TW

stichting
mathematisch
centrum

MC

AFDELING TOEGEPASTE WISKUNDE TW 122/70 OKTOBER

P.J. VAN DER HOUWEN

ONE-STEP METHODS FOR LINEAR INITIAL VALUE PROBLEMS II
APPLICATIONS TO STIFF EQUATIONS

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

Contents

1. Introduction

In a previous paper [11] we discussed polynomial methods for the numerical solution of differential equations of the type

$$(1.1)\qquad \frac{d\tilde{U}}{dt} = D\tilde{U} + F,$$

where $\tilde{U}$ and F are (vector) functions of the variable t, and D is a matrix with constant coefficients. In particular, we were interested in initial value problems for equations of type (1.1) which arise from partial discretization of linear partial differential equations. A number of difference schemes were developed which are appropriate for such equations. Here, we concentrate on another important class of differential equations of type (1.1), the so-called stiff equations, i.e. equations of which the solution can be distinguished in slowly and rapidly varying functions. The rapidly varying components correspond to eigenvalues of D which have a large modulus, the slowly varying ones to eigenvalues with a small modulus. When Runge-Kutta methods are applied to such equations small steps are required in order to represent accurately the fast components, while large steps are allowed for the slow components. Although the fast-components vanish after some time in the analytical solution, they always are present in a numerical approximation because of round-off errors. Therefore, Runge-Kutta methods are not attractive in solving stiff equations.

The first numerical treatment of stiff equations seems to be given by Curtiss and Hirschfelder [5] in 1952. For the case of systems of equations, however, Moretti [19] reported some short-comings of their algorithm. After the paper of Curtiss and Hirschfelder, particularly in recent years, a large number of numerical methods have been developed (see the reference list at the end of this paper). The greater part of these methods are based on some representation of the solution, which takes into account that the general solution is composed of a slowly and a rapidly varying component. We mention the papers of Certaine [4] (1960), Pope [20] (1963), Treanor [22] (1966), Fowler and Warten[7] (1967), Liniger [16] (1968), and Lawson [17] (1967). Except for the method of Liniger, these methods are all explicit one- or multi-step methods, which allow considerably larger steps than the Runge-Kutta methods. In this paper we propose two further explicit methods which may be useful in solving stiff equations. The first methods is based on the polynomials given in [11]. However, instead of a polynomial with fixed coefficients we now use a polynomial with coefficients depending on the step length. The step is prescribed by accuracy censiderations and the coefficients are selected in such a way that the corresponding polynomial is stable for this step.

The second method exploits the locations of the eigenvalues of the matrix D. Therefore, a rough knowledge of these locations is necessary. We discuss generating polynomials which are appropriate in cases where the eigenvalues are known to lie either in two widely seperated clusters centered at the negative real axis, or in three clusters two of which being conjugate complex, the third being near the origin. These polynomials are, in fact, two-point approximations of the exponential operator, whereas the polynomials generating the first method are one-point approximations.

Another possibility is the use of implicit difference schemes. We refer to the papers of Dahlquist [6] (1963), Calahan [3] (1967), Liniger and Willoughby [17] (1967) and Gear [9] (1969). Such schemes allow, in general, unrestricted time steps but this has to be paid for by solving at each integration step a system of algebraic equations. It depends, therefore, on the particular problem whether an implicit method should be used or not.

In section 6 we give the general generating function (a rational function instead of a polynomial) of implicit difference schemes, the consistency conditions, and the stability criteria of some schemes.

Numerical applications of the methods proposed in this paper will be given in reference [12] (to appear).

Finally, the author wishes to acknowledge the work done by Mr. IJsselstein who wrote the plotting-program by which the figures 4.3 and 4.4 were obtained.

2. Stiff equations

2.1 General considerations

The first numerical approach to a (single) stiff equation seems to be given by Curtiss and Hirschfelder [5] in 1952. In their paper, a single differential equation is called stiff in the neighbourhood of the point $t = t_k$, when in this neighbourhood, the integral curves of the equation have the behaviour as shown in figure 2.1.

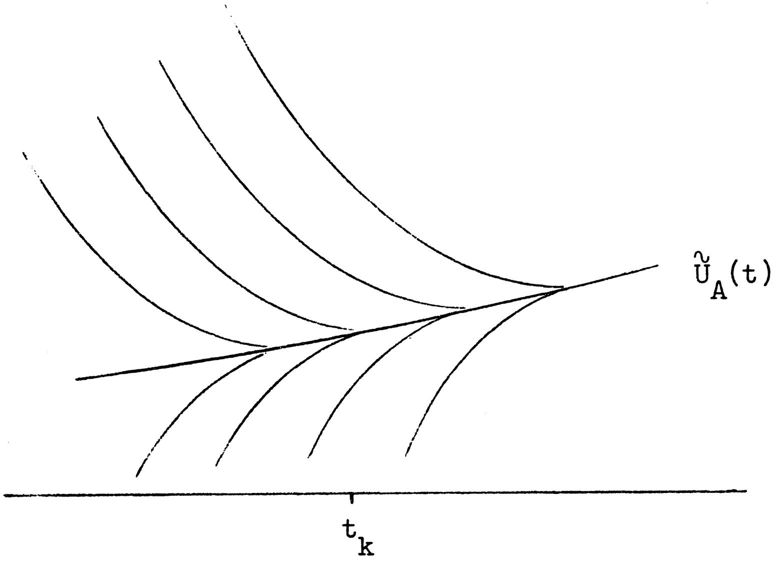


fig. 2.1 Integral curves of a single stiff equation

Thus, a stiff equation has one slowly varying solution and further only solutions which converge to (or diverge from, depending on the direction of integration) this particular solution. Hence, the integral curve $y = \tilde{U}(t)$ corresponding to the given initial value problem, finally converges to a, let us say, asymptotic integral curve. We shall denote the asymptotic solution by $\tilde{U}_A(t)$. This leads us to the following representation of the solution of a single stiff equation as the sum of an asymptotic part and a perturbation part:

$$(2.1) \qquad \tilde{U}(t) = \tilde{U}_A(t) + e^{D(t-t_k)} [\tilde{U}(t_k) - \tilde{U}_A(t_k)].$$

In the following we shall assume that the perturbation part decreases for increasing values of t, i.e. D is assumed to be negative.

A typical example of a stiff equation is given by

$$(2.2)\qquad \frac{d}{dt}\tilde{U} = -\,1000\,\tilde{U} + t^2,\ \tilde{U}(0) = \tilde{U}_0.$$

The general solution of (2.2) is given by

$$(2.3)\qquad \tilde{U}(t) = 10^{-3}(t^2-2.10^{-3}t+2.10^{-6}) + e^{-1000t}(\tilde{U}_0-2.10^{-9}).$$

Evidently, here the asymptotic solution is

$$(2.4)\qquad \tilde{U}_A(t) = 10^{-3}(t^2 - 2.10^{-3}t + 2.10^{-6}).$$

This example leads us to the following non-geometrical definition of a (single) stiff differential equation (compare Curtiss and Hirschfelder [5]):

Definition 2.1

Equation (1.1) is said to be stiff when $|D| >> 1$ and $F(t)$ is a well-behaved function of t, i.e. $F(t)$ varies with t considerable more slowly than $\exp(Dt)$.

In order to extend this definition to sets of equations we uncouple these equations by introducing the new dependent variable $\tilde{V} = Q\tilde{U}$, where Q is a non-singular matrix. This yields

$$Q^{-1}\dot{\tilde{V}} = DQ^{-1}\tilde{V} + F \text{ or } \dot{\tilde{V}} = QDQ^{-1}\tilde{V} + QF.$$

Let us suppose that D has a complete set of eigenvectors. Then Q can be chosen in such a way that QDQ^{-1} becomes a diagonal matrix, i.e. the set of equations is uncoupled. We shall call a system of equations stiff if one or more of the uncoupled equations is stiff in the sense of definition 2.1, or equivalently

Definition 2.2

Equation (1.1) is said to be stiff when $F(t)$ is a well-behaved function of t and the real parts of some of the eigenvalues of D are very large in magnitude.

Since the solution of each uncoupled equation can be written in the form (2.1), the solution of the original set of equations can also be represented in the form (2.1). However, the asymptotic solution is not unique, but depends on the initial conditions. To illustrate this we consider the following example of a set of stiff equations (compare Fowler and Warten [7]):

$$(2.5)\qquad \begin{cases} \dot{\tilde{U}} = D\tilde{U} + \tilde{F}\ ,\ \tilde{U}(0) = \tilde{U}_0, \\ D = \begin{pmatrix} -500.5 & 499.5 \\ 499.5 & -500.5 \end{pmatrix},\ F = \begin{pmatrix} 2 \\ 2 \end{pmatrix},\ \tilde{U}_0 = 2(a+1)\begin{pmatrix} 1 \\ 1 \end{pmatrix} + b\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \end{cases}$$

where a and b are real constants.
It is easily verified that the solution of (2.5) is given by

$$(2.6)\qquad \tilde{U}(t) = 2(ae^{-t}+1)\begin{pmatrix} 1 \\ 1 \end{pmatrix} + be^{-1000t}\begin{pmatrix} -1 \\ 1 \end{pmatrix} .$$

From this expression it is seen that the asymptotic solution depends on the value of a and only for very large values of t becomes independent of a.

In figure 2.2 some integral curves of equation (2.5) are shown.

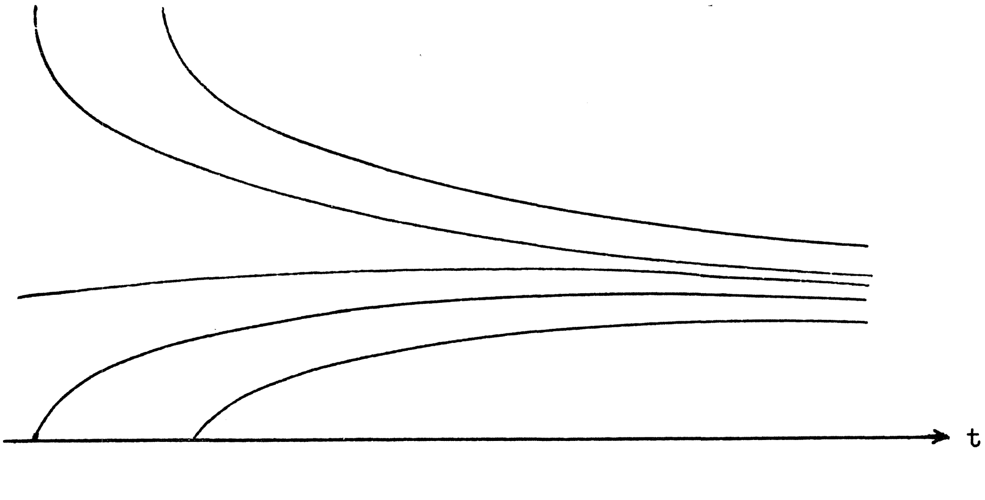


fig. 2.2 Integral curves of two coupled stiff equations

2.2 The local discretization error

It was shown in [11], section 2.2, that the local discretization error of an integration method based on a p-th order approximation of exp(τD) is approximately given by

$$\rho_k(\tau) \sim \left[\frac{1}{(p+1)!} - \beta_{p+1}\right] \tau^{p+1} \tilde{c}_k^{(p+1)}, \quad \tilde{c}_k^{(p+1)} = \left.\frac{d^{p+1}}{dt^{p+1}} \tilde{U}(t)\right|_{t=t_k} .$$

For a stiff equation, the solution of which can be represented by (2.1), this may be written in the form

$$(2.7) \qquad \rho_k(\tau) \sim \left[\frac{1}{(p+1)!} - \beta_{p+1}\right]\tau^{p+1} \{D^{p+1}(\tilde{U}(t_k) - \tilde{U}_A(t_k)) + \left.\frac{d^{p+1}}{dt^{p+1}} \tilde{U}_A(t)\right|_{t=t_k} \}.$$

Firstly, it follows from (2.7) that $\rho_k(\tau)$ converges rapidly to the function

$$(2.7') \qquad \rho_k(\tau) \sim \left[\frac{1}{(p+1)!} - \beta_{p+1}\right] \tau^{p+1} \left.\frac{d^{p+1}}{dt^{p+1}} \tilde{U}_A(t)\right|_{t=t_k} .$$

This demonstrates, as was already observed in [11], section 2.5, that in the first stage of the calculations (initial phase) a variable time step should be used.

Secondly, since $\tilde{U}_A(t)$ is a slowly varying function, $\rho_k(\tau)$ becomes very small in the asymptotic region, so that, so far as accuracy is concerned, large time steps are allowed even when low order polynomials are used. However, when Runge-Kutta methods are used, the large spectral radius, which is characteristic for a stiff equation, implies small time steps. Therefore, the methods given in [11] are not appropriate for solving stiff equations.

In order to overcome this difficulty one may use implicit difference schemes, which are, in general, stable for unrestricted time steps as will be shown in section 6.

Another possibility is to exploit the smallness of $\rho_k(\tau)$ and to use polynomials of order zero in the asymptotic region. It will be shown in section 3 that such polynomials have very flexible stability properties.

A further remark about the local error $\rho_k(\tau)$ in connection with stiff equations is the following. During the numerical integration process it is not $\rho_k(\tau)$ which is available, but the error $\rho'_k(\tau)$ defined by (see figure 2.3)

$$\rho_k'(\tau) = \tilde{U}'_{k+1} - u_{k+1},$$

where $\tilde{U}'_{k+1} = \tilde{U}'(t_k+\tau)$, $\tilde{U}'(t)$ being the solution of the differential equation through the point (t_k, u_k). For small values of $\tau^j c_k^{(j)}$, where

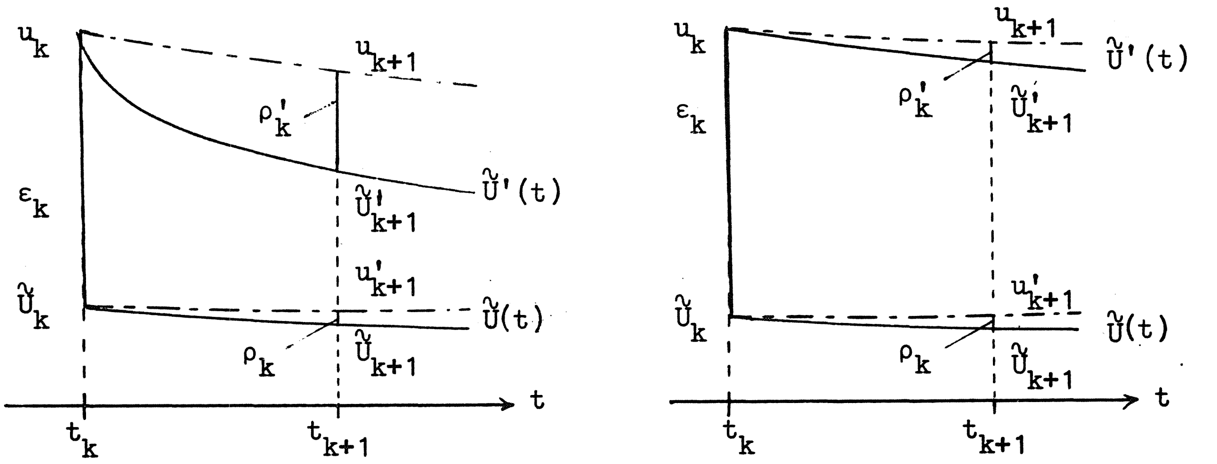
$$c_k^{(j)} = \frac{d^j}{dt^j} \tilde{U}'(t)\Big|_{t_k} ,$$

$\rho_k'(\tau)$ can be approximated by (compare [11], section 2.2)

$$(2.8) \qquad \rho_k'(\tau) \sim \left[\frac{1}{(p+1)!} - \beta_{p+1}\right] \tau^{p+1}\, c_k^{(p+1)}.$$

Suppose now that the global error ε_k contains eigenvectors of D corresponding to the large eigenvalues. Then $c_k^{(j)}$ is of order $[\sigma(D)]^j$ and τ should be chosen less than $1/\sigma(D)$ in order to get a correct value of $\rho'_k(\tau)$ by (2.8).

Furthermore, in the asymptotic region this value of $\rho_k'(\tau)$ is considerably larger than the value of $\rho_k(\tau)$. Hence, step size control based on the value of $\rho_k'(\tau)$ is too pessimistic in the asymptotic region.



(a) ε_k containing "late" eigenvectors (b) ε_k containing no "late" eigenvectors

fig. 2.3 The error $\rho_k'(\tau)$

In figure 2.3 (a) the errors $\rho_k' = \tilde{U}_{k+1}' - u_{k+1}$ and $\rho_k = \tilde{U}_{k+1} - u_{k+1}'$ are illustrated. Here u_{k+1}' denotes the numerical solution which is obtained when the integration starts in the point $(t_k, \tilde{U}_k)$.
Next, suppose that ε_k does not contain "late" eigenvectors. Then $c_k^{(j)}$ becomes small in the asymptotic region and $\rho_k'(\tau)$ is comparable with $\rho_k(\tau)$. (see figure 2.3 (b)). Our conclusion is that in order to get a reliable step size prediction in the asymptotic region, which also holds for $\tau > 1/\sigma(D)$, it is recommended to use polynomials which eliminates the "late" eigenvectors from ε_k as soon as the asymptotic region is reached.

2.3 Eigenvalues distributed over widely spaced clusters

The stiff systems of differential equations arising in physics and chemistry often have eigenvalues which are distributed over a few widely separated clusters. When a rough idea of the eigenvalue locations is available one can try to adapt the scheme to the problem to be solved. In this paper we consider the case where the eigenvalues occur in two clusters with center at the negative real axis and the case where they occur in three clusters two of which with conjugate complex centers and one with its center at the negative real axis.

Two-cluster case

Suppose that it is known that the eigenvalues are distributed as illustrated in figure 2.4 (shaded region).

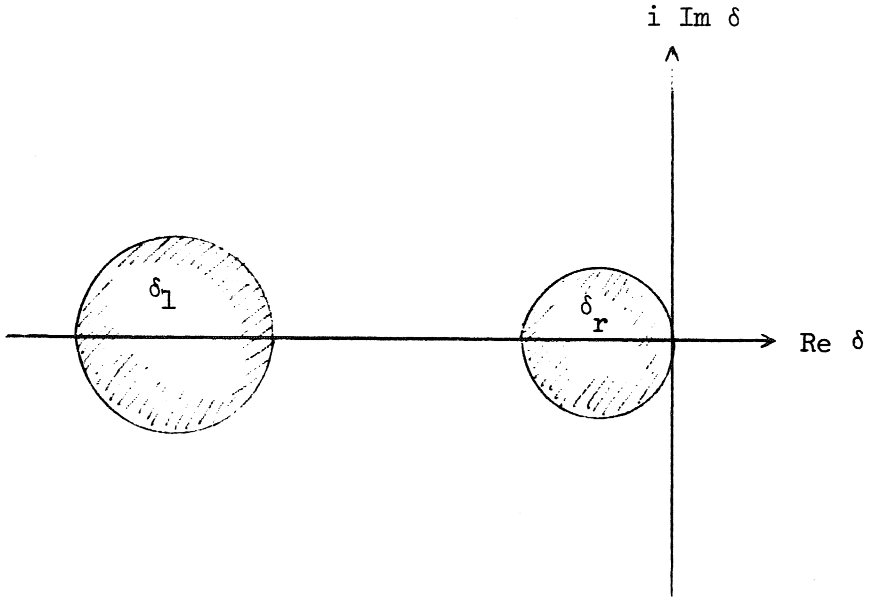


fig. 2.4 Eigenvalue locations in the two-cluster case

Furthermore, suppose that the cluster centers δ_l and δ_r are widely separated. Then, the asymptotic solution $\tilde{U}_A(t)$ will be composed of eigenvectors corresponding to the eigenvalues lying in the right hand cluster, whereas the vector $\tilde{U}(t_k) - \tilde{U}_A(t_k)$, occurring in the perturbation part of representation (2.1), consists of eigenvectors corresponding to eigenvalues of the left hand cluster. Therefore, in the neighbourhood of $t=t_k$ any solution of the differential equation can be approximated by

$$A_k + (t-t_k)B_k + e^{\delta_l(t-t_k)} C_k, \tag{2.9}$$

where A_k, B_k and C_k are appropriately chosen vectors only depending on k and not on τ. In particular, the solution $\tilde{U}'(t)$ starting in the point (t_k,u_k) can be represented in this way.
The two-cluster method, discussed in section 4.1, is in fact based on this representation.

Three-cluster case

When the eigenvalues are distributed as shown in figure 2.6, we may write for the solution $\tilde{U}'(t)$

$$\tilde{U}'(t) \approx A_k + (t-t_k)B_k + e^{\delta_l(t-t_k)} C_k' + e^{\bar{\delta}_l(t-t_k)} \bar{C}_k',$$

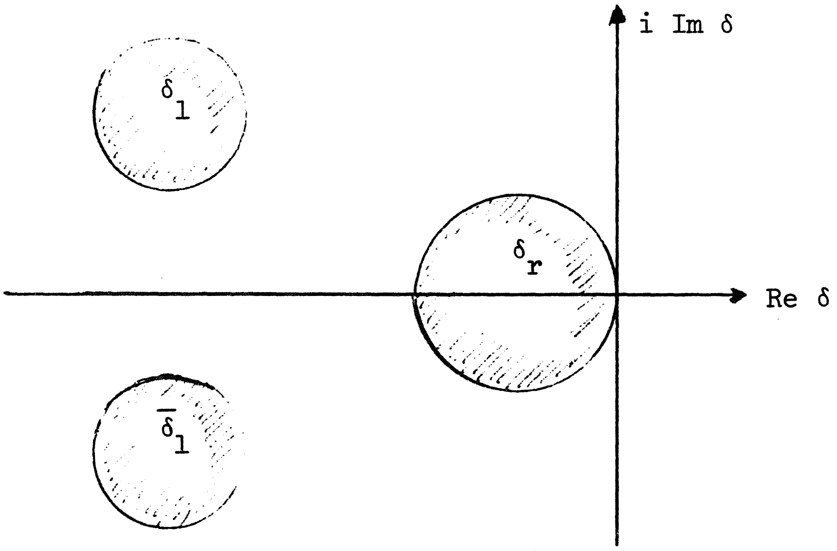


fig. 2.6 Eigenvalue locations in the three-cluster case

or equivalently

$$(2.10) \qquad \tilde{U}'(t) \simeq A_k + (t-t_k)B_k + \left[e^{\delta_1(t-t_k)+i\gamma} + e^{\bar{\delta}_1(t-t_k)-i\gamma} \right] C_k,$$

where A_k, B_k and C_k are appropriately chosen vectors and γ is a diagonal matrix with the arguments of the complex components of the vector C_k' as diagonal entries.

The three-cluster method discussed in section 4.4 uses this representation for the local analytical solution $\tilde{U}'(t)$.

3. Difference schemes with changing stability regions

The polynomials which generate the difference schemes discussed in reference [11], have constant coefficients β_j and, therefore, have a fixed stability region, i.e. the values of $\tau\delta$ are restricted to a fixed domain in the complex $z = \tau\delta$-plane.

In this section we allow the coefficients β_j to be functions of τ. This means that the local analytical solution $\tilde{U}'(t)$ at $t = t_k$ is no longer assumed to behave like a polynomial of degree n in $\tau = t-t_k$, but is represented in the more general form

$$(3.1) \qquad \tilde{U}'(t) = \tilde{U}'(t_k+\tau) \simeq u_k + \beta_1(\tau)\tau c_k^{(1)} + \beta_2(\tau)\tau^2 c_k^{(2)} + \dots + \beta_n(\tau)\tau^n c_k^{(n)}.$$

First, the problem is considered to choose the coefficients $\beta_j(\tau)$ such that for a given time step τ, prescribed by accuracy considerations, the scheme is stable, while the local discretization error is as small as possible. In section 4 we derive functions $\beta_j(\tau)$ which take into account the distribution of the eigenvalues of D.

3.1 Amplification of the stability region

Consider a generating polynomial $P_n(z)$ with stability region S, i.e.

$$(3.2) \qquad z \in S \rightarrow |P_n(z)| \leq 1.$$

Obviously, we have

$$(3.2') \qquad z \in \frac{1}{\beta_1} S \rightarrow |P_n(\beta_1 z)| \leq 1.$$

From this we immediately obtain

Theorem 3.1

Let $P_n(z)$ generate a difference scheme which is for $\tau \leqq \tau_0$ a stable scheme with respect to the given inital value problem. Then, the polynomial $P_n(\beta_1 z)$ with $\beta_1 = \tau_0/\tau_1$ generates a scheme which is stable for $\tau \leqq \tau_1$.

As an application of theorem 3.1 we consider the polynomial $A_4(z)$. In [11], section 3.1, it was shown that any initial value problem, in which D has eigenvalues with non-positive real parts, can be solved by this polynomial, provided that $\tau \leq 2.63/\sigma(D)$, $\sigma(D)$ being the spectral radius of D. According to the theorem, polynomial $A_4(2.63z/\tau\sigma(D))$ is stable for any time step τ.

Of course, we have to pay heavily for this property of unconditional stability. The accuracy of polynomials $P_n(\beta_1 z)$, $\beta_1 < 1$, is only of order zero, i.e.

$$(3.3) \qquad \rho_k(\tau) \sim (1-\beta_1)\tau\, \tilde{c}_k^{(1)}.$$

Thus, only when $c_k^{(1)}$ is small we may use values of β_1 less than 1. As was observed in the preceding section, stiff equations do have solutions for which in the asymptotic region $\tilde{c}_k^{(1)}$ is small. Hence polynomials of the above type may be useful. This will be illustrated for example (2.2) given in section 2.

From (2.3) it follows that

$$\tilde{U}(t_k) \sim 10^{-3} t^2{}_k \; , \; \tilde{c}_k^{(1)} \sim 2{\cdot}10^{-3} t_k \; , \; t_k >> 10^{-3}.$$

Hence, the local error satisfies the inequality

$$\rho_k(\tau) < \tau \tilde{c}_k^{(1)} = 2{\cdot}10^{-3} \; \tau t_k$$

and the relative local error the inequality

$$\frac{\rho_k(\tau)}{\tilde{U}_k} < \frac{2\tau}{t_k} .$$

Suppose that in the interval $1 \leq t_k \leq 20$ it is required to have locally an absolute error less than 10^{-9} and a relative error less than 1%. Then we see that the allowed time step in the interval $1 \leq t_k \leq 20$ increases linearly from .005 at $t_k = 1$ to .05 at $t_k = 10$ and decreases as $1/2t_k$ from .05 to .0025 at $t_k = 20$. It is easily verified that it requires less than 500 steps to integrate the interval $1 \leq t \leq 20$ with polynomials of type $P_n(\beta_1 z)$, whereas the polynomials $P_n(z)$ requires $19000/\beta(n)$ steps, $\beta(n)$ being a stability parameter associated with $P_n(z)$ (see [11], formula (2.25")). For instance, Euler's method ($\beta(n) = \beta(1) = 2$) requires 9500 steps, a factor 19 as much.

3.2 Weak and strong stability

According to the stability definition given in [11], section 2.4, the scheme generated by $P_n(\beta_1 z)$ is stable when

$$(3.4) \qquad \sigma(P_n(\beta_1 \tau D)) \leq 1.$$

Furthermore, it was shown that in cases where $\sigma(P_n(\beta_1 \tau D)) = 1$ the global error may increase at least linearly with k and is proportional to the maximal local error. Nowever, when $\sigma(P_n(\beta_1(\tau D)) < 1$, the global and local error are of the same order of magnitude as $k \to \infty$. These cases are sometimes distinguished as the weakly stable and strongly stable case (compare O'Brien, Hyman and Kaplan [10]). In using polynomials $P_n(\beta_1 z)$ with $\beta_1 < 1$, which produce relative large local errors, it may be advisable to choose $P_n(z)$ such that $\sigma(P_n(\beta_1 \tau D))$ is less than 1 in order to make the scheme strongly stable.

For example, when D has only eigenvalues with negative real parts, the polynomial $A_4(\beta_1 z)$ generates a strongly stable scheme, provided that β_1 is less than, instead of equal to, $2.63/\tau\sigma(D)$.

In cases where D has real negative eigenvalues it was recommended in [11] to use the polynomial $T_n(1+x/n^2)$. We shall give a theorem which defines a strongly stable modification of $T_n(1+x/n^2)$. For that purpose, we need the concept of the condition of a matrix.

Definition 3.1

The spectral condition number s(D) of a matrix D is defined by

$$(3.5)\qquad s(D) = \frac{|\delta|_{\max}}{|\delta|_{\min}} .$$

Furthermore, we introduce the quantities

$$(3.6)\qquad w_0 = \frac{\sqrt{s(D)}+1}{\sqrt{s(D)}-1} , \; w_1 = \frac{\sqrt{s(D)}}{n^2(\sqrt{s(D)}-1)} \cdot \frac{\frac{2n}{\sqrt{s(D)}}}{\tanh\left[\frac{2n}{\sqrt{s(D)}}\right]} .$$

Theorem 3.2

The polynomial

$$P_n(x) = \frac{T_n(w_0+w_1 z)}{T_n(w_0)}$$

has the following properties:

(a) It is first order exact.

(b) It satisfies the stability condition $\tau \leq \beta(n)/\sigma(D)$, where

$$\beta(n) \sim 2n^2 \; \frac{\tanh\left[\frac{2n}{\sqrt{s(D)}}\right]}{\frac{2n}{\sqrt{s(D)}}} \quad \text{as } s(D) \to \infty .$$

(c) The spectral radius of $P_n(\tau D)$, τ satisfying the stability condition, is given by

$$\sigma(P_n(\tau D)) \sim \frac{1}{\cosh\left[\frac{2n}{\sqrt{s(D)}}\right]} .$$

(d) Of all first order exact polynomials of degree n in x, $P_n(x)$ has the smallest maximum norm over the interval $-\beta(n) \leq x \leq -\beta(n)/s(D)$.

Proof See [10], p.39 ff.

For a discussion of this theorem we again refer to [10], p.39 ff. Here, only the explicit expressions for the polynomials $P_2(x)$ and $P_3(x)$ are given:

$$(3.7)\qquad \begin{cases} P_2(x) = 1 + x + \dfrac{2w_1^2}{2w_0^2 - 1}\, x^2, \\[2ex] P_3(x) = 1 + x + \dfrac{12w_1^2}{4w_0^2 - 3}\, x^2 + \dfrac{4w_1^3}{w_0(4w_0^2 - 3)}\, x^3. \end{cases}$$

3.3 Matching accurate and stable schemes together

As was already observed at the beginning of this section our aim is to choose the coefficients $\beta_j(\tau)$ such that for an arbitrary time step the scheme is stable and has a minimal local error as well. We are now in a position to be more specific. We have at our disposal difference schemes which are, for a given degree n of the generating polynomial, as accurate as possible (Runge-Kutta methods) and we possess unconditionally stable schemes. The next step is to match these schemes together in order to obtain a scheme which is as accurate as a Runge-Kutta method as $\tau \to 0$ and which is stable for unrestricted time steps.

Eigenvalues with negative real parts

Suppose that it is known that D has its eigenvalues in the negative half-plane. Further, let an estimate of $\sigma(D)$ be given. Then, the following polynomial can be used for an arbitrary time step τ:

$$(3.8)\qquad \begin{cases} P_4(z) \equiv A_4(z) \text{ if } \tau \leq \dfrac{\beta}{\sigma(D)}, \\[2ex] P_4(z) \equiv A_4\left(\dfrac{\beta z}{\tau\sigma(D)}\right) \text{ if } \tau \geq \dfrac{\beta}{\sigma(D)}, \end{cases}$$

where β is a constant less than 2.63.

Note that the coefficients β_j of $P_4(z)$ are changing continuously when τ increases, so that the local errors are also changing continuously.

Negative eigenvalues

Suppose that the eigenvalues of D are negative. We then can construct a more gradual transition from the Runge-Kutta type polynomials to the unconditionally stable polynomials.

We start with an n-th order Runge-Kutta method. In an actual computation, this scheme can be used until the time-step τ, prescribed by a local error analysis, exceeds the maximal step allowed by the stability condition of the scheme. Then, by changing the coefficient β_n the stability properties improve until a certain critical value is reached. Figure 3.1 shows how the new value of β_n is determined for a given value of τ.

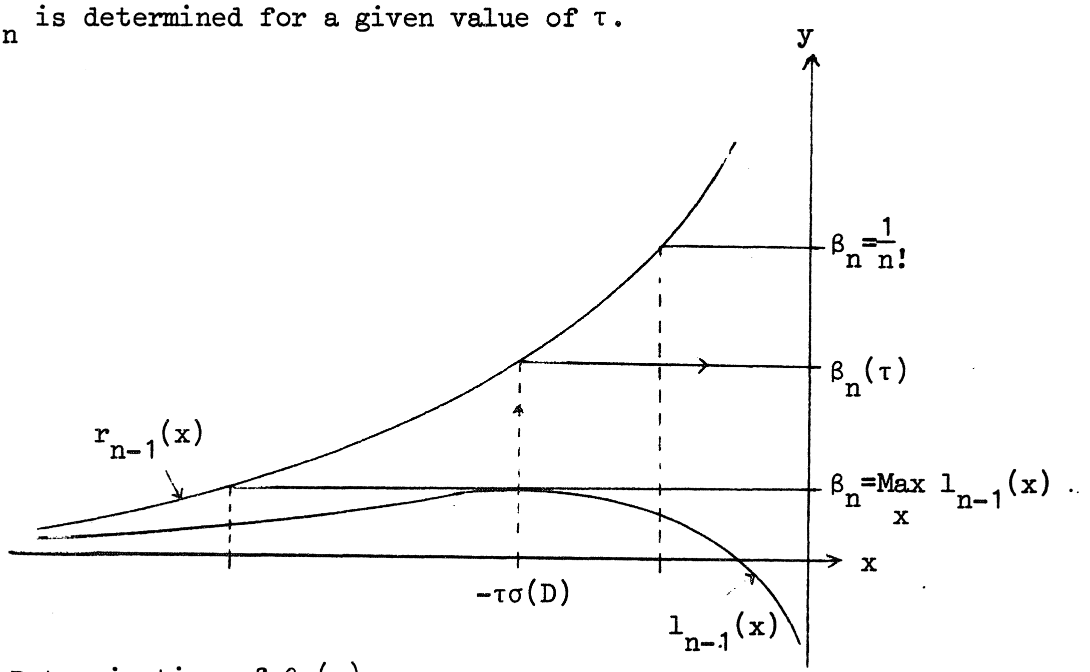


fig. 3.1 Determination of $\beta_n(\tau)$.

Let $r_{n-1}(x)$ and $l_{n-1}(x)$ be the functions defined in [11], formula (6.2) and let τ be the prescribed step. Then, we see from this figure that

$$(3.9)\qquad \begin{cases} \beta_n(\tau) = \frac{1}{n!} & \text{for } r_{n-1}(-\tau\sigma(D)) \geq \frac{1}{n!}\ , \\ \beta_n(\tau) = r_{n-1}(-\tau\sigma(D)) & \text{for } \underset{x}{\mathrm{Max}}\ l_{n-1}(x) \leq r_{n-1}(-\tau\sigma(D)) \leq \frac{1}{n!}\ . \end{cases}$$

The scheme which arises for the critical value $\beta_n = \underset{x}{\mathrm{Max}}\ l_{n-1}(x)$ is just the (n-1)-st order Runge-Kutta scheme with one stability term as treated in [11], section 6.2.

During the transition of β_n from 1/n! to $\underset{x}{\mathrm{Max}}\ l_{n-1}(x)$, the local error is approximately given by

$$(3.10)\qquad \rho_k(\tau) \sim \frac{1}{(n+1)!}\,\tau^{n+1}\,\tilde{c}_k^{(n+1)} + \left(\frac{1}{n!} - \beta_n\right)\tau^n\,\tilde{c}_k^{(n)}\,.$$

When the value $r_{n-1}(-\tau\sigma(D))$ drops below $\operatorname{Max}_x l_{n-1}(x)$ the scheme becomes unstable. To restore stability we have to change both β_n and β_{n-1}. Figure 3.2 explains how $\beta_n(\tau)$ and $\beta_{n-1}(\tau)$ are determined.

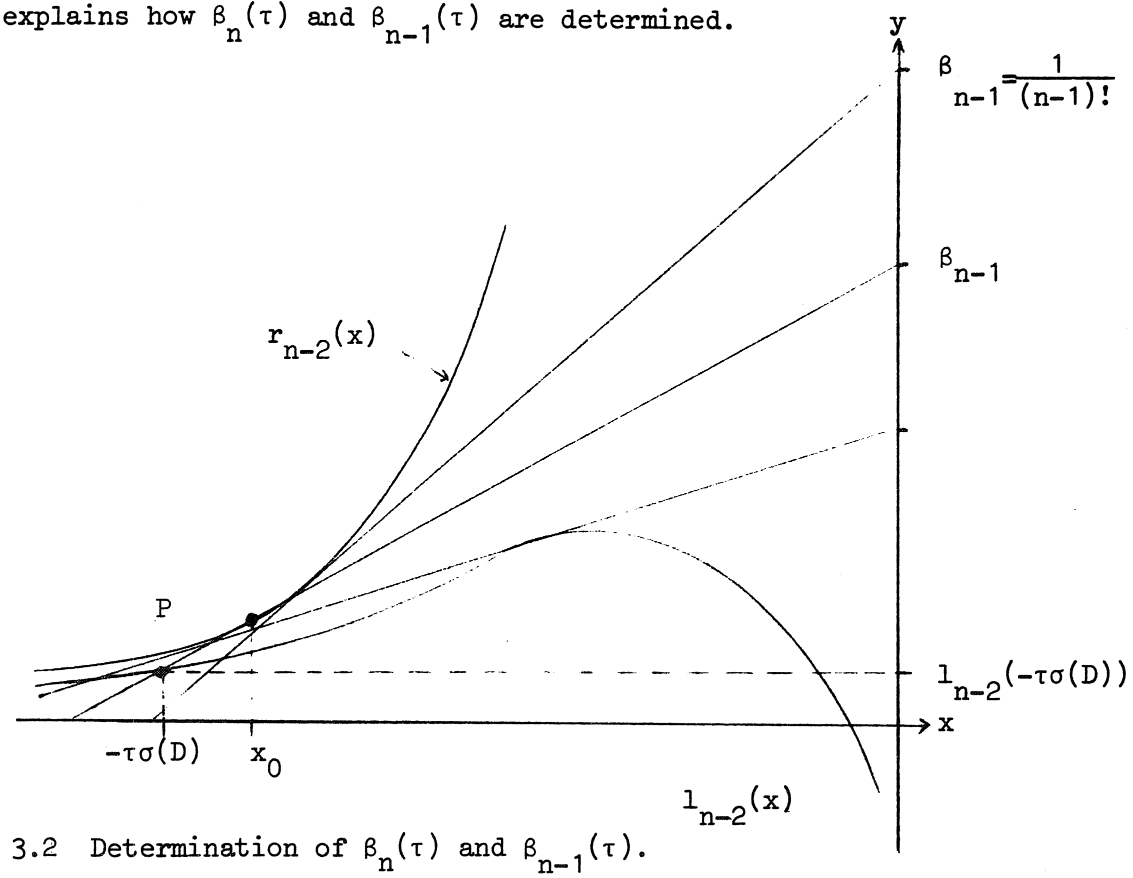


fig. 3.2 Determination of $\beta_n(\tau)$ and $\beta_{n-1}(\tau)$.

Let P be that point in figure 3.2 which has the coordinates $(-\tau\sigma(D), l_{n-2}(-\tau\sigma(D)))$, τ being the desired time step, and let $y = ax + b$ be the line through the point P which is tangent to the curve $y = r_{n-2}(x)$. Then the coefficients β_n and β_{n-1} are set equal to a and b, respectively. This yields

$$(3.11) \qquad \begin{cases} \beta_n(\tau) = r'_{n-2}(x_0), \\ \\ \beta_{n-1}(\tau) = l_{n-2}(-\tau\sigma(D)) + \tau\sigma(D)\, r'_{n-2}(x_0), \end{cases}$$

where x_0 is the solution of the equation

$$(3.12) \qquad r_{n-2}(x_0) = r'_{n-2}(x_0)x_0 + l_{n-2}(-\tau\sigma(D)) + \tau\sigma(D)r'_{n-2}(x_0).$$

When the line $y = \beta_n x + \beta_{n-1}$ touches both $r_{n-2}(x)$ and $l_{n-2}(x)$ we have the (n-2)-nd order Runge-Kutta method with two stability terms discussed in [11], section 6.3.

As soon as the line $y = \beta_n x + \beta_{n-1}$ intersects the curve $y = l_{n-2}(x)$ the scheme becomes unstable and one has to change the three coefficients β_n, β_{n-1} and β_{n-2}.

In this way we may proceed until we arrive at the unconditionally stable polynomials discussed in the preceding section. However, the considerations become increasingly more difficult. Therefore, we restrict our attention to the cases discussed above.

This section is concluded with an application of the preceding ideas to polynomials of second and third degree. In table 3.1 and 3.2 the coefficients $\beta_j(\tau)$ are explicitly given.

Table 3.1 $\quad P_2(x) = 1 + \beta_1(\tau)x + \beta_2(\tau)x^2$

Range of $\tau\sigma(D)$	$\beta_1(\tau)$	$\beta_2(\tau)$
$[0\ ,\ 2]$	1	$\frac{1}{2}$
$[2,\ \beta(2)]$	1	$\frac{1}{\tau\sigma(D)}$
$[\beta(2),\ \infty]$	$\frac{\beta(2)}{\tau\sigma(D)}$	$\beta_1^2(\tau)\ \frac{2w_1^2}{2w_0^2 - 1}$

The parameters w_0, w_1, $\beta(2)$ and $\beta(3)$ occurring in these tables are defined by formula (3.6) and theorem 3.2.

Table 3.2 $P_3(x) = 1 + \beta_1(\tau)x + \beta_2(\tau)x^2 + \beta_3(\tau)x^3$

Range of $\tau\sigma(D)$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_3(\tau)$
$[0.254]$	1	$\frac{1}{2}$	$\frac{1}{6}$
$[2.54, 6.27]$	1	$\frac{1}{2}$	$\frac{[\tau\sigma(D)]^2 - 2\tau\sigma(D) + 4}{2[\tau\sigma(D)]^3}$
$[6.27, \beta(3)]$	1	$\frac{2[\tau\sigma(D) - 2]}{\tau\sigma(D)[\tau\sigma(D) - \sqrt{2\tau\sigma(D)}]}$	$\frac{1}{4}\beta_2^2(\tau)$
$[\beta(3), \infty]$	$\frac{\beta(3)}{\tau\sigma(D)}$	$\beta_1^2(\tau)\ \frac{12w_1^2}{4w_0^2 - 3}$	$\beta_1^3(\tau)\ \frac{4w_1^3}{w_0(4w_0^2 - 3)}$

4. Two-point approximations of the exponential operator

The method of changing parameters takes into account that the solution of the differential equation may behave quiet differently in the interval of integration and is, for that reason, appropriate to integrate equations which have widely separated eigenvalues. In many cases, however, the eigenvalues of the matrix D are not only widely separated, but moreover, they can be placed in two groups, a group of "small" eigenvalues, and a group of "large" eigenvalues. By using this property one can again save computer time. We shall consider two important cases; firstly, the case where the eigenvalues occur in two clusters which are situated near the origin of the δ-plane and at a point $x = -b$, $b >> 0$; secondly, the case where the cluster at $x = -b$ is split up into two clusters of conjugate complex eigenvalues.

4.1 The two-cluster method

We assume the generating polynomial $P_n(x)$ to be of the form

$$(4.1)\qquad \begin{cases} P_n(x) = A_p(x) + x^{p+1}\, B_q(x),\ n = p + q + 1, \\ A_p(x) = 1 + x + \frac{1}{2!}x^2 + \ldots + \frac{1}{p!}x^p, \end{cases}$$

and we recall that the generated scheme is stable when

$$(4.2)\qquad l_p(x) \le B_q(x) \le r_p(x),\ x = \tau\delta_j,\ j = 1,2,\ldots,m,$$

where the δ_j are the eigenvalues of D (compare [11], formula (6.2)).

Let τ be prescribed by an accuracy condition and suppose that the values of $\tau\delta_j$ can be placed in two clusters centered at $x = \tau\delta_r \sim 0$ and $x = \tau\delta_1$, respectively. Let $b = -\tau\delta_1$, then the polynomial $B_q(x)$ should be chosen in such a way that it remains between $r_p(x)$ and $l_p(x)$ in the neighbourhood of x = 0 and x = -b (see figure 4.1).

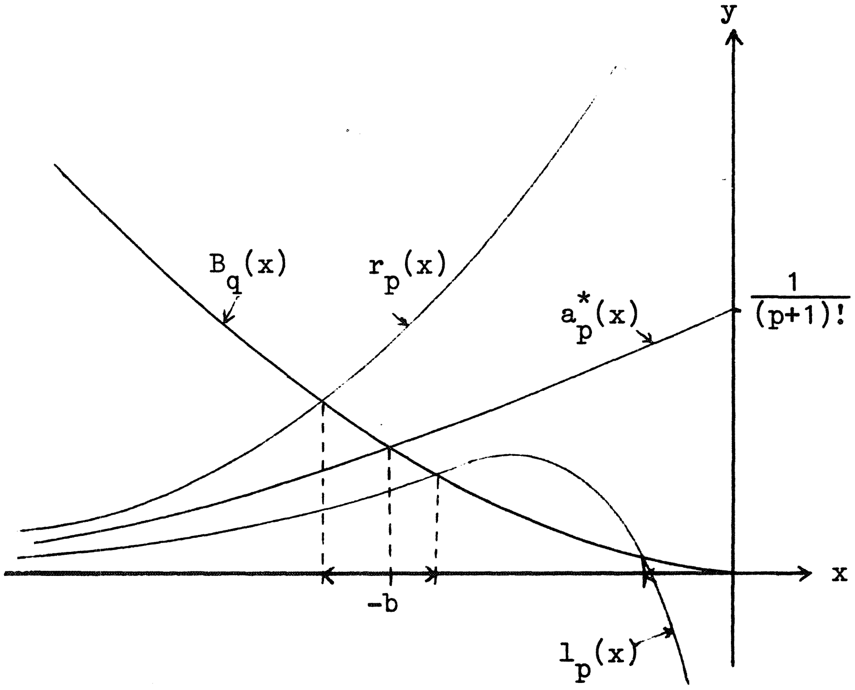


fig. 4.1 The functions $a_p^*(x)$ and $B_q(x)$

An additional condition to be imposed on $B_q(x)$ is that for $b \to 0$ $B_q(x)$ is asymptotically equivalent to the function

$$(4.2) \qquad a_p^*(x) = \frac{e^x - A_p(x)}{x^{p+1}}, \quad -b \leq x \leq 0.$$

Otherwise, the scheme is not a consistent approximation of the differential equation. In order to meet these conditions, we identify $B_q(x)$ with the first q+1 terms of the Taylor expansion of $a_p^*(x)$ at $x = -b$. Note that for large negative values of x the functions $a_p^*(x)$ resembles the average function $a_p(x)$ introduced in [11], section 6.4. Here, we cannot use $a_p(x)$ itself, because it becomes singular as $b \to 0$.

For future reference some of the corresponding polynomials $P_n(x)$ are given explicitly.

$$(4.3) \qquad \begin{cases} P_2(x) = 1 + x + \dfrac{e^{-b} -1 + b}{b^2} x^2, \\ P_3(x) = 1 + x + \dfrac{e^{-b}(b+3) - 3 + 2b}{b^2} x^2 + \dfrac{e^{-b}(b+2) - 2 + b}{b^3} x^3, \\ P_3(x) = 1 + x + \frac{1}{2} x^2 + \dfrac{-2e^{-b} + 2 - 2b + b^2}{2b^3} x^3, \\ P_4(x) = 1 + x + \frac{1}{2} x^2 + \dfrac{-e^{-b}(b+4) + b^2 - 3b + 4}{b^3} x^3 + \\ \qquad\qquad + \dfrac{-2e^{-b}(b+3) + b^2 - 4b + 6}{2b^4} x^4. \end{cases}$$

In fact, the choice of $B_q(x)$ as given above means that we approximate e^x by a polynomial $P_n(x)$ which is a two-point Taylor expansion of e^x at $x = 0$ and $x = -b$ (see figure 4.2). The idea of such an "exponential fitting" was already used by Pope [20] in 1963, Fowler and Warten [7] in 1967, and Liniger and Willoughby [17] also in 1967. The methods of Pope and Fowler-Warten are both based on the polynomial $P_2(x)$ mentioned above. However, Pope used a Taylor expansion of the operator $e^{\tau D}$ instead of the scalar e^{-b} and Fowler and Warten used a diagonal matrix representation of $e^{\tau D}$. Liniger and Willoughby considered "exponential fitting" in connection with implicit one-step methods.

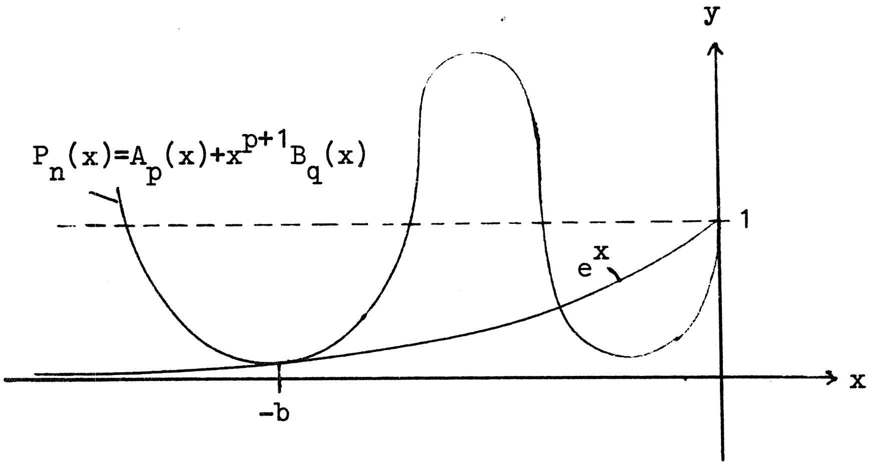


fig. 4.2 Generating polynomial $P_n(x)$

4.2 Regions of stability of the two-cluster method

It is of interest to know how far the stability regions, located at $x = -b$ and $x = 0$, extend along the x-axis for values of b. We shall consider the cases $p = 1$ and $p = 2$.

Since b is supposed to be large we may replace $a_p^*(x)$ by $a_p(x)$ and apply the method of analysis given in [11], section 6.4. In this way we get information about the stability region at $x = -b$.

When $p = 1$ we may write

$$(4.4)\qquad a_1^*(x) \sim a_1(x) = \sum_{j=0}^{\infty} \frac{b-j-1}{b^2}\left(\frac{x+b}{b}\right)^j \sim B_q(x) + R_{q+1}(x),$$

where the remainder $R_{q+1}(x)$ has the properties

$$(4.5)\qquad |R_{q+1}(x)| \leq \begin{cases} \frac{b-q-2}{b^2}\left(\frac{b}{-x}\right)\left(\frac{x+b}{b}\right)^{q+1} & , x \geq -b \\ \frac{b-q-2}{b^2}\left(-\frac{x+b}{b}\right)^{q+1} & , x \leq -b. \end{cases}$$

The stability region at $x = -b$ is determined by those two zeroes x_1 and x_2 of

(4.6) $$|R_{q+1}(x)| = \frac{1}{x^2}$$

which are directly to the left and to the right of x = -b. From (4.5) and (4.6) we have

(4.7) $$\begin{cases} x_2 = -b\left[1 - \left(\frac{b}{-x_2(b-q-2)}\right)^{\frac{1}{q+1}}\right] \sim -b\left[1 - b^{\frac{-1}{q+1}}\right], \\ x_1 = -b\left[1 + \left(\frac{b^2}{x_1^2(b-q-2)}\right)^{\frac{1}{q+1}}\right] \sim -b\left[1 + b^{\frac{-1}{q+1}}\right]. \end{cases}$$

Hence, the width $d_l(b)$ of the left stability region is approximately given by

(4.8) $$d_l(b) = |x_l - x_r| \sim 2\, b^{\frac{q}{q+1}}, \; b >> 1.$$

When this criterion is applied to a cluster of eigenvalues δ_j with its cénter at δ_l and a maximal diameter $d\delta_l$ we obtain the stability condition

(4.8') $$\tau \leq \frac{2^{q+1}|\delta|^q}{(d\delta_l)^{q+1}}.$$

This condition guarantees that the eigenfunction components in the discretization error, which correspond to eigenvalues of the cluster $(\delta_l, d\delta_l)$, do not increase in the numerical calculation.

For p > 1 a similar analysis can be given. When p = 2 the final result is

(4.9) $$d_l(b) \sim 2\, b^{\frac{q-1}{q+1}}, \; b >> 1,$$

(4.9') $$\tau \leq \frac{2^{\frac{q+1}{2}}|\delta_l|^{\frac{q-1}{2}}}{(d\delta_l)^{\frac{q+1}{2}}}.$$

Next, the width of the stability region at x = 0 is estimated. From formula (4.4) and [11], formula (6.16) it follows that for large values of b and $|x| \ll b$

$$(4.10) \qquad B_q(x) \sim \begin{cases} \frac{q+1}{b} & \text{if } p = 1, \\ \frac{q+1}{2b} & \text{if } p = 2. \end{cases}$$

Hence, the stability region covers at least the stability region of $A_p(x)$, $p = 1,2$, i.e. the interval $[-2,0]$ (compare [11], figure 3.1). For a cluster of eigenvalues of diameter $d\delta_r$ this gives rise to the stability condition

$$(4.11) \qquad \tau \leq \frac{2}{d\delta_r} .$$

Before we draw conclusions from the results given above we first consider the accuracy of the difference schemes.

4.3 Order of accuracy of the two-cluster method

From (4.3), expressions for the local discretization error $\rho_k(\tau)$, as defined in [11], formula (2.14), are easily derived. In table 4.1 the first terms of the Taylor expansion of these expressions are given, together with the stability conditions derived in the preceding subsection.

Table 4.1 Discretization errors and stability conditions for the polynomials (4.3).

(n,p)	$\rho_k(\tau)$	stability conditions	
(2.1)	$\frac{1}{6}\tau^3[\tilde{c}_k^{(3)} - \delta_1\tilde{c}_k^{(2)}]$	$\tau \leq \frac{2}{d\delta_1}$.	$\tau \leq \frac{2}{d\delta_r}$
(3.1)	$\frac{1}{24}\tau^4[\tilde{c}_k^{(4)} - 2\delta_1\tilde{c}_k^{(3)} + \delta_1^2\tilde{c}_k^{(2)}]$	$\tau \leq \frac{4\|\delta_1\|}{(d\delta_1)^2}$	
(3.2)	$\frac{1}{24}\tau^4[\tilde{c}_k^{(4)} - \delta_1\tilde{c}_k^{(3)}]$	$\tau \leq \sqrt{\frac{2}{\|\delta_1\|d\delta_1}}$	
(4.2)	$\frac{1}{120}\tau^5[\tilde{c}_k^{(5)} - 4\delta_1\tilde{c}_k^{(4)} + \delta_1^2\tilde{c}_k^{(3)}]$	$\tau \leq \frac{2}{d\delta_1}$	

Table 4.1 shows that the corresponding difference schemes are respectively of order 2, 3, 3 and 4. However, the error constants may be very large due to the large value of $|\delta_l|$. For instance, of the schemes characterized by (n,p) = (3.1) and (n,p) = (3.2), although both of order 3, the latter will in general have a considerable smaller error constant.

The stability conditions associated with these schemes are very weak, provided that the cluster diameters $d\delta_l$ and $d\delta_r$ are relatively small. As an illustration we consider the case where $d\delta_l$ and $d\delta_r$ are at most 1/10 of the spectral radius $\sigma(D)$. We then get for the maximal allowed effective time step $(\tau_{eff} = \tau/n)$, respectively $(\sigma(D) \sim |\delta_l|)$

$$(4.12) \qquad \frac{10}{\sigma(D)}, \ \frac{6.67}{\sigma(D)}, \ \frac{1.5}{\sigma(D)} \text{ and } \frac{5}{\sigma(D)}\ .$$

If, however, the cluster diameters are only 5% of the spectral radius $\sigma(D)$ we find the considerably larger steps:

$$(4.12') \qquad \frac{20}{\sigma(D)}, \ \frac{13.3}{\sigma(D)}, \ \frac{2.1}{\sigma(D)} \text{ and } \frac{10}{\sigma(D)}\ .$$

Further, we observe that for comparable diameters of the two clusters, it is the right hand cluster which prescribes the maximal allowable time step in the cases (3.1) and (3.2).

Finally, we remark that the eigenvalues δ_j may have imaginary parts. For the cluster at the origin this immediately follows from the stability regions of $A_1(z)$ and $A_2(z)$. Only when purely imaginary eigenvalues are present, the two-cluster methods described above become unstable unless b is sufficiently small. To obtain stability for large values of b we have to start with polynomials in which $p \geq 3$. Complex eigenvalues in the left cluster leads us to the three cluster method.

4.4 The three-cluster method

In this section the case is considered where the eigenvalues can be placed in a cluster at $\delta = \delta_r \sim 0$ and in two clusters centered at $\delta = \delta_l$ and $\delta = \bar{\delta}_l$. We try to fit the polynomial

$$P_3(z) = 1 + z + \beta_2 z^2 + \beta_3 z^3$$

with the exponential function e^z at $z = \tau\delta_1$, thus

$$e^{\tau\delta_1} = P_3(\tau\delta_1). \tag{4.13}$$

Note that two coefficients β_2 and β_3 are needed in order to satisfy (4.13) with real values of β_2 and β_3.
Equation (4.13) may be written in the form

$$\beta_2 + \beta_3\,\tau\delta_1 = \frac{e^{\tau\delta_1} - 1 - \tau\delta_1}{\tau^2\delta_1{}^2} \tag{4.13'}$$

Writing $\tau\delta_1$ as $b\,e^{i\phi}$ and equating real and imaginary parts we obtain

$$\left\{\begin{aligned} \beta_3 &= \frac{1}{b^2} + \frac{2\cos\phi}{b^3} + e^{b\cos\phi}\cdot\frac{\sin(b\sin\phi - 2\phi)}{b^2\sin\phi}, \\ \beta_2 &= \frac{-2\cos\phi}{b} - \frac{4\cos^2\phi - 1}{b^2} - e^{b\cos\phi}\cdot\frac{\sin(b\sin\phi - 3\phi)}{b^2\sin\phi}. \end{aligned}\right. \tag{4.14}$$

When $b\cos\phi << 0$ the last term in these expressions for β_2 and β_3 may be neglected. From a practical point of view this means a considerable saving of computertime, as the exponential function is very expensive.

Further, when b is small we compute β_2 and β_3 by means of the following approximations:

$$\left\{\begin{aligned} \beta_2 &\sim \frac{1}{2} - \frac{1}{24}b^2 = \frac{1}{2} - \frac{1}{24}|\delta_1|^2\tau^2, \\ \beta_3 &\sim \frac{1}{6} + \frac{1}{12}b\cos\phi = \frac{1}{6} + \frac{1}{12}\operatorname{Re}\delta_1\,\tau. \end{aligned}\right. \tag{4.14'}$$

These approximations may also be used to estimate the local discretization error as $\tau \to 0$. We have

$$(4.15) \qquad \rho_k(\tau) \sim (\tfrac{1}{2} - \beta_2)\tau^2 \, \tilde{c}_k^{(2)} + (\tfrac{1}{6} - \beta_3)\tau^3 \, \tilde{c}_k^{(3)} + \tfrac{1}{24}\,\tau^4 \, c_k^{(4)}$$

$$\sim \tfrac{1}{24}\,(|\delta_1|^2 \, \tilde{c}_k^{(2)} - 2 \operatorname{Re} \delta_1 \, \tilde{c}_k^{(3)} + \tilde{c}_k^{(4)})\tau^4.$$

Thus, (4.14) defines a third order accurate difference scheme.

In order to determine the stability region at $z = \tau\delta_1$ we write $P_3(z)$ in the equivalent form

$$P_3(z) = P_3(\tau\delta_1) + (z - \tau\delta_1)\Big[(1 + 2\beta_2\,\tau\delta_1 + 3\beta_3\tau^2\delta_1^2)$$

$$+ (\beta_2 + 3\beta_3\,\tau\delta_1)(z - \tau\delta_1) + \beta_3(z - \tau\delta_1)^2\Big].$$

First, the case $\sin\phi = 0$, $b \cos\phi << 0$ will be considered. From (4.13) and (4.14) it follows that in the neighbourhood of $z = -b$, $P_3(z)$ can be written as

$$P_3(z) \sim (z + b)\Big[(-\tfrac{1}{b} + \tfrac{3}{b^2})(z + b) + (\tfrac{1}{b^2} - \tfrac{2}{b^3})(z + b)^2\Big].$$

Formula (4.8) indicates that the stability region contains a circle of radius $O(\sqrt{b})$. Therefore, let us restrict z to the circle

$$(4.16) \qquad |z + b| \leq \sqrt{b}.$$

In this circle $P_3(z)$ can be approximated by

$$P_3(z) \sim -\frac{(z + b)^2}{b}.$$

Hence, $|P_3(z)| \leq 1$ in circle (4.16). This means that (4.16) defines the stability region when $\sin\phi = 0$. Note that (4.16) has the same diameter as the stability interval derived for the case of real eigenvalues (compare formula (4.8) with $q = 1$).

Next, we consider the case $\sin\phi \neq 0$, $b\cos\phi << 0$. From (4.14) it follows that

$$P_3(z) \sim (z - \tau\delta_1)\left[(1 + 2\beta_2\,\tau\delta_1 + 3\beta_3\tau^2\delta_1^2) - O(\tfrac{1}{b})\right]$$

$$= (z - \tau\delta_1)\left[2\,i\,e^{i\phi}\sin\phi + O(\tfrac{1}{b})\right].$$

This leads to the stability circle

$$(4.17)\qquad |z - \tau\delta_1| \leq \frac{1}{2|\sin\phi|}.$$

When this criterion is applied to a cluster of diameter $d\delta_1$ we obtain the stability condition

$$(4.18)\qquad \tau \leq \frac{1}{d\delta_1|\sin\phi|}\,.$$

This condition indicates that the maximal allowed time step becomes smaller when the cluster of eigenvalues is at a larger distance of the real axis. For example, when $\phi = 2\pi/3$ condition (4.16) becomes $\tau \leq 1.14/d\delta_1$, while for $\phi = 5\pi/6$ it becomes $\tau \leq 2/d\delta_1$.
As for the cluster near the origin, the polynomial behaves as $1 + z$ ($b >> 1$), hence the stability region at this part of the z-plane equals the stability region of Euler's method. This means that purely imaginary eigenvalues near the origin cannot be treated by the three-cluster method as described here. As was already observed in the preceding section we then have to start with polynomials of the type $A_p(z) + \beta_{p+1}z^{p+1} + \beta_{p+2}z^{p+2}$, where $p \geq 3$.

In figure 4.3 and 4.4 the stability regions of the polynomial $P_3(z)$ are given for some values of b and ϕ.

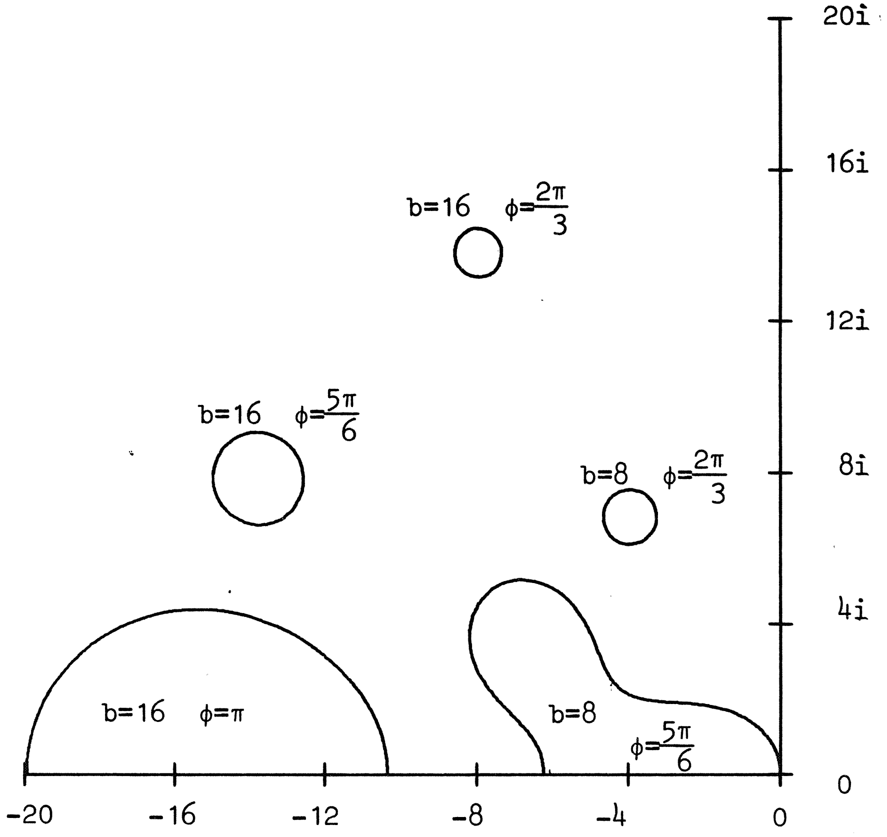


fig. 4.3 Left hand stability regions

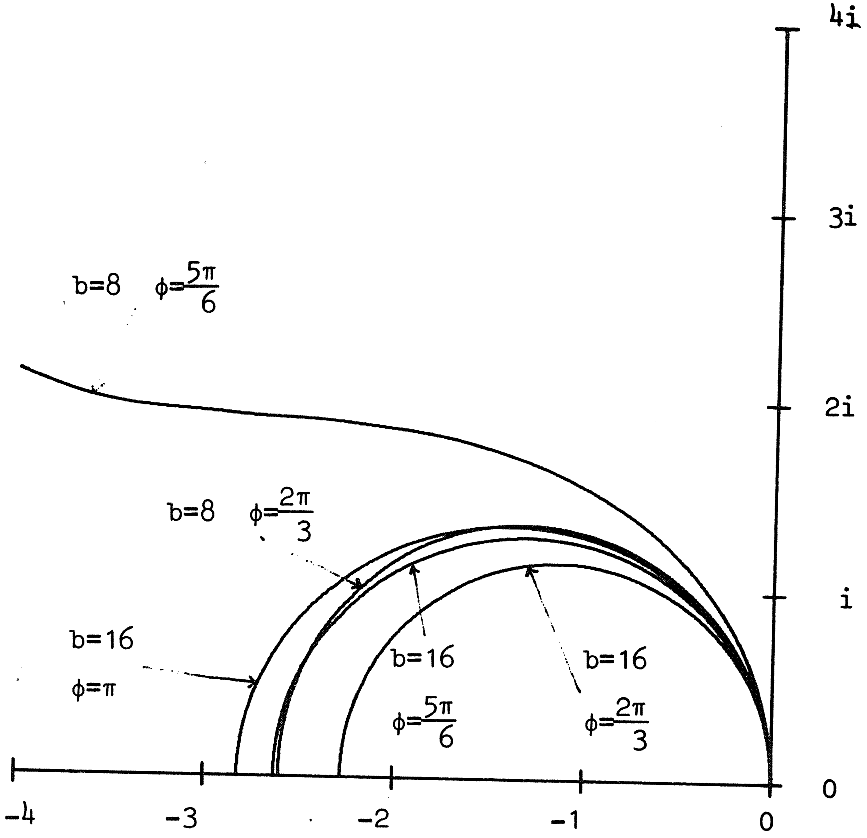


fig. 4.4 Right hand stability regions

5. Remarks on the location of the eigenvalues of the operator D

The application of the methods described in [11] and the preceding sections requires some information about the eigenvalues of D. For instance, when Chebyshev polynomials are used the eigenvalues have to be real or "almost real" and, in addition, a rough knowledge of the value of $\sigma(D)$ is required. The application of the schemes described in [11], section 5. presupposes purely imaginary eigenvalues and again an estimate of $\sigma(D)$ is desired. In the two- and three- cluster methods the location of the left hand clusters is required with some accuracy.

In the analysis of matrices a large number of useful theorems is developed which give such information. We mention the works of Bodewig [1], Zurmühl [24], Marcus [18].

In this section we give an additional method which enables us to calculate approximately the center δ_1 of a left hand cluster in the first step of the integration process. This method only applies to cases where the eigenvalues can reasonably be placed in two of three clusters in the sense as described in the preceding sections.

5.1 Determination of δ_1 in the two-cluster method

Our starting point is the representation

$$(5.1)\qquad \tilde{U}'(t_k+\tau) \cong A_k + B_k\tau + e^{\delta_1\tau}\, C_k$$

for the integral curve $\tilde{U}'$ through the point (t_k,u_k). In section 2 it was shown that this representation holds for the initial part of the curve, provided that the eigenvalues of D occur in two widely spaced clusters near the real axis.

Theorem 5.1

When (5.1) is a reasonable representation of the local analytical solution, then the value of δ_1 satisfies the relation

$$(5.2)\qquad \delta_1\, c_k^{(2)} \cong c_k^{(3)}.$$

Proof

From the definition of $c_k^{(2)}$ and $c_k^{(3)}$, and formula (5.1) we have

$$(5.3) \qquad \begin{cases} c_k^{(2)} = \dfrac{d^2}{dt^2} \tilde{U}'(t_k) \cong \delta_1^2 \, C_k, \\[2ex] c_k^{(3)} = \dfrac{d^3}{dt^3} \tilde{U}'(t_k) \cong \delta_1^3 \, C_k. \end{cases}$$

From these relations (5.2) immediately follows.
Note that $c_k^{(0)}$ (= u_k) and $c_k^{(1)}$ determine the values of A_k and B_k.

Formula (5.2) determines δ_1 more accurate as (5.1) is a better representation.

When we deal with a single differential equation we simply have

$$(5.2') \qquad \delta_1 \cong \frac{c_k^{(3)}}{c_k^{(2)}} .$$

In dealing with systems of equations (5.3) is an overdetermined set of equations. One possibility is to choose δ_1 such that the value of

$$\| \, \delta_1 \, c_k^{(2)} - c_k^{(3)} \|$$

is minimized. For the Euclidean norm $\| \; \|_2$ this leads to

$$(5.2') \qquad \delta_1 \cong \frac{(c_k^{(2)}, c_k^{(3)}}{\|c_k^{(2)}\|_2^2}$$

Another possibility is to allow δ_1 to be a diagonal matrix instead of a scalar. The derivation of (5.2) is not changed when the scalar δ_1 is replaced by a matrix δ_1. Hence (5.3') now determines the diagonal entries of δ_1.

It may be remarked that the calculation of δ_1 does not depend on the particular difference scheme used, but only on the value of the constants $c_k^{(2)}$ and $c_k^{(3)}$.

For linear systems, considered in this paper, δ_1 can be determined once and for all, because it does not depend on k. In practice, one calculates δ_1 in the first integration step. In dealing with non-linear equations, δ_1 depends on k and is to be determined every integration step.

We now give some examples of the calculation of δ_1.

Example 5.1

Consider again the initial value problem (see section 2.1)

$$(5.4) \qquad \tilde{U} = -\ 1000\ \tilde{U} + t^2,\ \tilde{U}(0) = \tilde{U}_0,$$

where $\tilde{U}_0$ is the initial value of the solution $\tilde{U}(t)$.
Obviously, the value of δ_1 is - 1000.
Let u_k be the computed solution at $t = t_k$. Then we have from [11], formula (2.7)

$$c_k^{(2)} = -\ 1000\ c_k^{(1)} + 2\ t_k = 10^6\ [u_k - 10^{-3}t^2_k + 2{\cdot}10^{-6}t_k],$$

$$c_k^{(3)} = -\ 1000\ c_k^{(2)} + 2.$$

Hence, one would find according to formula (5.3)

$$(5.5) \qquad \delta_1 \cong -\ 1000 + \frac{2}{c_k^{(2)}} = -\ 1000 + \frac{2{\cdot}10^{-6}}{u_k - 10^{-3}t_k^2 + 2{\cdot}10^{-6}t_k}.$$

This approximation becomes incorrect for small values of $c_k^{(2)}$ or more precisely, when it happens that

$$(5.6) \qquad u_k \cong 10^{-3}(t^2_k - 210^{-3}t_k).$$

Now it is easily verified that the analytical solution of (5.4) is given by

$$\tilde{U} = e^{-1000t}\ (\tilde{U}_0 - 210^{-9}) + 10^{-3}(t^2 - 210^{-3}t + 210^{-6}).$$

Thus, when the difference scheme used is an accurate one, as it should be, the calculations finally lead to a situation in which (5.6) is satisfied, namely

$$u_k \overset{\sim}{=} \tilde{U}(t_k) \overset{\sim}{=} 10^{-3}(t_k^2 - 2{\cdot}10^{-3}t_k) + 2{\cdot}10^{-9},\ t_k >> 10^{-3}.$$

From this we draw the conclusion that δ_1 is to be determined in the first steps of the integration process.

Example 5.2

Next we consider a initial value problem for two differential equations, namely (compare problem (2.5))

$$(5.7)\qquad \begin{cases} \tilde{U} = D\tilde{U} + F,\ \tilde{U}(0) = \tilde{U}_0, \\ D = \begin{pmatrix} -500.5 & 499.5 \\ 499.5 & -500.5 \end{pmatrix},\ F = \begin{pmatrix} 2 \\ 2 \end{pmatrix},\ \tilde{U}_0 = \frac{1}{10}\begin{pmatrix} -1 \\ 1 \end{pmatrix}. \end{cases}$$

The eigenvalues of D are $\delta_r = -1$ and $\delta_1 = -1000$.
From the definition of $c_k^{(2)}$ and $c_k^{(3)}$ we deduce that

$$c_k^{(2)} = D^2u_k - F,\quad c_k^{(3)} = D^3u_k + F.$$

For k = 0 this leads to

$$c_0^{(2)} = \begin{pmatrix} -10^5 - 2 \\ 10^5 - 2 \end{pmatrix},\quad c_0^{(3)} = \begin{pmatrix} 10^8 + 2 \\ -10^8 + 2 \end{pmatrix}.$$

Applying formula (5.2') we should find

$$(5.8)\qquad \delta_1 \overset{\sim}{=} -1000\ (1 - 4{\cdot}10^{-10}),$$

while interpreting δ_1 as a diagonal matrix we should find

$$(5.8')\qquad \delta_1 \overset{\sim}{=} -1000 \begin{pmatrix} 1 - 210^{-5} & 0 \\ 0 & 1 + 210^{-5} \end{pmatrix}.$$

5.2 Determination of δ_1 in the three-cluster method

When the eigenvalues of D occur in one cluster near the origin and in two other complex clusters with large negative real parts, the local analytical solution behaves as (compare formula (2.10))

$$\tilde{U}'(t_k+\tau) \cong A_k + B_k\tau + \left[e^{\delta_1\tau+i\gamma} + e^{\bar{\delta}_1\tau-i\gamma}\right] C_k . \tag{5.9}$$

Theorem 5.2

When (5.9) is a reasonable representation of the local analytical solution, then Reδ_1 is determined by the relation

$$\text{Re}\ \delta_1 \cong \frac{|\delta_1|^2\ c_k^{(2)} + c_k^{(4)}}{2\ c_k^{(3)}} . \tag{5.10}$$

Further, if Im$\delta_1 \neq 0$

$$|\delta_1| \cong \left[\frac{[c_k^{(4)}]^2 - c_k^{(3)}\ c_k^{(5)}}{[c_k^{(3)}]^2 - c_k^{(2)}\ c_k^{(4)}}\right]^{\frac{1}{2}} \tag{5.11}$$

Proof

Proceeding in the same way as in the preceding section we find

$$c_k^{(j)} \cong C_k \left[\delta_1^{\ j}\ e^{\ i\gamma} + \bar{\delta}_1^{\ j}\ e^{-i\gamma}\right] ,\ j \geq 2. \tag{5.12}$$

There are four unknowns in these equations, namely C_k, γ, δ_1 and $\bar{\delta}_1$. Hence we need the four equations arising for j = 2, 3, 4 and 5. The equations corresponding to j = 2, 3 give

$$C_k\ e^{i\gamma} \cong \frac{c_k^{(3)} - \bar{\delta}_1 c_k^{(2)}}{\delta_1^2(\delta_1 - \bar{\delta}_1)} ,\ C_k\ e^{-i\gamma} \cong \frac{c_k^{(3)} - \delta_1\ c_k^{(2)}}{\bar{\delta}_1^2(\delta_1 - \bar{\delta}_1)} .$$

where we temporally assume that $\delta_1 \neq \bar{\delta}_1$ or Im $\delta_1 \neq 0$.

By substituting these expressions into equations j = 4 and 5 we obtain

$$(\delta_1 - \bar{\delta}_1)\left[\quad \delta_1\bar{\delta}_1\, c_k^{(2)} - (\delta_1 + \bar{\delta}_1)\, c_k^{(3)} + c_k^{(4)}\right] \overset{\sim}{=} 0$$

$$(\delta_1 - \bar{\delta}_1)\left[\ \delta_1\bar{\delta}_1(\delta_1 + \bar{\delta}_1)\, c_k^{(2)} - (\delta_1^2 + \bar{\delta}_1^2 + \delta_1\bar{\delta}_1)\, c_k^{(3)} + c_k^{(5)}\right] \overset{\sim}{=} 0.$$

From the first equation relation (5.10) immediately follows, and by substituting (5.10) into the second equation relation (5.11) is derived. Further, if Im δ_1 = 0 we have(compare the proof of theorem 5.1) that $\delta_1 \overset{\sim}{=} c_k^{(j+1)}/c_k^{(j)}$, so that (5.10)also holds in the real case. Note that in the real case (5.11) becomes undetermined.

In actual computations, an estimate of $|\delta_1|$, for instance one may use the spectral radius $\sigma(D)$, is often available, so that only relation (5.10) is necessary in order to locate the cluster center δ_1. As a consequence, the computation of $c_k^{(5)}$ is not necessary.

6. Implicit difference schemes

6.1. The generating function

Instead of employing polynomials $P_n(z)$ to generate a difference method, one may use rational functions $P_n(z)/Q_m(z)$ as generating functions. Let

$$P_n(z) = 1 + \beta_1 z + \beta_2 z^2 + \ldots + \beta_n z^n,$$

(6.1)

$$Q_m(z) = 1 + \alpha_1 z + \alpha_2 z^2 + \ldots + \alpha_m z^m.$$

Then, the function

(6.2) $$R_{n,m}(z) = \frac{P_n(z)}{Q_m(z)}$$

generates the difference scheme

$$
\begin{aligned}
&u_{k+1} + \alpha_1 \tau \, c_{k+1}^{(1)} + \ldots + \alpha_m \tau^m \, c_{k+1}^{(m)} = \\
(6.3)\quad & \\
&= u_k + \beta_1 \tau \, c_k^{(1)} + \ldots + \beta_n \tau^n \, c_k^{(n)} \, .
\end{aligned}
$$

Since the vector $c_{k+1}^{(j)}$ depends on u_{k+1} (compare [11], formula (2.5)), scheme (6.3) is implicit in u_{k+1}. In order to calculated u_{k+1}one first computes the righthand side of (6.3) to obtain a vector h_k, and then solves the equation

$$(6.4) \qquad Q_m(\tau D)u_{k+1} = h_k - \tau g_{k+1}^{(m)},$$

where $g_{k+1}^{(m)}$ is defined by (compare [11], formula (2.9'))

$$(6.5) \qquad g_{k+1}^{(m)} = \left[\alpha_1 E_0 + \alpha_2 \tau E_1 + \ldots + \alpha_m \tau^{m-1} E_{m-1}\right] f_{k+1}.$$

6.2 Consistency and Padé rational approximations

By expanding the vectors $c_{k+1}^{(j)}$ in Taylor series with respect to the point $t = t_k$ we may write (6.3) in the form

$$
\begin{aligned}
(6.6) \qquad u_{k+1} &= u_k + (\beta_1 - \alpha_1)\tau \, c_k^{(1)} + (\beta_2 - \alpha_1 - \alpha_2)\tau^2 \, c_k^{(2)} \\
&\quad + (\beta_3 - \tfrac{1}{2}\alpha_1 - \alpha_2 - \alpha_3)\tau^3 \, c_k^{(3)} + \ldots \, .
\end{aligned}
$$

We have first order accuracy when $\beta_1 - \alpha_1 = 1$, second order accuracy when, in addition, $\beta_2 - \alpha_1 - \alpha_2 = \frac{1}{2}$, and so on. There are n+m coefficients α_j and β_j, hence the maximal order of accuracy is n+m and is obtained when in (6.6) the coefficients of $c_k^{(j)}$ are equal to $\frac{1}{j!}$ for j = 1,2, ... n+m. The corresponding generating function $R_{n,m}(z)$ turns out to be a Padé rational approximation of exp(z) compare Varga [23], p.266 and Calahan [3]). In the following table several Padé approximations are listed:

Table 6.1 Padé rational approximations of exp(z)

	n = 0	n = 1	n = 2
m = 0	1	$1 + z$	$1 + z + \frac{1}{2}z^2$
m = 1	$\frac{1}{1 - z}$	$\frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$	$\frac{1 + \frac{2}{3}z + \frac{1}{6}z^2}{1 - \frac{1}{3}z}$
m = 2	$\frac{1}{1 - z + \frac{1}{2}z^2}$	$\frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}$	$\frac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2}$

Of course, other approximations, which are consistent without possessing a maximal order of accuracy, are possible. For instance, the idea of exponential fitting can also be carried through for rational approximations (see Liniger and Willoughby [17], and Liniger [16]).

6.3 Regions of stability

Most implicit difference schemes are stable for unrestricted time steps. To illustrate this we give for the case where D has negative eigenvalues a table of values of the parameter β(n,m) occurring in the stability condition

$$(6.7) \qquad \tau \leq \frac{\beta(n,m)}{\sigma(D)} \; .$$

Table 6.2 Values of $\beta(n,m)$ in the real case

	n = 0	n = 1	n = 2
m = 0	0	2	2
m = 1	∞	∞	6
m = 2	∞	∞	∞

From this table we see that not every implicit scheme is unconditionally stable. Here, the scheme corresponding with n = 2, m = 1 has a relatively severe condition.

In practice, implicit schemes are recommended in those cases where the computation time necessary to solve equation (6.4) does not nullify the advantage of taking steps which are not limited by stability considerations.

References

[1] Bodewig, E., Matrix calculus, North-Holland publishing Company, Amsterdam, (1956).

[2] Brayton, R.K., F.G. Gustavson, W. Liniger, A numerical analysis of the transient behaviour of a transistor circuit, I.B.M. Journal,(1966), p. 292 ff.

[3] Calahan, D.A., Numerical solutions of linear systems with widely seperated time constants, Proc. of the IEEE, (1967), p. 2026.

[4] Certaine, J., The solution of ordinary differential equations with large time constants, Mathematical methods for digital computers, chapter 11, (1960), p. 128-132. (Edited by A. Ralston and S. Wilt, John Wiley & Sons, Inc., New York).

[5] Curtiss, C.F., J.O. Hirschfelder, Integration of stiff equations, Proc. Math. Acad. Sci., U.S.A. 38 (1952), p. 235-243.

[6] Dahlquist, G.G., A special stability problem for linear multistep methods, BIT, 3 (1963), p. 27-43.

[7] Fowler, M.E., R.M. Warten, A numerical integration technique for ordinary differential equations with widely seperated eigenvalues, I.B.M. Journal, (1967), 537-543.

[8] Gear, C.W., The control of parameters in the automatic integration of ordinary differential equations, Report COO-1469-0077, Department of Comp. Sc. University of Illinois, Urbana, Illinois 61801, (1968).

[9] Gear, C.W., The automatic integration of stiff ordinary differential equations, Information processing 68 - North-Holland publishing company - Amsterdam, (1969), p. 187-193.

[10] Houwen, P.J. van der, Finite difference methods for solving partial differential equations, MN tract 20, Mathematisch Centrum, Amsterdam, (1968).

[11] Houwen, P.J. van der, One-step methods for linear initial value problems I. Polynomial methofs, TW report 119, Mathematisch Centrum, Amsterdam, (1970).

[12] Houwen, P.J. van der, One-step methods for linear initial value problems II. Numerical examples, TW report, Mathematisch Centrum, Amsterdam, (to appear)

[13] Lee, H.B., Matrix filterig as an aid to numerical integration, Proc. of the IEEE, vol. 55, no. 11, (1967), p. 1826-1831.

[14] Lawson, J.D., Generalized Runge-Kutta processes for stable systems with large Lipschitz constants, SIAM J. Numer. Anal., vol. 4, no. 3, (1967).

[15] Liniger, W., A criterion for A-stability of linear multi-step integration formulae, Computing 3, (1968), p. 280-285.

[16] Liniger, W., Optimization of a numerical integration method for stiff systems of ordinary differential equations, I.B.M. Research Report RC 2198, (1968).

[17] Liniger, W., R. Willoughby, Efficient integration of stiff systems of ordinary differential equations, I.B.M. Research Report RC 1970, (1967).

[18] Marcus, M., Basic theorems in matrix theory, U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series 57, (1964).

[19] Moretti, G., The chemical kinetics problem in the numerical analysis of nonequilibrium flowes, Proc. I.B.M. Sci. Comp. Symp., Large-scale problems in Physics, (1963), p. 167-182.

[20] Pope, D.A., An exponential method of numerical integration of ordinary differential equations, Communications of the ACM, Numerical Analysis, vol. 6, no. 8, (1963).

[21] Richards, P.I., W.D. Lanning, M.D. Torrey, Numerical integration of large, highly damped non-linear systems, SIAM Review, vol. 7, no. 3, (1965), p. 376 ff.

[22] Treanor, C.E., A method for the numerical integration of coupled first-order differential equations with greatly different time constants, Math. Comp., 20, (1966), p. 39-45.

[23] Varga, R.S., Matrix interative analysis, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, (1962).

[24] Zurmühl, R., Matrizen, Springer Verlag, Berlin, (1950).