linear multistep formulas where the coefficients of the formula do not depend on the coefficients of the differential equation. In fact we see no such distinction. In the classical case the coefficients of the differential equation enter into the method when it is used to approximate the differential equation e.g. when $\dot{y}_{n-i}$ is replaced by $f(y_{n-i}, t_{n-i})$. It is essential after all that the numerical method at some point be dependent on the equation to be solved. In our case this dependence occurs at the outset in the determination of coefficients and in the error analysis. In the classical case it enters in the error analysis and in the use of the methods.

14.8. Illustrative computations

We now apply the six sets of methods labeled I, II, ..., IV in 14.6 respectively, to the sample problem

$$\ddot{x} + \lambda^2 x = \lambda^2 \sin t$$

14.51) $$x(0) = 0, \qquad x'(0) = \frac{\lambda}{2} + \frac{1}{1 - 1/\lambda^2} .$$

Runs are made over the interval $[0,T] = [0,\pi]$. In

Method	L λ	1	2	3	1	2	7
I	10	.273	.108	.112	.133	.126	.126
	10^3	.113	.00217	.0611	.0283	.00683	.0083
	10^5	.112	.00209	.0611	.0111	.000106	.00627
II	10	.122	.133	.155	.126	.127	.128
	10^3	.00125	.0622	.177	.0241	.00926	.0136
	10^5	.00104	.0621	.177	.000118	.00627	.0125
III	10	.242	.111	.0872	.136	.126	.126
	10^3	.0032	.00422	.00317	.0294	.00684	.00546
	10^5	.0034	.00419	.00313	.00023	.00112	.89E-6
IV	10	.123	.111	.0938	.126	.126	.126
	10^3	.00627	.0144	.0244	.0241	.00684	.00546
	10^5	.00623	.0144	.0244	.000133	.000179	.000264
V	10	.144	.152	.156	.127	.127	.128
	10^3	.0657	.094	.119	.0249	.0116	.0136
	10^5	.0657	.0939	.119	.0063	.00942	.0125
VI	10	.758E4	.66E11	.124	.195E1	.471E1	.11E2
	10^3	.0447	.0639	.244	.0246	.00901	.0253
	10^5	.0447	.0639	.244	.00421	.00629	.0251
h			.1			.01	

$||e||_{l_2}$

Table 14.1

table 14.1 we display the l_2-norm of the cumulative error

$$14.52) \qquad ||e||_{l_2} \equiv \left[h \sum_{n=0}^{[\pi/h]} e_n^2 \right]^{1/2},$$

for a set of various combinations of $h = .1, .01$, $\lambda = 10, 10^3, 10^5$ and $L=1, 2, 3$ and for each of the six methods cited.

To illustrate both the favorable and unfavorable effects in our methods table 14.1 contains cases for which the methods are designed to operate well along with cases to which correspond poor or nonsensical results.

For example although the cases corresponding to $\lambda=10$ give fair results, these cases are not stiff and we should not expect good results. When h is decreased improvement should occur but only for the stiff cases. The cases $\lambda = 10^3$ and $h = .01$ are not stiff and improvement with decreasing h does not always occur in these cases. Method VI is used in some unstable cases. Examining (14.27) we see that R_p is proportional to L^p. Thus in some cases as L increases we see an improvement due to improving the averaging (i.e. increasing Δ), but ultimately a degradation due to the L dependence of R_p. The stiff cases for moderate L give extremely good results as we expect.

14.9 The non linear case and the case of systems

In [14.1] a discussion of the extension of the results described in sections 14.1 - 14.8 to the nonlinear case and to the case of systems is given. We will give some highlights of that discussion.

In the nonlinear case, f_{n-i} in the multistep formula (14.7) is replaced by $f(y_{n-i}, t_{n-i})$ since $f_{n-i} = f(x_{n-i}, t_{n-i})$ can not be computed as we proceed along the mesh. This results in a degradation of the error estimate (14.39) to the following:

$$14.53) \qquad ||e_N|| \leqslant const[h^p + L\varepsilon_1 v_m].$$

Here

$$14.54) \qquad \varepsilon_1 = \max_j |h^2 d_j \mathbb{L}|$$

and

$$14.55) \qquad v_m = \frac{1}{2}\left[\int_{-\infty}^{\infty} \frac{|\omega|^2}{|P_m(\omega)|^2}\, d\omega\right]^{1/2}.$$

Remark 14.4: The two terms in the estimate (14.53) are not comparable in orders of h. The first term which corresponds to the local truncation error is small for h small. The second term is the error by which a function may be approximated by its average. We may expect the latter to be small if λ is large. (14.53) may be viewed as the statement that modulo the error made in replacing a function by its average, the numerical method is globally h^p.

In the systems case, the differential equations (14.3) is replaced by the second order system

$$14.56) \qquad \ddot{x} + \Lambda^2 x = f(x, t).$$

Here x and f are q-vectors and Λ is a $q \times q$ matrix. The coefficients c_j (and s_j) and d_j of the numerical method are replaced by $q \times q$ matrices (denoted by the same symbols). Many such formal replacements of the scalar development follow. For example the first two moments become

$$14.57) \qquad \begin{aligned} m_0 &= \left(\sum_{j=0}^{r} s_j - h^2\Lambda^2 \sum_{j=0}^{s} d_j\right)\zeta_q \\ m_1 &= \left(\sum_{j=0}^{r} j s_j + \frac{L}{2}\sum_{j=0}^{r} s_j - \Lambda^2 \sum_{j=0}^{s} j d_j\right)\zeta_q \end{aligned}$$

(compare (14.27)), where ζ_q is the q-vector all of whose components are unity.

The error analysis proceeds similarly (using some of the matricial arguments of § 8 leading to an estimate of the global error which is similar to the one described in Theorem 14.2).

We conclude this summary of the systems case with the following two remarks.

Remark 14.5: Referring to Remark 14.3 and the dependence of the coefficients of the numerical method on the coefficients of the differential equation, we see from (14.57) the way in which the dependence appears in terms of the matrix Λ^2, for the coefficients determined by generalized moment conditions. It is important to take note that the coefficients depend on the matrix Λ^2 and not explicitly on eigenvalues of Λ^2. Thus, if we know that a system is stiff, with highly oscillatory components, we may use the methods described here without having to calculate the eigenvalues of Λ^2 which cause this stiffness.

Remark 14.6: In the usual systems case for the numerical treatment of differential equations the methods frequently used are the scalar methods with the scalar coefficients simply multiplied by I_q. We suspect that the methods developed here in the scalar case would work in the same way with the simple additional requirement of replacing λ or λ^{-1} by Λ or Λ^{-1} respectively. At present this remark is only a conjecture and we defer for a further study its verification.

REFERENCES

[14.1] Miranker, W.L. and Wahba, G., "An Averaging Method for the Stiff Highly Oscillatory Problem", IBM Research Center Report RC 5528 (7/21/75), to appear Math. Comp.