LIST OF ERRATA

Page	Line	Now reads in part	Should read
I	15	attached	attacked
10	19	$\|\|\|u - \tilde{u}\|\|\| \leq$	$\|\|\|u^{*} - \tilde{u}\|\|\| \leq$
14	5	$U = R^{-1}(U_0, F, \Phi)$	$U = \tilde{R}^{-1}(U_0, F, \Phi)$
14	7	$\lim_{\tau\to 0} \lim_{\tau'\to 0} \|\|\|[U-\tilde{U}]_{d(\tau')}\|\|\| =$	$\lim_{\tau\to 0} \|\|\|[U-\tilde{U}]_{d}\|\|\| =$
		$= \lim_{\tau\to 0} \|\|\|[U-\tilde{U}]_{d}\|\|\|$	$= \lim_{\tau\to 0} \lim_{\tau'\to 0} \|\|\|[U-\tilde{U}]_{d(\tau')}\|\|\|$
31	8	$\|\|S\|\| \|\|S^{-1}\|\|$	$\|\|S\|\| \|\|S^{-1}\|\|$
38	20	$a < c < b$	$a < c \leq \frac{1}{2}(b+a-(b-a)\cos\frac{\pi}{n})$
39	1	$\delta = -\overline{\delta}$ be	$-\overline{\delta} = -\frac{1}{2}(b+a-(b-a)\cos\frac{\pi}{n})$, i.e.
39	17	$-\overline{\delta} \leq -c$	$-\overline{\delta} < -c$
39	18	$-c < -\overline{\delta}$	$-\overline{\delta} = -c$
40	10	$\text{th}(\ln \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}})$	$\text{th}(n \ln \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}})$
48	1	$Q_3(0) = 1$ and $Q_3'(0) \leq 0$	$Q_3'(0) \leq 0$ and $Q_3''(0) \leq 0$
52	28	$\gamma = \frac{2}{1+2c}$	$\gamma = \frac{2\tau}{1+2c}$
73	22	should read: decreases from ∞ to $2a$, from (4.20) and from the fact that the function $\delta_1^2 + \delta_2^2$ satisfies in the points $(\omega_1\xi, \omega_2\eta) = (\frac{\pi}{2}, 0)$, $(\frac{\pi}{2}, \pi)$, $(0, \frac{\pi}{2})$, $(\pi, \frac{\pi}{2})$ the inequality $\max_{\vec{\omega}} (\delta_1^2 + \delta_2^2) \geq \frac{\max [1, (\frac{b-2a}{b+2a})^2]}{\min (\xi^2, \eta^2)}$.	

MATHEMATICAL CENTRE TRACTS
20

FINITE DIFFERENCE METHODS

for solving partial differential equations

BY

P. J. VAN DER HOUWEN

MATHEMATISCH CENTRUM AMSTERDAM
1968

I

PREFACE

It is a well-known fact that the majority of partial differential equations cannot be integrated analytically. In these cases it is necessary to employ some method of approximation. There exist a large number of different approximation methods for solving partial differential equations the most important of which is the method of finite differences.

Finite difference methods were discussed in 1928 in the celebrated paper of Courant, Friedrichs and Lewy, but only in recent years, with the development of high-speed computing machines, these methods were applied in practical problems on a large scale. Although digital computers perform just the same operations as can be performed by hand, their speed and capacity make it possible to deal with problems which are out of the question by hand calculation.

As the speed of the computer has been increasing, one has attached still more complicated problems. However, many of these problems have turned out to be very time consuming. In such cases it is desirable to construct more efficient difference methods.

The interest in this problem and related subjects during the last years led to a seminar on the stability of difference schemes, which was organized in 1965/'66 at the Mathematisch Centrum at Amsterdam, under the supervision of Prof.dr. H.A. Lauwerier and Prof.dr.ir. A. van Wijngaarden. The present monograph presents the worked-out lectures given by the author at this seminar.

In the first chapter the basic concepts of the theory of difference schemes approximating initial boundary value problems are discussed. By presenting the material from an abstract point of view it was found possible to give a very compact description of the theory. The main result of this chapter is an equivalence theorem for convergence and stability which holds for an extensive class of difference schemes.

Chapter II is devoted to the problem of stability of two-level difference equations. Several methods are described by which it is possible to weaken the stability conditions of a given difference scheme.

In chapter III numerical solution methods for the North Sea Problem are investigated. Apart from a few drastically simplified models, this problem cannot be solved analytically. However, when finite difference methods are used, arbitrary configurations of coasts and oceans and arbitrary depth functions may be introduced. By applying the theory of the preceding chapters a number of difference schemes are constructed which appear to be acceptable with respect to their stability properties.

The last chapter deals with elliptic boundary value problems. The solution of such a problem is interpreted as the stationary solution of an appropriate initial boundary value problem, and, therefore, elliptic boundary value problems may be solved numerically by applying methods discussed in chapter II. In connection with elliptic differential equations these methods are called iterative methods. Our considerations are restricted to a special iterative method which is called Richardson's method. Some accelerating procedures are given which were successfully applied on a computer.

The author expresses his gratitude to the Board of Directors of the "Stichting Mathematisch Centrum" for giving him the opportunity to carry out the investigations presented in this monograph, and for publishing this study in the series "Mathematical Centre Tracts".

Further, it is with great pleasure that the author thanks his promotor Prof.dr. H.A. Lauwerier of the University of Amsterdam for his stimulating criticism which has been a great help in preparing this text.

He also is indebted to Mr. G.J.R. Förch for many valuable discussions.

Finally, he acknowledges Mr. C.W. de Jager and Mr. M. Murenbeeld for correcting the English text, Mrs. H. Roqué for the typing of the manuscript and Mr. D. Zwarst for the printing and the binding.

CONTENTS

Chapter I

INITIAL BOUNDARY VALUE PROBLEMS

1. Introduction

In this chapter the main features of the theory of finite difference methods are described. It will turn out that the fundamental problems encountered in this theory are those of consistency, convergence and stability.

One is faced with the problem of consistency in approximating the continuous problem by a discrete problem. It is natural to require that when refining the finite difference approximation, in the limit the discrete and continuous problems become equivalent, i.e. the finite difference approximation is required to be consistent with the continuous problem.

However, the consistency of a difference scheme does not guarantee that the difference solution approximates the analytical solution. Here, the convergence problem arises by way of the conditions for which the difference solutions converge to the analytical solution if the difference approximation is refined.

Theoretically, it suffices to construct consistent and convergent difference schemes in order to solve the analytical problem numerically. However, in actual computation one cannot find the difference solution exactly, as one is faced with the phenomenon of round-off errors which give rise to a numerical solution instead of the true difference solution. The numerical solution may differ considerably from the difference solution. Therefore, it is desirable to employ difference schemes which are more or less insensitive to such external influences. This leads to the problem of the stability of a difference scheme.

This chapter is concluded by an example which illustrates the theory of preceding sections.

It may be remarked that there already exists extensive literature on finite difference methods, for instance the treatises of Forsythe and Wasow [1960], Fox [1962], Godunov and Rjabenki [1964], Richtmyer [1957], Rjabenki and Filippov [1960] and Saul'yev [1964]. In this chapter, however, the material is presented from a more abstract point of view than was done in these works. This has the advantage of permitting a more compact description of the finite difference method.

2. Definition of initial boundary value problems

In this section we shall give an abstract formulation of a differential equation with initial and boundary conditions.

As an example we consider the diffusion equation

$$U_t - D(x,t)\, U_{xx} = H(x,t), \tag{2.1}$$

for $0 < x < 1$ and $0 < t \leq T$, with the initial condition

$$U = U_0(x), \tag{2.2}$$

for $t = 0$, and the boundary condition

$$U = \Phi(x,t), \tag{2.3}$$

for $x = 0$ and $x = 1$.

This initial boundary value problem may be interpreted as a mapping of the unknown function U onto the triple function (U_0,H,Φ), or if we wish to include the dependence of U on the difference coefficient D, we may describe equations (2.1) - (2.3) by a mapping of U onto (U_0,F,Φ), where F is a vector function with components H and D.

We shall describe a general initial boundary value problem by such a mapping.

Let us consider a real interval $[0,T]$, an Euclidean space $\mathfrak{R}_m$ of dimension m and a domain G with boundary Γ in $\mathfrak{R}_m$. Let $E(\bar{G})$, $E(G)$ and $E(\Gamma)$ be linear normed spaces of scalar or vector functions, respectively defined on the sets of points $\bar{G} \times [0,T]$, $G \times [0,T]$ and $\Gamma \times [0,T]$, where $\bar{G} = G \cup \Gamma$ and where $\bar{G} \times [0,T]$ denotes the Cartesian

product of $\bar{G}$ and $[0,T]$. The elements of $E(\bar{G})$ and $E(G)$ will be denoted by Latin capitals, the elements of $E(\Gamma)$ by Greek capitals. Further, we consider a linear normed space $E_0(\bar{G})$ of functions U_0 defined on $\bar{G}$. The linear operations in the spaces $E_0(\bar{G})$, $E(\bar{G})$, $E(G)$ and $E(\Gamma)$ will be denoted by the customary addition and multiplication operations and the norms of the elements of these spaces by $\| \ \|$, $\|\| \ \|\|$, $\|\| \ \|\|_G$ and $\|\| \ \|\|_\Gamma$ respectively.

The spaces $E_0(\bar{G})$, $E(G)$ and $E(\Gamma)$ constitute the space $E_0(\bar{G}) \times E(G) \times E(\Gamma)$ of elements (U_0,F,Φ) with $U_0 \in E_0(\bar{G})$, $F \in E(G)$ and $\Phi \in E(\Gamma)$, which is a linear normed space with respect to the linear operations

$$(U_0,F,\Phi) + (U_0',F',\Phi') = (U_0 + U_0',\ F + F',\ \Phi + \Phi') \tag{2.4}$$

and

$$a(U_0,F,\Phi) = (aU_0,aF,a\Phi), \tag{2.5}$$

where a is a scalar, and with respect to the norm

$$\|(U_0,F,\Phi)\| = \|U_0\| + \|\|F\|\|_G + \|\|\Phi\|\|_\Gamma. \tag{2.6}$$

Definition 2.1

The problem of finding the inverse of a given mapping L of an unknown function U of $E(\bar{G})$ onto a known element (U_0,F,Φ) of $E_0(\bar{G}) \times E(G) \times E(\Gamma)$ will be called an initial boundary value problem.

Initial boundary value problems will be described by the equation

$$LU = (U_0,F,\Phi). \tag{2.7}$$

The domain of definition and the range of the operator L are denoted by D_L and Δ_L respectively.

The diffusion equation considered above is an example of an initial boundary value problem in the sense of definition 2.1. The domain G is the open interval $0 < x < 1$ and Γ consists of the two boundary points $x = 0$ and $x = 1$. The spaces $E(\bar{G})$ and $E_0(\bar{G})$ are both function spaces of scalar functions depending respectively on the

variables x and t and on the variable x. Further, we may choose $E(\Gamma) = \{\phi \in E(\bar{G}) \mid \phi = 0 \text{ on } G\}$ and $E(G) = \{F \in E(\bar{G}) \mid F = 0 \text{ on } \Gamma\}$ if $F = H$ or $E(G) = \{H \in E(\bar{G}) \mid H = 0 \text{ on } \Gamma\} \times \{D \in E(\bar{G}) \mid D = 0 \text{ on } \Gamma\}$ if we also wish to include $D(x,t)$ into the data, i.e. if $F = (H,D)$.

The functions U_0, F and Φ will be called here the initial function, the interior function and the boundary function respectively. We remark that an initial value problem for ordinary differential equations may be considered as a mapping of an element U of $E(\bar{G})$ onto an element (U_0,F) of $E_0(\bar{G}) \times E(G)$. In that case the domain G consists of only one point and Γ is empty.

In this paper we shall restrict our considerations to well-posed problems (compare Hadamard [1923] and Lavrientiev [1967]).
In our notation such problems may be defined as follows.

Definition 2.2

The problem $LU = (U_0,F,\Phi)$ is said to be well-posed with respect to the norms in $E(\bar{G})$ and $E_0(\bar{G}) \times E(G) \times E(\Gamma)$ if L has a unique inverse L^{-1} which is continuous in the point (U_0,F,Φ).

3. Definition of difference schemes

In general, problem (2.7) cannot be solved in an explicit way. Therefore one associates to (2.7) a discrete problem which can be solved by elementary algebraic manipulations. We shall now define the discrete analogue of an initial boundary value problem.

First, we replace the continuous interval $[0,T]$ by the discrete set $\{t_k \mid 0 = t_0 < t_1 < \ldots < t_N = T\}$ and we define for $k = 0, 1, \ldots, N-1$

$$\tau_k = t_{k+1} - t_k, \quad \tau = \max_{0 \le k \le N-1} \tau_k. \tag{3.1}$$

Together with the set of points $\{t_k\}_{k=0}^{N}$ we take a finite set of points $G_\tau \subset G$ and a finite set of points $\Gamma_\tau \subset \Gamma$. These three point sets constitute a grid or net Q_τ in $\bar{G} \times [0,T]$, i.e.

(3.2) $$Q_\tau = \overline{G}_\tau \times \{t_k\}_{k=0}^{N},$$

where $G_\tau = G_\tau \cup \Gamma_\tau$.

Let us assume that a sequence of nets Q_τ is defined with the property that

(3.3) $$\lim_{\tau \to 0} Q_\tau \text{ is dense in } \overline{G} \times [0,T].$$

At this point we introduce linear normed spaces $E_0(\overline{G}_\tau)$, $E(\overline{G}_\tau)$, $E(G_\tau)$ and $E(\Gamma_\tau)$ for each net Q_τ, in the same way as we previously introduced the spaces $E_0(\overline{G})$, $E(\overline{G})$, $E(G)$ and $E(\Gamma)$. The elements of these spaces are defined on the sets $\overline{G}_\tau$, $\overline{G}_\tau \times \{t_k\}_{k=0}^{N}$, $G_\tau \times \{t_k\}_{k=0}^{N}$ and $\Gamma_\tau \times \{t_k\}_{k=0}^{N}$ respectively. They will be called net functions or grid functions and are denoted by small letters u_0, u, f and ϕ.

Definition 3.1

A mapping R of an unknown net function u of $E(\overline{G}_\tau)$ onto a known element (u_0,f,ϕ) of $E_0(\overline{G}_\tau) \times E(G_\tau) \times E(\Gamma_\tau)$, which is defined for each net Q_τ, will be called a difference scheme.

Difference schemes will be described by the equation

(3.4) $$Ru = (u_0,f,\phi).$$

We denote the domain of definition and the range of the operator R by D_R and Δ_R. It will be assumed that D_R and Δ_R are linear spaces and that R has a unique inverse R^{-1} which is continuous in Δ_R for every $\tau \neq 0$.

4. Consistency

So far we have not brought into relation the problems $LU = (U_0,F,\Phi)$ and $Ru = (u_0,f,\phi)$. We now investigate the conditions for which the discrete problem is an approximation of the continuous problem.

Let $[\ ]_d$ be the operator which associates to an element U of $E(\bar{G})$ the values of U in the points of the net Q_τ. These values establish the net function $[U]_d$. In the same manner we associate to the functions $U_0 \in E_0(\bar{G})$, $F \in E(G)$ and $\Phi \in E(\Gamma)$ the net functions $[U_0]_d$, $[F]_d$ and $[\Phi]_d$. The operator $[\ ]_d$ will be called the discretization operator.

The discretized elements $[U]_d$, $[U_0]_d$, $[F]_d$ and $[\Phi]_d$ are assumed to be elements of the spaces $E(\bar{G}_\tau)$, $E_0(\bar{G}_\tau)$, $E(G_\tau)$ and $E(\Gamma_\tau)$ respectively.

We shall write

$$[U]_d = u,\ [U_0]_d = u_0,\ [F]_d = f,\ [\Phi]_d = \phi. \tag{4.1}$$

Further, it will be assumed that

$$[D_L]_d = \{[U]_d \mid U \in D_L\} \subset D_R. \tag{4.2}$$

We are now in a position to compare a discrete problem and a continuous problem.

Let $\tilde{U}$ be the solution of the differential problem

$$LU = (U_0, F, \Phi), \tag{4.3}$$

and let u be the solution of the discrete problem

$$Ru = (u_0, f, \phi). \tag{4.4}$$

If equation (4.4) is a reasonable approximation of equation (4.3), one may expect that the net function $\tilde{u} = [\tilde{U}]_d$ satisfies a difference equation which closely resembles difference equation (4.4). From (4.2) we see that $\tilde{u} \in D_R$, hence there exists an element $(\tilde{u}_0, \tilde{f}, \tilde{\phi})$ such that

$$R\tilde{u} = (\tilde{u}_0, \tilde{f}, \tilde{\phi}). \tag{4.5}$$

It follows from (4.3) and (4.4) that

$$Ru = (\tilde{u}_0, \tilde{f}, \tilde{\phi}) + [L\tilde{U}]_d - R\tilde{u}. \tag{4.6}$$

Equations (4.4) and (4.5) differ by the term $[L\tilde{U}]_d - R\tilde{u}$. To evaluate this term we introduce a norm in the space $E_0(\bar{G}_\tau) \times E(G_\tau) \times E(\Gamma_\tau)$.

Let $\|\ \|$, $\|\|\ \|\|_{G_\tau}$ and $\|\|\ \|\|_{\Gamma_\tau}$ denote the norms in the spaces $E_0(\bar{G}_\tau)$, $E(G_\tau)$ and $E(\Gamma_\tau)$. We then define the norm

$$\|(u_0,f,\phi)\| = \|u_0\| + \|\|f\|\|_{G_\tau} + \|\|\phi\|\|_{\Gamma_\tau}. \tag{4.7}$$

Further, we require that for the elements of $[E_0(\bar{G})]_d$, $[E(G)]_d$ and $[E(\Gamma)]_d$ the relations

$$\|u_0\| \to \|U_0\|,\ \|\|f\|\|_{G_\tau} \to \|\|F\|\|_G,\ \|\|\phi\|\|_{\Gamma_\tau} \to \|\|\Phi\|\|_\Gamma \tag{4.8}$$

hold as $\tau \to 0$ (compare Rjabenki and Filippov [1960], p. 12).

Definition 4.1

The value of $\|[L\tilde{U}]_d - R\tilde{u}\|$ is called the error of the approximation.

Definition 4.2

A difference scheme is said to be consistent with an initial boundary value problem if the error of the approximation converges to zero as $\tau \to 0$.

In concrete cases our consistency condition reduces to the conditions generally imposed upon the difference scheme in literature (cf. Forsythe and Wasow [1960], p. 17, Rjabenki and Filippov [1960], p. 12). In practice, consistency in the sense of definition 4.2 is easily verified (see section 7 of this chapter). In connection with this we note that Lax has given a different definition of consistency (compare Lax and Richtmyer [1956] or Richtmyer [1957], p. 43). However, for the particular class of problems they consider, their consistency definition reduces to consistency in the sense of definition 4.2 (cf. Richtmyer [1957], p. 56).

This section is concluded with a figure, which may clarify the ideas described above.

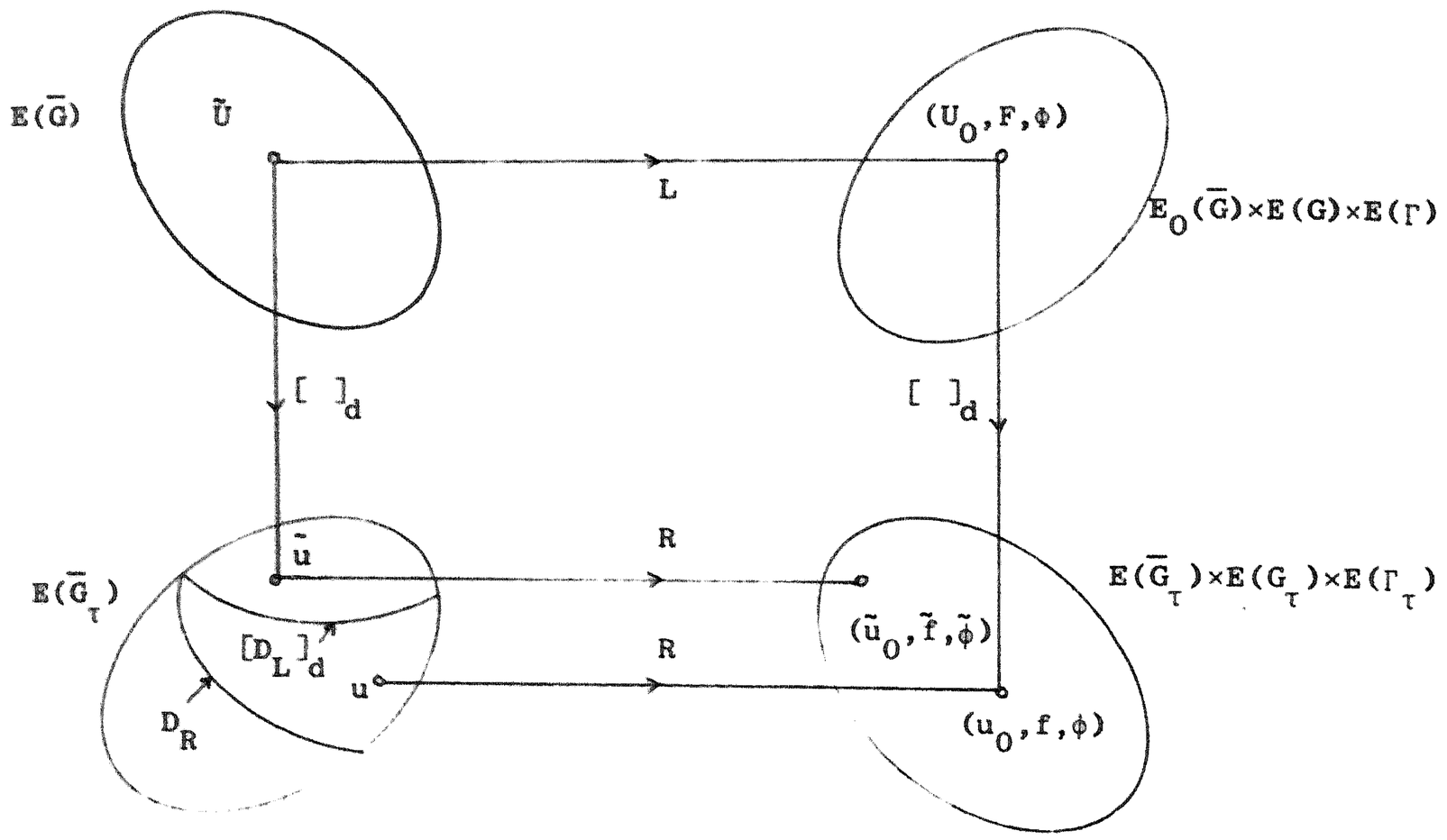

fig. 4.1

5. Convergence

The consistency of a difference scheme does not guarantee the convergence of the difference solutions to the analytical solution. The approximating difference scheme is only a formal approximation. The following definitions determine a convergent difference scheme.

Definition 5.1

The value of $\|\|u - \tilde{u}\|\|$ is called the discretization error.

Definition 5.2

A difference scheme is said to be convergent if the discretization error converges to zero as $\tau \to 0$.

Further, it will be assumed that for all elements u of $[E(\bar{G})]_d$

$$\|\|u\|\| \to \|\|U\|\| \tag{5.1}$$

as $\tau \to 0$ (compare relations (4.8)).

According to Rjabenki and Filippov [1960], p. 16 we give the following definition which is the discrete analogue of the corresponding definition 2.2 for the continuous case.

Definition 5.3

The difference scheme $Ru = (u_0, f, \phi)$ is said to be well posed with respect to the norms in $E(\bar{G}_\tau)$ and $E_0(\bar{G}_\tau) \times E(G_\tau) \times E(\Gamma_\tau)$, if for each net Q_τ, R has a unique inverse which is uniformly continuous as $\tau \to 0$ in the point (u_0, f, ϕ).

From (4.5) and (4.6) we have immediately

Theorem 5.1

A consistent and well-posed difference scheme is convergent.

If R is a linear operator a stronger statement can be made.

Theorem 5.2

Let $\|R^{-1}\| = O(\tau^{-q})$ and $\|[L\tilde{U}]_d - R\tilde{u}\| = O(\tau^p)$ as $\tau \to 0$. Then a consistent, linear difference scheme is convergent for all $(u_0, f, \phi) \in \Delta_R$ if $q < p$.

Proof

From (4.5) and (4.6) we have for linear difference schemes the relation

$$u - \tilde{u} = R^{-1}([L\tilde{U}]_d - R\tilde{u}), \tag{5.2}$$

so that

$$\|u - \tilde{u}\| \leq \|R^{-1}\| \, \|[L\tilde{U}]_d - R\tilde{u}\|, \tag{5.3}$$

where

$$\|R^{-1}\| = \sup_{(u_0, f, \phi) \in \Delta_R} \frac{\|R^{-1}(u_0, f, \phi)\|}{\|(u_0, f, \phi)\|} . \tag{5.4}$$

From the assumptions of the theorem we derive that

$$\|\|u - \tilde{u}\|\| = O(\tau^{p-q}) \tag{5.5}$$

as $\tau \to 0$. It may now be readily seen that the scheme is convergent for $p > q$.

Equation (5.5) gives a measure for the rate of convergence in the mean. For a discussion of the local rate of convergence we refer to Rjabenki and Filippov [1960], p. 20.

6. Stability

In the preceding sections we have given conditions for which a difference scheme is convergent. In actual computation, however, one cannot construct the difference solution exactly, as one is faced with the phenomenon of round-off errors which give rise to a numerical solution u^* instead of the difference solution u. In some cases the numerical solution may differ considerably from the difference solution.

Definition 6.1

The value of $\|\|u - u^*\|\|$ is called the numerical error.

In practice we would of course like the difference between the analytical and the numerical solution to be small. From the inequality

$$\|\|u - \tilde{u}\|\| \leq \|\|u^* - u\|\| + \|\|u - \tilde{u}\|\| \tag{6.1}$$

we see that both the discretization error and the numerical error must be small for the net Q_τ used. In this section we shall discuss the numerical error.

We assume that the numerical solution u^* satisfies the scheme

$$Ru^* = (u_0^*, f^*, \phi^*). \tag{6.2}$$

The numerical error may be interpreted as the result of a perturbation of the data (u_0, f, ϕ). To ensure that the numerical error is small, we

require that the difference scheme is more or less insensitive to perturbations of the data. This leads to the concept of the stability of difference schemes. We distinguish stability with respect to the initial condition (initial stability), the interior function (inner stability) and the boundary conditions (boundary stability) by restricting the perturbations to the spaces $E_0(\bar{G}_\tau) \times 0 \times 0$, $0 \times E(G_\tau) \times 0$ and $0 \times 0 \times E(\Gamma_\tau)$ respectively - compare Rjabenki and Filippov [1960], p. 15. In literature, many definitions of stability are met, each of which differs by the condition imposed upon the behaviour of the operator R^{-1} as a function of τ and T. In this monograph we shall consider stability in the sense of Forsythe and Wasow (F-W stability), Rjabenki and Filippov (R-F stability), Lax and Richtmyer (L-R stability), and O'Brien, Hyman and Kaplan (B-H-K stability).

6.1 Stability in the sense of Forsythe and Wasow

In our notation the stability definition of Forsythe and Wasow ([1960], p. 32) takes the form of

Definition 6.1

A linear difference scheme is F-W stable if $\|R^{-1}(\tau,T)\| = O(\tau^{-q})$ as $\tau \to 0$ with $q > 0$ and T constant.

From (5.3) it follows that a F-W stable scheme is convergent when the error of the approximation behaves as a certain power p of τ where $p > q$.

In practice, F-W stability implies that the numerical error behaves as a negative power of τ, since we have

$$(6.3) \qquad \|u^* - u\| \leq \|R^{-1}(\tau,T)\| \, \|(u_0^* - u_0,\ f^* - f,\ \phi^* - \phi)\| .$$

According to Forsythe and Wasow [1960], p. 32, such a behaviour is acceptable in actual computation. In addition, they noted that in practice the departure of the numerical solution from the difference solution is, in order of magnitude, either a low power of τ^{-1} or an

exponential function of τ^{-1}. For a particular class of problems, this statement was proved in Kreiss [1962], p. 163.

6.2 Stability in the sense of Rjabenki and Filippov

Rjabenki and Filippov required that the effect of a perturbation of the data upon the difference solution does not increase if the net Q_τ is refined.

Definition 6.2

A difference scheme is R-F stable if $R^{-1}(\tau,T)$ is uniformly continuous in the point (u_0,f,ϕ) as $\tau \to 0$ with T constant.

Thus an R-F stable scheme is identical to a well-posed scheme (cf. definition 5.3). From this and theorem 5.1 it follows that a consistent difference scheme which is R-F stable is also convergent.

When the difference scheme is linear, the R-F stability definition states that the operators $\{R^{-1}(\tau,T)\}_\tau$ are uniformly bounded as $\tau \to 0$ with T constant. In fact, this is the stability condition Godunov and Rjabenki ([1964], p. 45) imposed upon the difference scheme.

6.3 Stability in the sense of Lax and Richtmyer

For a particular class of linear difference schemes described by homogeneous step-by-step methods Lax and Richtmyer [1956] have given a stability definition which is related to the definition of Rjabenki and Filippov. We shall extend the definition of Lax and Richtmyer to difference schemes of the more general type $Ru = (u_0,f,\phi)$. For that purpose we assume the existence of a set of bounded linear operators $\{\tilde{R}^{-1}(\tau,T)\}_\tau$ with domain and range in $E_0(\bar{G}) \times E(G) \times E(\Gamma)$ and $E(\bar{G})$ respectively, and such that

$$[\tilde{R}^{-1}(U_0,F,\Phi)]_d = R^{-1}(u_0,f,\phi) \tag{6.4}$$

for each net Q_τ. In practice one starts with the definition of $\tilde{R}^{-1}(\tau,T)$ and then $R^{-1}(\tau,T)$ is defined by formula (6.4) - compare Richtmyer [1957], p. 41 .

Definition 6.3

A linear difference scheme is L-R stable if $\tilde{R}^{-1}(\tau,T)$ is uniformly bounded as $\tau \to 0$ with T constant.

For the particular class of problems considered by Lax and Richtmyer, Lax has proved an important theorem which is known as the equivalence theorem of Lax (see Richtmyer [1957], p. 45). This theorem states that convergence and L-R stability are equivalent. We shall prove a similar theorem for schemes of type $Ru = (u_0, f, \phi)$. The conditions for which this theorem is valid will be the same as those required by Lax, apart from the consistency condition, compare section 4.

Theorem 6.1

Let the domain of definition of the operator $\tilde{R}^{-1}(\tau, T)$ be a Banach space B. Given a linear initial boundary value problem $L\tilde{U} = (U_0, F, \Phi)$ which is well posed, and a linear difference scheme $Ru = (u_0, f, \phi)$ which is a consistent approximation of the continuous problem for all elements of B, then L-R stability is a necessary and sufficient condition for convergence of the difference scheme for all elements of B.

Proof

First we prove that a convergent scheme is necessarily L-R stable, i.e. the operators $\{\tilde{R}^{-1}(\tau, T)\}_\tau$ are uniformly bounded as $\tau \to 0$ with T constant. Using the theorem of Banach-Steinhaus it is sufficient to prove that $\tilde{R}^{-1}(\tau,T)(U_0, F, \Phi)$ is uniformly bounded as $\tau \to 0$, where (U_0, F, Φ) is an arbitrary element of the Banach space B.
Let $[\ ]_{d(\tau')}$ denote the discretization operator corresponding to τ'. Then it follows from formula (6.4) together with the convergence of the scheme and the condition that the initial boundary value problem is well posed, that

$$(6.5) \qquad \lim_{\tau \to 0} \left|\left| \tilde{R}^{-1}(\tau, T)(U_0, F, \Phi) \right|\right| = \lim_{\tau \to 0} \lim_{\tau' \to 0} \left|\left| [\tilde{R}^{-1}(\tau, T)(U_0, F, \Phi)]_{d(\tau')} \right|\right| =$$

$$= \lim_{\tau\to 0} \left|\left|\left| [\tilde{R}^{-1}(\tau,T)(U_0,F,\Phi)]_d \right|\right|\right| = \lim_{\tau\to 0} \left|\left|\left| u \right|\right|\right| = \left|\left|\left| \tilde{U} \right|\right|\right| < \infty$$

for each element of B. This proves the L-R stability of the difference scheme.

We now prove that, conversely, L-R stability implies convergence. Let $\tilde{U} = \tilde{R}^{-1}(\tilde{U}_0,\tilde{F},\tilde{\Phi})$ and $U = R^{-1}(U_0,F,\Phi)$. Then we have by the consistency and L-R stability of the scheme

$$(6.6) \qquad \lim_{\tau\to 0} \left|\left|\left| u - \tilde{u} \right|\right|\right| = \lim_{\tau\to 0} \lim_{\tau'\to 0} \left|\left|\left| [U - \tilde{U}]_{d(\tau')} \right|\right|\right| = \lim_{\tau\to 0} \left|\left|\left| [U - \tilde{U}]_d \right|\right|\right| =$$

$$= \lim_{\tau\to 0} \left|\left|\left| U - \tilde{U} \right|\right|\right| \leq \lim_{\tau\to 0} \left|\left| \tilde{R}^{-1}(\tau,T) \right|\right| \left|\left| (U_0 - \tilde{U}_0,\ F - \tilde{F},\ \Phi - \tilde{\Phi}) \right|\right| \leq$$

$$\leq \lim_{\tau\to 0} \left|\left| \tilde{R}^{-1}(\tau,T) \right|\right| \lim_{\tau\to 0} \left|\left| (u_0 - \tilde{u}_0,\ f - \tilde{f},\ \phi - \tilde{\phi}) \right|\right| = 0.$$

From this theorem it follows that R-F stability implies L-R stability, so that one should require L-R stability instead of R-F stability in order to guarantee convergence. However, in all examples known to the author in which L-R stability is proved, one has actually proved R-F stability (compare for instance the examples given by Richtmyer [1957]). Therefore, it should be of interest to know when L-R stability and R-F stability are equivalent. For instance, it is possible to prove the following theorem.

Theorem 6.2

Let the conditions of theorem 6.1 be satisfied and let for each $\tau > 0$ A be a subspace of B such that $[A]_d = \{(u_0,f,\phi) \mid \|(u_0,f,\phi)\| = 1\}$. Further, let $\left|\left|\left| [U]_d \right|\right|\right| \leq c \left|\left|\left| U \right|\right|\right|$ where c does not depend on τ and U. Then L-R stability and R-F stability are equivalent when the subspaces A are uniformly bounded in τ.

Proof

We have

$$\left|\left| R^{-1}(\tau,T) \right|\right| = \sup_{(u_0,f,\phi)\in[A]_d} \left|\left|\left| R^{-1}(u_0,f,\phi) \right|\right|\right| =$$

$$= \sup_{(U_0,F,\Phi)\in A} \left|\left|\left| [\tilde{R}^{-1}(U_0,F,\Phi)]_d \right|\right|\right| \leq$$

$$\leq c \sup_{(U_0,F,\Phi)\in A} \left|\left|\left| \tilde{R}^{-1}(U_0,F,\Phi) \right|\right|\right| \leq$$

$$\leq c \left\| \tilde{R}^{-1}(\tau,T) \right\| \sup_A \| (U_0,F,\Phi) \| .$$

From this we see that L-R stability implies R-F stability.
The converse follows from theorem 6.1.

As an example we consider the case where $E_0(\bar{G})$, $E(G)$ and $E(\Gamma)$ are the spaces of all functions which are continuous on $\bar{G}$, $\bar{G} \times [0,T]$, and on $\Gamma \times [0,T]$ respectively. These spaces are Banach spaces with respect to the norms

$$(6.7) \quad \|U_0\| = \max_{\bar{G}} |U_0|, \quad \left|\left|\left| F \right|\right|\right|_G = \max_{\bar{G}\times[0,T]} |F|, \quad \left|\left|\left| \Phi \right|\right|\right|_\Gamma = \max_{\Gamma\times[0,T]} |\Phi| .$$

It can easily be verified that $B = E_0(G) \times E(G) \times E(\Gamma)$ is a Banach space with respect to the norm defined by (2.6) and (6.7).
In the same manner we define maximum norms in the spaces $E_0(\bar{G}_\tau)$, $E(G_\tau)$, $E(\Gamma_\tau)$, $E(\bar{G}_\tau)$ and $E(\bar{G})$.
For this example the conditions of theorem 6.2 can easily be satisfied.

6.4 Stability in the sense of O'Brien, Hyman and Kaplan

If the net Q_τ is refined and T is kept constant, the F-W, R-F and L-R stability conditions guarantee a certain insensitivity to perturbations of the data. However, nothing is said about the calculation of the difference solution u on a fixed net Q_τ. Let us consider step-by-step methods, i.e. processes in which u is constructed by successively calculating its level functions u_k which are formed by the values of u in the points $\bar{G}_\tau \times t_k$. For large values of N it is important to know the behaviour of the perturbations as a function of t_k in order to derive conditions which prevent an accumulation of round-off errors at the end of the step-by-step method. To describe the develop-

ment of the round-off errors we consider the behaviour of $\|R^{-1}(\tau,T)\|$ as a function of T, where τ is kept constant (cf. O'Brien, Hyman and Kaplan [1951]).

Definition 6.4

A difference scheme is B-H-K stable if $R^{-1}(\tau,T)$ is uniformly continuous in the point (u_0,f,ϕ) as $T \to \infty$ with τ constant.

This stability condition prevents an accumulation of the perturbations at the end of a step-by-step method and is, therefore, very important from the practical point of view.

In their paper, O'Brien, Hyman and Kaplan distinguished weak and strong stability. These forms of stability may be interpreted as initial stability and inner or boundary stability.

Finally, we note that the concept of linear instability introduced by O'Brien, Hyman and Kaplan, may be expressed by the formula

$$\|R^{-1}(\tau,T)\| = O(T^{+q}) \text{ as } T \to \infty \tag{6.8}$$

with $q > 0$ and τ constant. This concept may be compared with the F-W stability where $\|R^{-1}(\tau,T)\|$ behaves as a negative power of τ as $\tau \to 0$.

7. An example

In this section we shall illustrate the theory of the preceding sections by a simple example. It will be shown how one may construct difference schemes and improve the accuracy in certain cases. Further, a difference scheme will be given which yields the exact analytical solution of the differential problem, but which is an unstable scheme in the sense of O'Brien, Hyman and Kaplan. Finally, it will be shown that a stability condition which guarantees stability with respect to one norm, may lead to instability with respect to another norm.

Let us consider the differential equation

$$\tilde{U}_t + \tilde{U} = F \tag{7.1}$$

for $0 < t \leq T$ with the initial condition $\tilde{U} = U_0$ for $t = 0$. In this example the domain G consists of only one point, Γ is empty, $E(G)$ is a space of scalar functions depending on t, and $E_0(\bar{G})$ is the real axis or, if desired, the complex plane. Evidently, the problem is of the type $L\tilde{U} = (U_0, F)$.

The net Q_τ will be defined by points $t_k = k\tau$, where $k = 0,1,\ldots,N$ and $\tau = T/N$, and the difference scheme $Ru = (u_0, f)$ is defined by the equations

$$(7.2)\qquad \begin{cases} u_0 = [U_0]_d, \\ u_1 = [\tilde{U}_1]_d + O(\tau^p) \text{ as } \tau \to 0, \\ \alpha u_{k+1} + \beta u_k + \gamma u_{k-1} = f_k, \quad k = 1, 2, \ldots, N-1. \end{cases}$$

In these equations we have $u_k = u(t_k)$, $\tilde{U}_1 = \tilde{U}(t_1)$ and $p > 0$; α, β and γ are parameters which have to be determined in such a way that the scheme is a consistent approximation of the initial value problem. We remark that the value of u_1 may be found within any order of accuracy p by using a Taylor expansion of $\tilde{U}$ in the point $t = 0$ and by determining the coefficients from equation (7.1).

7.1 Consistency

Let us assume that D_L consists of functions which are differentiable a sufficient number of times. Then, by substituting $\tilde{U}$ into (7.2) and expanding $\tilde{U}$ in the point $t = t_k$, we find for $k = 1, 2, \ldots, N-1$ the equation

$$(7.3)\quad \left[\alpha\left(1 + \tau\frac{\partial}{\partial t} + \frac{1}{2}\tau^2\frac{\partial^2}{\partial t^2} + \ldots\right) + \beta + \gamma\left(1 - \tau\frac{\partial}{\partial t} + \frac{1}{2}\tau^2\frac{\partial^2}{\partial t^2} - \ldots\right)\right]\tilde{u}_k = \tilde{f}_k,$$

where

$$\frac{\partial^n}{\partial t^n}\tilde{u}_k = \left(\left[\frac{\partial^2}{\partial t^n}\tilde{U}\right]_d\right)_{t=t_k}.$$

From this we obtain

$$\tilde{f}_k - f_k = (\alpha + \beta + \gamma - 1)\tilde{u}_k + (\alpha - \gamma - \frac{1}{\tau})\tau \frac{\partial}{\partial t} \tilde{u}_k + \tag{7.4}$$

$$+ \frac{1}{2} (\alpha + \gamma)\tau^2 \frac{\partial^2}{\partial t^2} \tilde{u}_k + \frac{1}{3!} (\alpha - \gamma)\tau^3 \frac{\partial^3}{\partial t^3} \tilde{u}_k + \ldots .$$

The error of the approximation is given by

$$\| (u_0, \tilde{f}) - (u_0, f) \| \leq \| \tilde{f} - f \|, \tag{7.5}$$

so that $\tilde{f}_k - f_k$ must converge to zero as $\tau \to 0$ for all k. Since the difference scheme does not depend on the values of f_k for $k = 0$ and $k = N$, we may define $\tilde{f}_0 = f_0$ and $\tilde{f}_N = f_N$. The values of $\tilde{f}_k - f_k$ for $k = 1, 2, \ldots, N-1$ depend on the class C of functions $\tilde{U}$ under consideration.

Let N_L be the space of functions which are solutions of the homogeneous differential equation, i.e.

$$N_L = \{\tilde{U} | L\tilde{U} = (U_0, 0)\}.$$

We consider the following two cases:

(a) N_L is contained in C. Let us choose

$$\alpha + \beta + \gamma - 1 = 0, \quad \alpha - \gamma - \frac{1}{\tau} = 0. \tag{7.6}$$

Substitution of (7.6) into (7.4) yields as $\tau \to 0$ the expression

$$\tilde{f}_k - f_k = \frac{1}{2} (2\alpha\tau - 1)\tau \frac{\partial^2}{\partial t^2} \tilde{u}_k + O(\tau^2). \tag{7.7a}$$

From this we conclude that the choices $2\alpha\tau - 1 = O(1)$ as $\tau \to 0$ and $p \geq 1$ together with (7.6) lead to a first order approximation, and the choices $2\alpha\tau - 1 = O(\tau)$ and $p \geq 2$ together with (7.6) lead to a second order approximation.

(b) C is contained in N_L. Here we use the fact that $\tilde{U}_t = -\tilde{U}$. From (7.4) we obtain

(7.7b) $$\tilde{f}_k - f_k = \tilde{f}_k = [(\alpha + \beta + \gamma - 1) - (\alpha - \gamma - \frac{1}{\tau})\tau +$$

$$+ \frac{1}{2!}(\alpha + \gamma)\tau^2 - \frac{1}{3!}(\alpha - \gamma)\tau^3 + \ldots]\tilde{u}_k.$$

In this simple case it is possible to choose α, β and γ in such a way that the error of the approximation is zero for all τ, i.e.

(7.8) $$\alpha = 1, \quad \beta = -\exp(-\tau), \quad \gamma = 0.$$

The difference scheme becomes

(7.9) $$u_{k+1} = \exp(-\tau)u_k,$$

which leads to the analytical solution $\tilde{u}$. However, in the homogeneous case the coefficients α, β and γ as defined by (7.8), give rise to only a first order approximation.

It is possible to construct a difference scheme which solves the homogeneous case exactly and which approximates the inhomogeneous case with second order accuracy. Substitute (7.6) into (7.7b), then

$$\tilde{f}_k = [\frac{2\alpha\tau - 1}{2!}\tau - \frac{1}{3!}\tau^2 + \frac{2\alpha\tau - 1}{4!}\tau^3 - \frac{1}{5!}\tau^4 + \ldots]\tilde{u}_k.$$

It is easily verified that $\tilde{f}_k = 0$ for

(7.10) $$2\alpha\tau - 1 = \frac{\frac{\tau^2}{3!} + \frac{\tau^4}{5!} + \ldots}{\frac{\tau}{2!} + \frac{\tau^3}{4!} + \ldots} = \frac{\sinh\tau - \tau}{\cosh\tau - 1}.$$

On the other hand, we have

(7.11) $$2\alpha\tau - 1 \sim \frac{1}{3}\tau \text{ as } \tau \to 0,$$

so that the inhomogeneous case is approximated with second order accuracy.

7.2 Stability

We shall study the stability of the scheme which arises from (7.6), i.e. the scheme

$$(7.12)\qquad (\rho + 1)u_{k+1} + 2(\tau - \rho)u_k + (\rho - 1)u_{k-1} = 2\tau f_k,$$

where

$$\rho = 2\alpha\tau - 1.$$

It is convenient to write this scheme in the equivalent form

$$(7.13)\qquad \begin{cases} u_{k+1} = 2\,\dfrac{\rho-\tau}{\rho+1}\,u_k + \dfrac{1-\rho}{1+\rho}\,b\,v_k + \dfrac{2\tau}{\rho+1}\,f_k, \\ v_{k+1} = \dfrac{1}{b}\,u_k, \end{cases}$$

where $k = 1, 2, \ldots, N-1$, $\rho \neq -1$ and where b is a parameter $\neq 0$. Introducing the vector $\vec{w}_k$ with components u_k and v_k, the vector $\vec{g}_k$ with components f_k and 0, and the matrix A, where

$$(7.14)\qquad A = \begin{pmatrix} 2\,\dfrac{\rho-\tau}{\rho+1} & b\,\dfrac{1-\rho}{1+\rho} \\ \dfrac{1}{b} & 0 \end{pmatrix},$$

and setting

$$(7.15)\qquad I = \frac{2\tau}{\rho+1}\,E,$$

where E is the identity matrix, we may write the difference scheme as

$$(7.13')\qquad \vec{w}_{k+1} = A\,\vec{w}_k + I\,\vec{g}_k.$$

This scheme is of the type $\vec{Rw} = (\vec{w}_1, \vec{g})$.

Next we define in the space $E(\bar{G}_\tau)$ of difference solutions $\vec{w}$ a norm in terms of the norm in $E_0(\bar{G}_\tau)$:

$$(7.16)\qquad |||\vec{w}||| = \operatorname*{Max}_k \|\vec{w}_k\|.$$

It can be proved that (see chapter II, section 3.1 and section 4)

$$(7.17)\qquad |||\vec{w}||| \leq \operatorname{Max}(1, \|A\|^N)\|\vec{w}_1\| + \|I\|\,\frac{1 - \|A\|^N}{1 - \|A\|}\,|||\vec{g}|||.$$

We now discuss the cases of initial and inner stability.

(a) Initial stability. It is clear from inequality (7.17) that the condition

$$\|A\| < 1 + O(\tau) \text{ as } \tau \to 0$$

guarantees R-F stability with respect to $\vec{w}_1$, and that

$$\|A\| \leq 1$$

guarantees B-H-K stability in the weak sense.

The norm of A depends on the norm we choose in $E_0(\bar{G}_\tau)$. Let us consider the norms

$$\|\vec{w}_k\|_p = \sqrt[p]{|u_k|^p + |v_k|^p}, \quad p = 2, \infty. \tag{7.18}$$

When p = 2 the norm of A is the spectral norm. By choosing $b = \sqrt{(1+\rho)/(1-\rho)}$ and $|\rho| < 1$, the matrix A becomes symmetric, so that $\|A\|_2$ is equal to the spectral radius $\sigma(A)$ of A. In this case we derive that

$$\begin{cases} \|A\|_2 \leq 1 + O(\tau) & \text{for} \quad 0 \leq \rho < \frac{1}{2}\tau, \\ \|A\|_2 = 1 & \text{for} \quad \rho = \frac{1}{2}\tau, \\ \|A\|_2 < 1 & \text{for} \quad \frac{1}{2}\tau < \rho < 1. \end{cases} \tag{7.19}$$

For $\rho < 0$ we have instability. The case $\rho \geq 1$ may be investigated by using the relation

$$\|A\|_2 = \sqrt{\sigma(AA^*)},$$

A^* being the conjugate transpose of A, which holds for any matrix A. However, in section 3.2 of chapter II we shall give a less laborious method of analysis.

When $p = \infty$ the norm of A is the maximum norm. Choosing b = 1 we find that

$$\begin{cases} \|A\|_\infty = 1 + 2\frac{\tau - 2\rho}{1+\rho} & \text{for} \quad 0 \leq \rho < \frac{1}{2}\tau, \\ \|A\|_\infty = 1 & \text{for} \quad \frac{1}{2}\tau \leq \rho \leq 1. \end{cases} \tag{7.20}$$

From (7.19) and (7.20) it follows that, with respect to the spectral norm as well as the maximum norm, there is R-F stability for $0 \leq \rho < 1$ and B-H-K stability for $\frac{1}{2}\tau < \rho < 1$. This implies that the scheme arising from (7.10) is not B-H-K stable, although it yields the analytical solution in the homogeneous case.

(b) Inner stability. From (7.17) we derive that the scheme is R-F stable when

$$\|A\| \leq 1 + O(\tau) \text{ as } \tau \to 0$$

and B-H-K stable in the strong sense when

$$\|A\| < 1.$$

Thus R-F stability with respect to the initial condition implies R-F stability with respect to the inhomogeneous term, irrespective whether the spectral or the maximum norm is used for A. However, the results (7.19) and (7.20) indicate that for $\frac{1}{2}\tau < \rho < 1$ the scheme is only B-H-K stable when the spectral norm is used. In fact, for $\frac{1}{2}\tau \leq \rho \leq 1$ the scheme is linearly unstable with respect to the maximum norm in the sense of O'Brien, Hyman and Kaplan. This shows that stability depends on the choice of norms in the function spaces $E_0(\bar{G}_\tau)$ and $E(\bar{G}_\tau)$.

Chapter II

TWO-LEVEL DIFFERENCE EQUATIONS

1. Introduction

In the preceding chapter we have discussed three important concepts, viz. consistency, convergence and stability, which provide conditions for the successful application of the method of finite differences to initial boundary value problems.
In this chapter we shall consider these conditions more closely for a particular class of difference schemes which are called step-by-step methods and which are described by two-level difference equations. In most cases the consistency condition is easily satisfied and we shall concentrate our attention to the R-F stability in order to guarantee convergence and to the B-H-K stability for preventing accumulation of round-off errors.

We begin with analysis of stability. For stationary step-by-step methods a simple algebraic criterium is derived which guarantees B-H-K stability. The case of R-F stability is more difficult and requires further attention. For general stationary schemes a necessary criterium for R-F stability is given, which is known as von Neumann's condition. By restricting the class of difference schemes in an appropriate way we are able to reduce the problem of R-F stability to an algebraic problem in matrix theory. This results in a number of criteria which are sufficient for R-F stability.

In many cases the R-F stability conditions which have to be imposed upon the net Q_τ in order to guarantee convergence are rather unattractive in actual computation. For instance, if we are not content with the accuracy we have obtained, we may decide to refine the net $\overline{G}_\tau$. Then it is often necessary to refine the set of points $\{t_k\}_{k=0}^{N}$ to a much greater extent than accuracy would require. For that reason it is desirable to

soften the stability conditions of the difference scheme. The main part of this chapter will be concerned with appropriate transformations of the difference scheme in order to make the stability conditions less stringent. We consider

(a) the method of non-uniform time steps,

(b) the method of implicit difference schemes,

(c) the method of dissipative terms.

(a) The use of non-uniform time steps will be investigated for schemes in which a certain difference operator has either real or imaginary eigenvalues. The case of real eigenvalues reduces essentially to a well-known method which uses Chebyshev polynomial operators and which was developed for the iterative solution of elliptic boundary value problems. See Flanders and Shortley [1950], Young [1953] or Forsythe and Wasow [1960], p. 227. This method was used by Yuan'Chzhao-Din [1958] to construct R-F stable schemes for the solution of self-adjoint parabolic initial boundary value problems (compare Saul'yev [1964] and Franklin [1959]). However, his method is not B-H-K stable in the strong sense and, as the method has the property that the perturbations of the difference solution, due to the round-off errors associated with the application of the Chebyshev polynomial operators, are not at random, this method may lead to a large accumulation of round-off errors. We shall slightly modify the method to guarantee B-H-K stability in the strong sense.

In the case of imaginary eigenvalues it turns out that the time steps have to be chosen complex. This implies that one has to use twice as much storage room as was needed for the original scheme. However, the stability conditions are considerably softened. The method can be applied to transport problems, for instance the North Sea Problem discussed in chapter III, and yields seemingly new difference schemes.

(b) It will be shown that a very general class of difference schemes can be transformed into implicit difference schemes which are unconditionally B-H-K stable and which unconditionally satisfy von Neumann's criterium. However, this practical advantage has to be paid for by solving a matrix equation at each time step. Fortunately, such matrix

equations are not ill conditioned so that fast converging iteration methods may be applied (see chapter IV).

(c) The effect of methods (a) and (b) may be interpreted as the introduction of viscosity terms of increasing order into the original scheme. This suggests the consideration of other types of dissipative terms. In this chapter the effect of an inertia term will be studied.

2. Two-level formulae

Let $E(G_\tau)$ and $E(\Gamma_\tau)$ be subspaces of $E(\bar{G}_\tau)$ consisting of functions which are zero on $\Gamma_\tau \times \{t_k\}_0^N$ and $G_\tau \times \{t_k\}_0^N$ respectively, and let $E_0(G_\tau)$ and $E_0(\Gamma_\tau)$ be subspaces of $E_0(\bar{G}_\tau)$ consisting of functions which are zero on Γ_τ and G_τ. Further, let the functions u_k, f_k and ϕ_k for $k = 0, 1, \ldots, N$ be net functions lying in $E_0(\bar{G}_\tau)$ and defined by the values of u, f and ϕ in the points of the net $\bar{G}_\tau \times t_k$. Clearly we have $f_k \in E_0(G_\tau)$ and $\phi_k \in E_0(\Gamma_\tau)$ for $k = 0, 1, \ldots, N$. The functions u_k, f_k and ϕ_k are called level functions.

We now define a difference scheme by the two-level formula

$$u_{k+1} = A_k u_k + I_k f_k + B_k \phi_k, \tag{2.1}$$

where A_k, I_k and B_k are linear operators uniformly bounded as $\tau \to 0$ with domain in $E_0(\bar{G}_\tau)$, $E_0(G_\tau)$ and $E_0(\Gamma_\tau)$ respectively and range in $E_0(\bar{G}_\tau)$. It is clear that (2.1) describes a difference scheme in the sense of definition 2.1 of chapter I.

As soon as the data u_0, f and ϕ are given one may step-by-step construct the difference solution u from the level functions u_k, f_k and ϕ_k. Most difference approximations of linear initial boundary value problems can be reduced to schemes of type (2.1).

In order to study the stability properties of (2.1) we write

$$u_{k+1} = P_{0k} u_0 + Q_k f + S_k \phi, \tag{2.2}$$

where P_{0k}, Q_k and S_k are operators defined by

$$
(2.3)\qquad
\begin{cases}
P_{lk} = \begin{cases} \prod_{m=k}^{l} A_m & \text{for } 0 \le l \le k \\ 1 & \text{for } l > k, \end{cases} \\
Q_k f = \sum_{l=1}^{k+1} P_{lk}\, I_{l-1}\, f_{l-1}, \\
S_k \phi = \sum_{l=1}^{k+1} P_{lk}\, B_{l-1}\, \phi_{l-1}.
\end{cases}
$$

Since the difference scheme is linear, the stability depends upon the behaviour of $\|R^{-1}\|$ as a function of τ and T. We shall relate $\|R^{-1}\|$ to the norms of P_{0k}, Q_k and S_k. For that purpose we define

$$
(2.4)\qquad |||u||| = \operatorname{Max}_k \|u_k\| .
$$

Throughout this monograph it will be assumed that the norm in $E(\overline{G}_\tau)$ is expressed in this way. Further, we define the quantities

$$
(2.5)\qquad |||P_0||| = \operatorname{Max}_k \|P_{0k}\|, \quad |||Q||| = \operatorname{Max}_k \|Q_k\|, \quad |||S||| = \operatorname{Max}_k \|S_k\|,
$$

where $\|P_{0k}\|$, $\|Q_k\|$ and $\|S_k\|$ denote the norms of P_{0k}, Q_k and S_k. From chapter I, formula (4.7), and formulae (2.2), (2.4) and (2.5) we find

$$
(2.6)\qquad \|R^{-1}\| \le \operatorname{Max}(|||P_0|||, |||Q|||, |||S|||).
$$

The quantities $|||P_0|||$, $|||Q|||$ and $|||S|||$ determine the initial, the inner and the boundary stability respectively. We shall investigate these quantities seperately in the following sections.

This section is concluded with some examples of difference schemes of type (2.1).

Example 2.1

Consider the scheme defined in section 7 of the preceding chapter. This scheme was reduced (compare formula (7.13')) to equation (2.1)

without boundary conditions.

Example 2.2

Many linear partial differential equations in time t and space coordinate $\vec{x}$ can be reduced to the form

$$\tilde{U}_t = \tilde{D}\tilde{U} + F, \tag{2.7}$$

where $\tilde{D}$ is a differential operator in $\vec{x}$ defined in a domain G. Let U_0 be an initial function defined on $\bar{G}$, let $F = 0$ for $\vec{x} \in \Gamma$ and let Φ be a boundary function such that $\Phi = 0$ for $\vec{x} \in G$ and

$$\tilde{U} = \Phi$$

for $\vec{x} \in \Gamma$.

Further, let we choose the points $k\tau$ on the t-axis and a net $G_\tau \cup \Gamma_\tau$ in $\bar{G}$. Then, by replacing the differential quotients in (2.7) by difference quotients defined by the values of $\tilde{U}$ in the netpoints of Q_τ, we may obtain the consistent difference scheme

$$u_{k+1} = (1 + \tau D)u_k + I\, f_k + E_+\, \phi_k, \tag{2.8}$$

where D is a difference analogue of $\tilde{D}$ in G_τ, $1 + \tau D = 0$ in Γ_τ, $I = \tau$ and $E_+\phi_k = \phi_{k+1}$.

Difference schemes of type (2.8) are very important and the greater part of our consideration will deal with such schemes.

3. Stability with respect to the initial condition

In this section stability will mean stability with respect to the initial condition. This form of stability is extensively investigated in literature and many authors restrict their stability considerations completely to stability with respect to the initial condition. We mention Esch [1960], Kreiss [1962], Richtmyer [1957], Saul'yev [1964] and Todd [1956].

We shall merely give an outline of the theory of stability analysis and we shall restrict our considerations to the reduction of the stability problem to a problem in matrix theory.

3.1 Non-stationary operators A_k

Let us define

$$|||A||| = \operatorname{Max}_k \|A_k\| . \tag{3.1}$$

Theorem 3.1

For difference scheme (2.1) we have the following stability criteria.

(a) $|||A||| \leq 1 + O(\tau \ln \tau)$ as $\tau \to 0$ ==> F-W stability.

(b) $|||A||| \leq 1 + O(\tau)$ as $\tau \to 0$ ==> R-F stability.

(c) $|||A||| \leq 1$ as $\tau \to 0$ ==> B-H-K stability (weak stability).

Proof

The theorem follows immediately from the inequality

$$|||P_0||| \leq \operatorname{Max}_k |||A|||^k . \tag{3.2}$$

Example 3.1

We shall investigate the equation

$$\tilde{U}_t - \tilde{U}_{xx} + a\tilde{U} = F \tag{3.3}$$

for $0 < x < 1$ and $0 < t \leq T$ with the initial condition $\tilde{U} = U_0$ for $t = 0$, and the boundary conditions $\tilde{U}_x + b\tilde{U} = \Phi$ in $x = 0$ and $\tilde{U}_x = \Phi$ in $x = 1$.
The parameters a and b are real.
We choose a net Q_τ in which the net points are given by $(x_j, t_k) = (j\xi, k\tau)$ where $\xi = 1/m$, $\tau = T/N$, $j = 0, 1, \ldots, m$ and $k = 0, 1, \ldots, N$.
The value of ξ is expressed in terms of τ by the relation

$$\xi = \sqrt{\frac{\tau}{r}} , \tag{3.4}$$

where r is a constant.

We may construct the following consistent difference approximation.

$$
(3.5)\quad \begin{cases} u_{k+1} = \sqrt{r}\,(\sqrt{r} - b\sqrt{\tau})^{-1}\,[ru_k + (1 - 2r - a\tau)X_+u_k + rX_+^2u_k] + \\ \qquad + \tau\sqrt{r}\,(\sqrt{r} - b\sqrt{\tau})^{-1}X_+f_k - \sqrt{\tau}(\sqrt{r} - b\sqrt{\tau})^{-1}E_+\phi_k \text{ in } j=0, \\ u_{k+1} = rX_-u_k + (1 - 2r - a\tau)u_k + rX_+u_k + \tau f_k \text{ in } j=1,2,\ldots,m-1, \\ u_{k+1} = rX_-^2u_k + (1 - 2r - a\tau)X_-u_k + ru_k + \tau X_-f_k + \sqrt{\tau/r}\,E_+\phi_k \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in } j=m, \end{cases}
$$

where X_+ is defined by $X_+u_k(j\xi) = u_k((j+1)\xi)$.
Choosing the maximum norm in $E_0(\bar{G}_\tau)$ we obtain for the matrix A the bound

$$
(3.6)\qquad \|A\|_\infty = \mathrm{Max}\{1, \sqrt{r}\,|\sqrt{r} - b\sqrt{\tau}|^{-1}\}(2r + |1 - 2r - a\tau|).
$$

Applying theorem 3.1 we may derive that for $b \leq 0$ the scheme is R-F stable when

$$
(3.7)\qquad r \leq \frac{1}{2}
$$

and B-H-K stable when

$$
(3.8)\qquad r \leq \frac{1}{2} - \frac{1}{2}a\tau,\ a \geq 0.
$$

There is no R-F stability for $b > 0$, but there is B-H-K stability in the cases

$$
(3.9)\qquad r \leq \frac{1}{2} - \frac{1}{2}a\tau,\ a \geq 0,\ \tau > \frac{4r}{b^2},
$$

$$
(3.10)\qquad r \leq \frac{1}{2} - \frac{1}{2}a\tau,\ a > 0,\ \frac{b^2}{a^2r} \leq \tau \leq \frac{4r}{b^2}.
$$

3.2 Stationary operators A

When A does not depend on k, the stability criteria given in theorem 3.1 may be weakened in a number of cases.

Throughout this section it will be assumed that in the finite dimensional space $E_0(\bar{G}_\tau)$ a norm is defined according to an inner-product in $E_0(\bar{G}_\tau)$. Then we may represent the bounded linear operator A with respect to an arbitrary orthonormal base in $E_0(\bar{G}_\tau)$ by a matrix which will also be denoted by A.

From (2.3) and (2.5) we have

$$(3.11) \qquad \|| P_0 ||| = \max_{0 \leq k \leq N-1} \| A^k \|,$$

where $N = T/\tau$. If $\| A^N \|$ is uniformly bounded as $\tau \to 0$ or $T \to \infty$, then $\|| P_0 |||$ is uniformly bounded as $\tau \to 0$ or $T \to \infty$. Note that the criteria given in theorem 3.1 were derived from the requirement that, as $\tau \to 0$ or $T \to \infty$, $\| A_k \|^N$ is uniformly bounded for all k.

When the matrix A is normal, i.e. when $AA^* = A^*A$ where A^* is the conjugate transpose of A, then this condition on $\| A^N \|$ reduces to the criteria of theorem 3.1, since

$$(3.12) \qquad \| A^N \| = \sigma^N(A) = \| A \|^N,$$

where $\sigma(A)$ is the spectral radius of A.

In general, however, we have

$$(3.13) \qquad \| A^N \| \leq \| A \|^N.$$

Therefore, we may find weaker stability criteria than indicated by theorem 3.1 in the stationary case.

Lemma 3.1

Let $N = T/\tau$ and let p be the largest order of all diagonal submatrices J_r of the Jordan normal form J of A with $\sigma(J_r) = \sigma(A)$. Then, as $\tau \to 0$ or $T \to \infty$

$$(3.14) \qquad \| A^N \| \sim \nu \; N^{p-1} \; [\sigma(A)]^{N-p+1},$$

where ν depends on τ, but does not depend on T.

Proof

According to Varga [1962], p. 64, we have for large values of N

$$\|A^N\| \sim \nu' \binom{N}{p-1} \sigma(A)^{N-p+1}, \tag{3.15}$$

where ν' is a positive constant which only depends on the matrix A. In fact, ν' satisfies the inequality

$$\frac{1}{\|S\| \|S^{-1}\|} \leq \nu' \leq \|S\| \|S^{-1}\|, \tag{3.16}$$

where S is a nonsingular matrix related to A by the equation $J = SAS^{-1}$. $\|S\| \|S^{-1}\|$ is called the condition number of the matrix A (cf. Varga [1962], p. 65). Thus ν' is bounded by a quantity which does not depend on T.

The binomial coefficient $\binom{N}{p-1}$ may be transformed by Stirling's formula. For large values of N we have

$$\binom{N}{p-1} \sim \nu'' N^{p-1}, \tag{3.17}$$

where ν'' is a constant not depending on N.
Substituting (3.17) into (3.15) and setting $\nu'\nu'' = \nu$ we obtain (3.14).

Theorem 3.2

Let $\mu(A)$ be the maximal multiplicity of the eigenvalues α_j of A with $|\alpha_j| = \sigma(A)$. Then scheme (2.1) is B-H-K stable if $\sigma(A) \leq 1$ and $\mu(A) = 1$, or if $\sigma(A) < 1$.

Proof

Applying lemma 3.1 with $\sigma(A) \leq 1$ and $\mu(A) = 1$, i.e. $p \leq \mu(A) = 1$, we see that $\|A^{T/\tau}\|$ is uniformly bounded for $\tau \to 0$. This is also the case if $\sigma(A) < 1$ and $\mu(A)$ arbitrary.

Note that we have linear instability if $\sigma(A) \leq 1$ and $p > 1$, and exponential instability if $\sigma(A) > 1$. Thus $\sigma(A) \leq 1$ is a necessary condition for stability.

This theorem will be needed in the stability analysis of a difference scheme for the North Sea Problem, which is discussed in chapter III.

When T is kept constant and τ tends to zero, formula (3.14) may be used to investigate F-W and R-F stability. From (3.14) and (3.16) we deduce for F-W and R-F stability the necessary criteria $\sigma(A) \leq 1 + O(\tau \ln \tau)$ and $\sigma(A) \leq 1 + O(\tau)$ respectively, as $\tau \to 0$. In order to obtain sufficient criteria we have to know something about the behaviour of ν as a function of τ. From formula (3.16) we see that this behaviour is determined by the matrix S.

Theorem 3.3

Let the condition number of the matrix A be uniformly bounded as $\tau \to 0$. Then scheme (2.1) is F-W stable if $\sigma(A) \leq 1 + O(\tau \ln \tau)$ as $\tau \to 0$ and R-F stable if $\sigma(A) \leq 1 + O(\tau)$ as $\tau \to 0$ and $\mu(A) = 1$.

Proof

The theorem follows immediately from lemma 3.1, formula (3.16) and the assumptions of the theorem.

In general, the analysis of the condition number of A as a function of τ is difficult. However, when R-F stability is investigated, the following lemma may simplify the analysis.

Lemma 3.2

Let $A = B(\tau) + \tau C(\tau)$ where $B^N(\tau)$ and $C(\tau)$ are uniformly bounded as $\tau \to 0$. Then A^N is uniformly bounded as $\tau \to 0$.

A proof of this lemma may be found in Strang [1964].

As an example we consider difference schemes in which the operator A is of the form

$$A(\tau) = A(0) + \tau C(\tau), \tag{3.18}$$

where $A(0)$ and $C(\tau)$ are matrices of constant order and $C(\tau)$ is uniformly bounded as $\tau \to 0$. Such difference schemes arise from ordinary differ-

ential equations. The condition number of A(0) does not depend on τ. Therefore, according to lemma 3.2 and theorem 3.3, the scheme is R-F stable provided that $\sigma(A(0)) \leq 1$ and $\mu(A(0)) = 1$.

In the case of difference schemes for partial differential equations the order of A is not constant, but tends to infinity as $\tau \to 0$. However, in some important cases it is possible to reduce the stability problem for the matrix A to an equivalent problem for a matrix $\hat{A}$ of constant order.

Let us consider the elements e_0 of $E_0(\bar{G}_\tau)$ of the form

$$e_0 = v_0 \, s(\vec{\omega},\vec{x}_j),$$

where v_0 is a (vector) function of $E_0(\bar{G}_\tau)$ with values for the, say n, components $v_0^{(1)}, \ldots, v_0^{(n)}$ which do not depend on the net point $\vec{x}_j$ of the net $\bar{G}_\tau$, and where $s(\vec{\omega},\vec{x}_j)$ is a scalar function defined on $\bar{G}_\tau$, which depends on a vector parameter $\vec{\omega}$ lying in the $\vec{x}$-plane. Further, let A be an operator with the property that

$$A \, v_0 \, s(\vec{\omega},\vec{x}_j) = s(\vec{\omega},\vec{x}_j) \, \hat{A}(\vec{\omega}) \, v_0, \tag{3.19}$$

where $\hat{A}(\vec{\omega})$ is a n × n matrix which depends on $\vec{\omega}$. $\hat{A}(\vec{\omega})$ is called the amplification matrix. For the scalar functions $s(\vec{\omega}_1,\vec{x}_j)$ and $s(\vec{\omega}_2,\vec{x}_j)$ we define an inner-product $(s(\vec{\omega}_1,\vec{x}_j),s(\vec{\omega}_2,\vec{x}_j))$ by summing the values of $s(\vec{\omega}_1,\vec{x}_j)\,\overline{s(\vec{\omega}_2,\vec{x}_j)}$ over all net points (the bar denotes the conjugate value), and for functions of $E_0(\bar{G}_\tau)$ we define an inner-product by summing the the inner-products between the corresponding components from 1 to n.

Theorem 3.4

Let A satisfy (3.19), where the functions $\{s(\vec{\omega},\vec{x}_j)\}_{\vec{\omega}}$ form a complete, orthonormal set for the components of the functions u_0 of $E_0(\bar{G}_\tau)$. Then the following inequality holds:

$$\|A^k\| \leq \sup_{\vec{\omega}} \hat{A}^k(\vec{\omega}) \, .$$

Proof

We may represent an element u_0 of $E_0(G_\tau)$ in the form

$$u_0 = \sum_{\vec{\omega}} v_0(\vec{\omega})\ s(\vec{\omega},\vec{x}_j).$$

From the definitions of the inner-products and the fact that the functions $\{s(\vec{\omega},\vec{x}_j)\}_{\vec{\omega}}$ are orthonormal, it follows that

$$\begin{aligned}
\|A^k u_0\|^2 &= (\sum_{\vec{\omega}} \hat{A}^k(\vec{\omega}) v_0(\vec{\omega}) s(\vec{\omega},\vec{x}_j), \sum_{\vec{\omega}} \hat{A}^k(\vec{\omega}) v_0(\vec{\omega}) s(\vec{\omega},\vec{x}_j)) \\
&= \sum_{\vec{\omega}} \|\hat{A}^k(\vec{\omega}) v_0(\vec{\omega}) s(\vec{\omega},\vec{x}_j)\|^2 \\
&\leq \operatorname{Sup}_{\vec{\omega}} \|\hat{A}^k(\vec{\omega})\|^2 \sum_{\vec{\omega}} \|v_0(\vec{\omega}) s(\vec{\omega},\vec{x}_j)\|^2 \\
&= \operatorname{Sup}_{\vec{\omega}} \|\hat{A}^k(\vec{\omega})\|^2 \sum_{\vec{\omega}} \sum_i \|v_0^{(i)}(\vec{\omega})\|^2 \\
&= \operatorname{Sup}_{\vec{\omega}} \|\hat{A}^k(\vec{\omega})\|^2 \|u_0\|^2.
\end{aligned}$$

Hence

$$\|A^k\| = \operatorname*{Sup}_{\|u_0\|=1} \|A^k u_0\| \leq \operatorname{Sup}_{\vec{\omega}} \|\hat{A}^k(\vec{\omega})\|.$$

By means of this theorem we may reduce the R-F and F-W stability problem to the analysis of the matrices $\hat{A}^k(\vec{\omega})$. Applying formula (3.14) to $\hat{A}(\vec{\omega})$ we still have a factor ν which depends on τ, but the matrices $\hat{S}$ and $\hat{S}^{-1}$, occurring in the upper bound of ν, are now of fixed order n. This may simplify the problem considerably. For instance, by applying theorem 3.3 to the amplification matrix $\hat{A}(\vec{\omega})$ we see that, when $\hat{A}(\vec{\omega})$ has eigenvectors which are uniformly independent as $\tau \to 0$ for all $\vec{\omega}$ (i.e. the determinant of the matrix having the normalized eigenvectors of $\hat{A}(\vec{\omega})$ as columns, is bounded away from zero uniformly in τ and $\vec{\omega}$), then ν is uniformly bounded as $\tau \to 0$ for all $\vec{\omega}$. We obtain the R-F stability condition $\sigma(\hat{A}(\vec{\omega})) \leq 1 + O(\tau)$ as $\tau \to 0$ for all $\vec{\omega}$.
This stability condition was given by Richtmyer ([1957], p. 64) for difference schemes satisfying equation (3.19) for the set

(3.20) $$s(\vec{\omega},\vec{x}_j) = \exp(i\vec{\omega}\cdot\vec{x}_j).$$

Here $\vec{\omega}\cdot\vec{x}_j$ denotes the inner-product of $\vec{\omega}$ and $\vec{x}_j$. The other stability criteria given by Richtmyer ([1957], chapter IV) also hold in the general case considered above. Each of these criteria lead to the condition

(3.21) $$\sigma(A) \leq 1 + O(\tau) \text{ as } \tau \to 0.$$

Thus, they merely indicate that the necessary condition (3.21) (von Neumann's condition) is also sufficient for R-F stability.
In the following subsections we restrict our considerations to sufficient conditions for B-H-K stability and to von Neumann's necessary condition. If R-F stability is considered, we shall neglect the terms $O(\tau)$ in A.

Example 3.2

Consider the scheme defined in chapter I by formulae (7.13'), (7.14) and (7.15). In this example the matrix A was given by

$$A = \begin{pmatrix} 2\,\frac{\rho-\tau}{\rho+1} & b\,\frac{1-\rho}{1+\rho} \\ \frac{1}{b} & 0 \end{pmatrix},$$

where $b \neq 0$.
The eigenvalues of A are given by

$$\alpha = \frac{\rho - \tau + \sqrt{1 - 2\rho\tau + \tau^2}}{\rho + 1}.$$

From this we may deduce that there is B-H-K stability for $\rho \geq \frac{1}{2}\tau$ (theorem 3.2). Applying theorem 3.3 we find R-F stability for $\rho \geq 0$. Recalling that a first analysis given in chapter I, section 7 led to the conditions $\frac{1}{2}\tau \leq \rho \leq 1$ and $0 \leq \rho < 1$ respectively, it may be concluded that the considerations in this section have led to weaker stability conditions.

3.3 The method of non-uniform real time steps

In this section we shall consider difference schemes of the special type (2.8), i.e. the matrix A is of the form

$$A = 1 + \tau D. \tag{3.21}$$

It will be assumed that D does not depend on k and that the eigenvalues δ_j, j = 1, ..., m, of D satisfy the inequality

$$-\sigma(D) = \delta_m \leq \delta_{m-1} \leq \ldots \leq \delta_1 = -\delta_0 < 0. \tag{3.22}$$

This type of difference schemes arises from initial boundary value problems for parabolic differential equations (compare example 3.3). Applying theorem 3.2 we obtain the result that the condition

$$\tau < \frac{2}{\sigma(D)} \tag{3.23}$$

is sufficient for B-H-K stability. If (3.23) holds as $\tau \to 0$, then von Neumann's condition, necessary for R-F stability, is also satisfied. For relatively large values of $\sigma(D)$, condition (3.23) will lead to inconveniently small time steps τ. In such cases it is desirable to improve the stability. We shall discuss a method by which the stability condition is considerably weakened.

From the theory of the iterative solution of elliptic boundary value problems we take the following method (see for instance Forsythe and Wasow [1960], p. 226). The operator A is replaced by a polynomial of degree n in τD, i.e.

$$P_n(\tau D) = 1 + \tau D + \beta_2 \tau^2 D^2 + \ldots + \beta_n \tau^n D^n, \tag{3.24}$$

where $\beta_2, \ldots \beta_n$ are real parameters which are uniformly bounded as $\tau \to 0$. Clearly, the scheme

$$u_{k+1} = P_n(\tau D) u_k \tag{3.25}$$

approximates the same continuous problem as the scheme $u_{k+1} = (1 + \tau D)u_k$. The eigenvalues of the operator $P_n(\tau D)$ are given by $P_n(\tau \delta_j)$, j = 1, 2, ..., m (see figure 3.1).

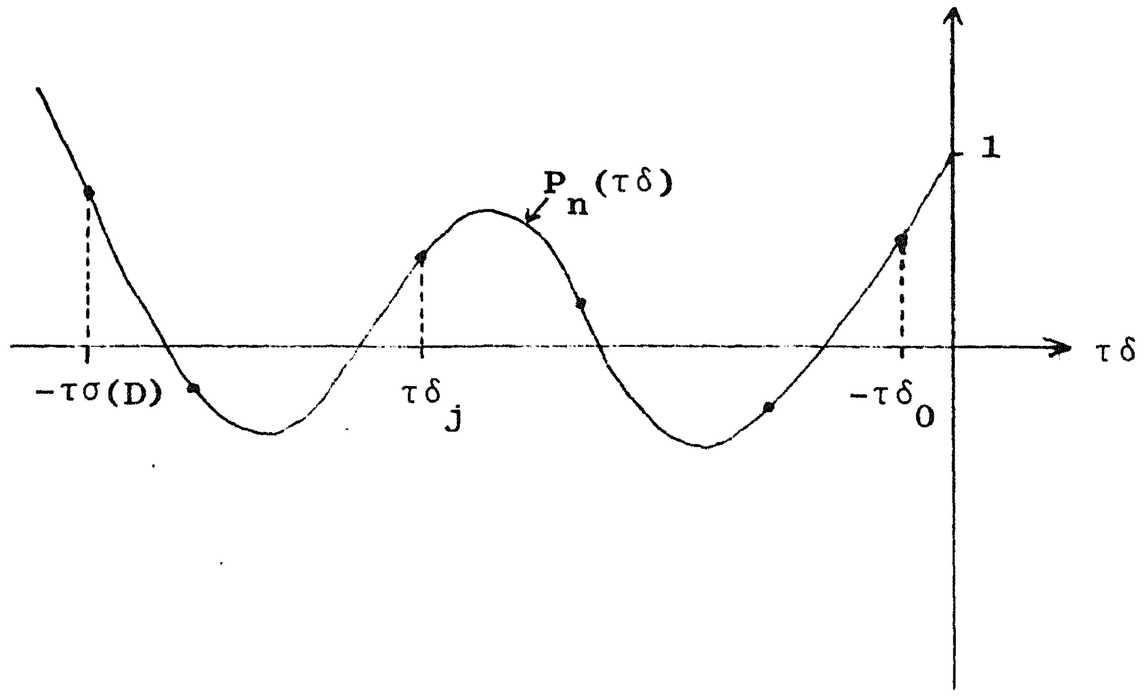

fig. 3.1

We require that $\sigma(P_n(\tau D)) < 1$ where τ is as large as possible. Observing that

$$\tau = [\frac{d}{d\delta} P_n(\tau\delta)]_{\delta=0},$$

we are led to the following problem.

Given are positive numbers c and b with $c < b$, and a positive number α_0 less than 1. The problem is to find a polynomial $Q_n(\delta)$ of degree n in δ such that

(3.26) $$Q_n(0) = 1,$$

(3.27) $$|Q_n(\delta)| \leq \alpha_0 \text{ for } -b \leq \delta \leq -c,$$

(3.28) $$Q_n'(0) \quad \text{as large as possible.}$$

Let $Q_n(\delta)$ be such a polynomial, then we set

(3.29) $$c = \delta_0,\ b = \sigma(D),\ P_n(\tau\delta) \equiv Q_n(\delta),$$

by which we obtain for $P_n(\tau D)$ the spectral radius

$$\sigma(P_n(\tau D)) = \alpha_0 < 1. \tag{3.30}$$

Scheme (3.25) satisfies the stability condition

$$\tau \leq Q_n'(0), \tag{3.31}$$

where $Q_n'(0)$ will appear to depend on the value of α_0.

In order to solve problem (3.26) - (3.28) we consider a related problem occurring in the theory of iterative processes, namely to find a polynomial $Q_n(\delta)$ satisfying (3.26) which has a minimal maximum norm over a given negative interval. The solution of this problem was given by Markov (see Forsythe and Wasow [1960], p. 227).

Theorem 3.5

Given the positive numbers a and b (a < b). The polynomial

$$C_n(a,b,\delta) = T_n^{-1}\left(\frac{b+a}{b-a}\right) T_n\left(\frac{b+a+2\delta}{b-a}\right),$$

where T_n is the Chebyshev polynomial cos(n arccos w), has of all polynomials $Q_n(\delta)$ of degree n in δ satisfying $Q_n(0) = 1$, a minimal maximum norm over the interval $-b \leq \delta \leq -a$.

We now prove that such a transformed Chebyshev polynomial is also the solution of problem (3.26) - (3.28).

Theorem 3.6

Given the positive numbers a, b and c, satisfying a < c < b, and the number $\alpha_0 = T_n^{-1}\left(\frac{b+a}{b-a}\right)$. Then the polynomial $C_n(a,b,\delta)$ has of all polynomials $Q_n(\delta)$ of degree n in δ, satisfying $Q_n(0) = 1$ and $|Q_n(\delta)| \leq \alpha_0$ for $-b \leq \delta \leq -c$, the largest derivative in $\delta = 0$.

Proof

The extrema of $C_n(a,b,\delta)$, including the boundary extrema in the points $\delta = -a$ and $\delta = -b$ are alternatively $+\alpha_0$ and $-\alpha_0$ (see figure 3.2).

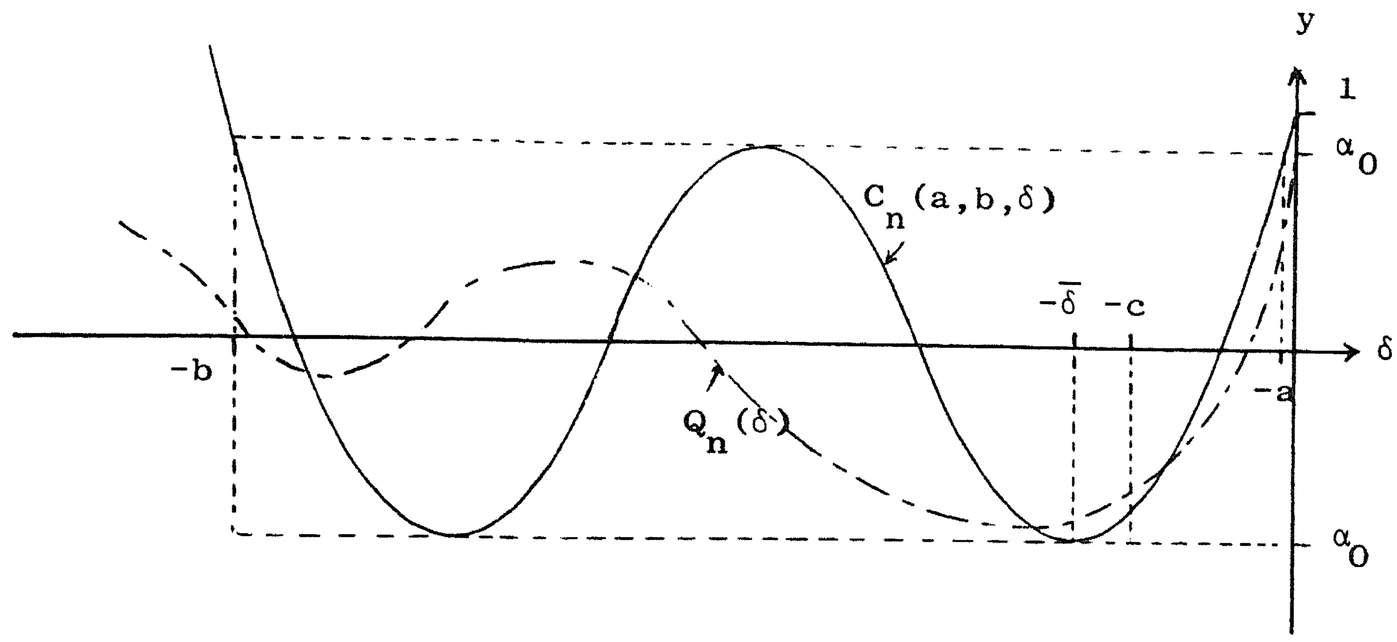

fig. 3.2

Let $\delta = -\bar{\delta}$ be the point in which the curve $y = C_n(a,b,\delta)$ assumes its first non-boundary extremum negatively from $\delta = -a$ moving along the δ-axis, and suppose that $-\bar{\delta} < -c < -a$. Then, the curve $y = C_n(a,b,\delta)$ divides the rectangular domain R bounded by the lines $y = \pm \alpha_0$, $\delta = -\bar{\delta}$ and $\delta = -b$ into n disjunct domains (see figure 3.2).
Let $Q_n(\delta)$ be a polynomial satisfying the conditions of the theorem and the inequality $Q_n'(\delta) > C_n'(a,b,\delta)$ in $\delta = 0$. Now $Q_n(\delta)$ intersects the curve $y = C_n(a,b,\delta)$ at least twice outside the domain R and intersects $y = C_n(a,b,\delta)$ n-1 times in R unless one or more points of intersection coincide with the non-boundary extrema of $C_n(a,b,\delta)$. Since $Q_n(\delta)$ cannot leave $\bar{R}$ for $-b < \delta < -\bar{\delta}$, we have in the latter case second order points of intersection. This implies that the polynomial $V(\delta) = C_n(a,b,\delta) - Q_n(\delta)$ has n+1 zeroes. This contradiction excludes the possibility that $Q_n'(\delta) > C_n'(a,b,\delta)$ in $\delta = 0$.
If the derivatives in $\delta = 0$ are equal and $Q_n(\delta) \not\equiv C_n(a,b,\delta)$, we have a second order point of intersection in $\delta = 0$, which leads to the same contradiction. This proves the theorem for $-\bar{\delta} \leq -c$.
The case $-c < -\bar{\delta}$ is proved analogously.

Theorem 3.7

Let a and b satisfy the inequalities $0 \leq a \leq \delta_0$ and $b \geq \sigma(D)$. Then the following approximations hold for the scheme $u_{k+1} = C_n(a,b,D)u_k$.

$$\tau \sim \frac{2n^2}{b} \frac{\operatorname{th}(2n\sqrt{\frac{a}{b}})}{2n\sqrt{\frac{a}{b}}}, \quad a << b, \tag{3.32}$$

$$\sigma(C_n(a,b,D)) \sim \cosh^{-1}(2n\sqrt{\frac{a}{b}}), \quad a << b. \tag{3.33}$$

Further, the scheme is B-H-K stable and satisfies von Neumann's condition if $a > 0$, or equivalently if

$$\tau < \frac{2n^2}{b} . \tag{3.34}$$

Proof

Applying theorem 3.6 we see that the polynomial $Q_n(\delta) \equiv C_n(a,b,\delta)$ solves problem (3.26) - (3.28). From the definition of $C_n(a,b,\delta)$ we have

$$\alpha_0 = \cosh^{-1}(n \ln \frac{\sqrt{b} + \sqrt{a}}{\sqrt{b} - \sqrt{a}}), \quad C_n'(a,b,0) = \frac{2n^2}{b} \frac{\operatorname{th}(\ln \frac{\sqrt{b} + \sqrt{a}}{\sqrt{b} - \sqrt{a}})}{2n\sqrt{\frac{a}{b}}} .$$

We now use the approximation

$$\ln \frac{\sqrt{b} + \sqrt{a}}{\sqrt{b} - \sqrt{a}} = 2\sqrt{\frac{a}{b}} + O(\frac{a}{b}), \quad a << b.$$

Substituting this into the expressions for α_0 and $C_n'(a,b,0)$ and using formulae (3.30) and (3.31) we are led to the approximations (3.33) and (3.32) respectively.

The second part of the theorem follows immediately from (3.32), (3.33) and theorem 3.2.

It may be remarked that (3.32) implies that τb is uniformly bounded as $\tau \to 0$, so that, as $\tau \to 0$, $P_n(\tau\delta)$ is a uniformly bounded polynomial on a uniformly bounded interval $-\tau b \leq \tau\delta \leq -\tau\delta_0 < 0$. Therefore, the coefficients of $P_n(\tau\delta)$ are uniformly bounded which proves the consistency of the scheme.

In order to compare the time steps of the original and the new scheme we introduce the effective time step

$$\tau_{eff} = \frac{\tau}{n} = \frac{2n}{b} \frac{\operatorname{th}(2n\sqrt{\frac{a}{b}})}{2n\sqrt{\frac{a}{b}}} \tag{3.35}$$

(for each time step the computational labour of the new scheme is roughly n times as great as the labour of the original scheme). Choosing $b = \sigma(D)$ we see that we have gained a factor $n \operatorname{th}(2n\sqrt{\frac{a}{b}})/(2n\sqrt{\frac{a}{b}})$ in computation time over the original scheme (compare condition (3.23)). In figure 3.3 the behaviour of $\operatorname{th}(2n\sqrt{\frac{a}{b}})/(2n\sqrt{\frac{a}{b}})$ and $\cosh^{-1}(2n\sqrt{\frac{a}{b}})$ as functions of $2n\sqrt{\frac{a}{b}}$ is illustrated.

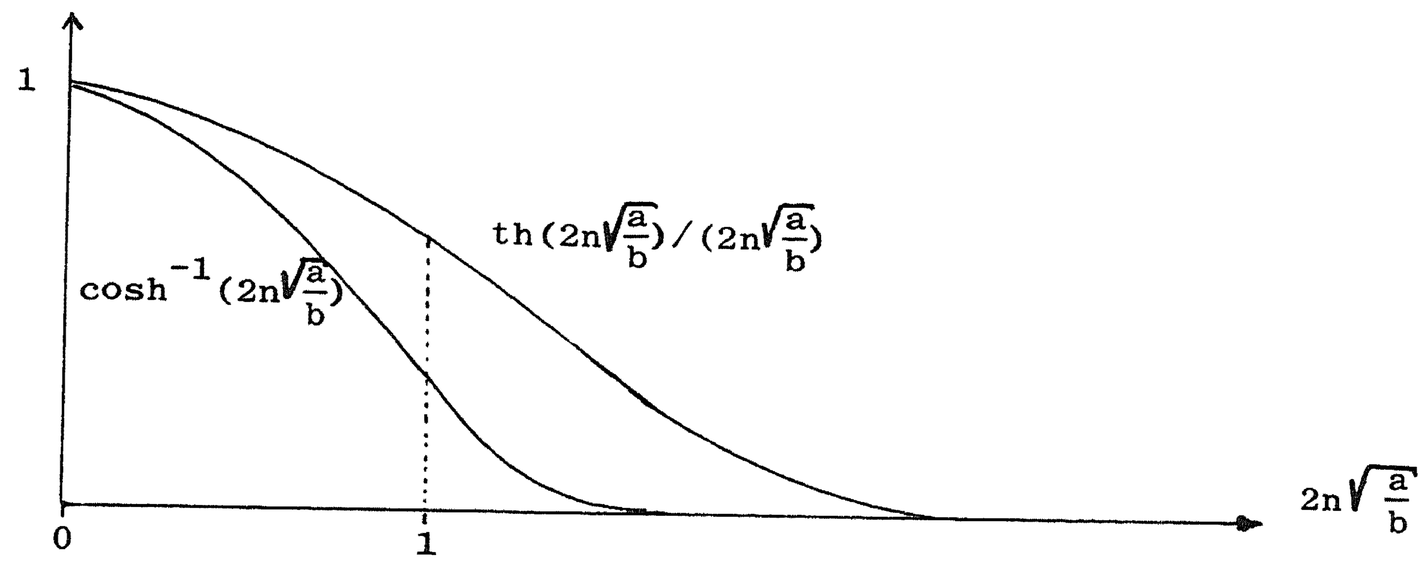

fig. 3.3

For $a = 0$ the gain factor assumes its maximal value n. The corresponding difference scheme $u_{k+1} = C_n(0, \sigma(D), D)u_k$ was employed by Yuan'Chzhao-Din (see Saul'yev [1964], p. 317) to solve initial boundary value problems for parabolic differential equations. Such schemes are R-F stable as well as B-H-K stable (in the weak sense) provided that the problem is self-adjoint. Otherwise, we are not sure that the scheme is B-H-K stable in the weak sense, but it is certainly unstable in the strong sense (see section 4). Therefore, we prefer to choose $a > 0$, which guarantees B-H-K stability in the weak sense as well as in the strong sense (see formula (3.33) or figure 3.3). Because $\operatorname{th}(2n\sqrt{\frac{a}{b}})/(2n\sqrt{\frac{a}{b}})$ varies slow-

ly for small values of $2n\sqrt{\frac{a}{b}}$, the effective time step τ_{eff} will not change much by choosing a small positive value for a.

In actual application of the method we shall use the fact that $C_n(a,b,\delta)$ has real zeroes, so that $C_n(a,b,\delta)$ can be factorized into real linear factors of the form $1 - \delta/z_l$, where z_l is a zero of $C_n(a,b,\delta)$. From the definition of $C_n(a,b,\delta)$ we derive that

$$z_l = -\frac{1}{2}(a+b) - \frac{1}{2}(a-b)\cos\frac{2l+1}{2n}\pi, \tag{3.36}$$

where $l = 0, 1, \ldots, n-1$. The difference scheme assumes the form

$$u_{k+1} = C_n(a,b,D)u_k = \prod_{r=0}^{n-1}(1 + \omega_r D)u_k, \tag{3.24'}$$

where the numbers ω_r (the so-called relaxation parameters) have the values $-z_l^{-1}$, $l = 0, 1, \ldots, n-1$. The scheme (3.24) is uniquely determined when we are given the correspondence $r = r(l)$ for $l = 0, 1, \ldots, n-1$. With regard to the storage room in the computing machine, the factorized form is very suitable.

Finally, we remark that scheme (3.24') may be interpreted as the original scheme with non-uniform real time steps ω_r.

Example 3.3

We investigate for the two-dimensional diffusion equation

$$\tilde{U}_t = \Delta\tilde{U} + F$$

the Dirichlet problem for the square of side π.
On a grid with rectangular meshes of sides ξ and η we define at the internal net points the difference operator D by the formula

$$D = \frac{Y_+ + \alpha + Y_-}{\alpha + 2}\,\frac{X_+ - 2 + X_-}{\xi^2} + \frac{X_+ + \beta + X_-}{\beta + 2}\,\frac{Y_+ - 2 + Y_-}{\eta^2}, \tag{3.37}$$

where $X_\pm$ and $Y_\pm$ represent translations $\pm\xi$ and $\pm\eta$ along the x-axis and the y-axis respectively (compare example 3.1). The parameters α and β are weight parameters. It is obvious from the structure of D

that (3.37) represents a difference analogue of the Laplace operator Δ. If $\alpha = \beta = \infty$, then (3.37) reduces to the usual five-point Laplace difference operator (cf. Forsythe and Wasow [1960], p. 192). However, it will turn out that more appropriate choices for α and β can be made.

We write the operator D in the form

$$D = L_1(X_+ + X_-)(Y_+ + Y_-) + L_2(X_+ + X_-) + L_3(Y_+ + Y_-) + L_4, \tag{3.38}$$

where the coefficients L_i are defined by

$$\begin{cases} L_1 = \frac{1}{2}\,\xi^{-2}(1 + \rho^2 - \gamma), \\ L_2 = \xi^{-2}(\gamma - \rho^2), \\ L_3 = \xi^{-2}(\gamma - 1), \\ L_4 = -\,2\,\xi^{-2}\gamma, \\ \gamma = \frac{\alpha}{\alpha + 2} + \rho^2\,\frac{\beta}{\beta + 2},\ \rho = \xi/\eta. \end{cases} \tag{3.39}$$

A consistent difference scheme for the initial boundary value problem under consideration is of the form (see example 2.1)

$$u_{k+1} = (1 + \tau D)u_k + \tau f_k + E_+\phi_k,$$

In order to apply the method of non-uniform real time steps the eigenvalues δ of D have to be negative. The eigenfunctions of D are given by

$$e_{n,m} = \sin nj\xi \sin ml\eta, \tag{3.40}$$

where $n = 1, 2, \ldots, \frac{\pi}{\xi} - 1$ and $m = 1, 2, \ldots, \frac{\pi}{\eta} - 1$, and the eigenvalues by

$$\delta_{n,m} = L_4 + 2L_2 \cos n\xi + 2L_3 \cos m\eta + 4L_1 \cos n\xi \cos m\eta. \tag{3.41}$$

We are interested in the extrema of $\delta_{n,m}$. Since $\delta_{n,m}$ is a harmonic function of the variables $\cos n\xi$ and $\cos m\eta$ in the region

$$-\cos\xi \leq \cos n\xi \leq \cos\xi,\ -\cos\eta \leq \cos m\eta \leq \cos\eta, \tag{3.42}$$

$\delta_{n,m}$ assumes its extrema at the boundary of this region, and because $\delta_{n,m}$ is a linear function of cos $n\xi$ and cos $m\eta$ along these boundaries, the extrema are assumed in the corner points A(cos ξ, cos η), B(- cos ξ, cos η), C(- cos ξ, - cos η) and D(+ cos ξ, - cos η). From (3.39) and (3.40) we find that

$$(3.43)\quad \begin{cases} \delta_A = -2 + [\dfrac{1+\rho^2-\gamma}{2\rho^2}\, pq\xi^2] \\ \delta_B = -4\xi^{-2} + [p + \dfrac{2+\rho^2-2\gamma}{\rho^2}\, q] \\ \delta_C = -4\rho^2\xi^{-2} + [p(1+2\rho^2-2\gamma) + q] \\ \delta_D = -4(2\gamma-1-\rho^2)\xi^{-2} - [(2+3\rho^2-3\gamma)p + \dfrac{3+2\rho^2-3\gamma}{\rho^2}\, q \\ \qquad\qquad - \dfrac{1+\rho^2-\gamma}{2\rho^2}\, pq\xi^2], \end{cases}$$

where

$$p = 2(1-\cos\xi)\xi^{-2},\quad q = 2(1-\cos\eta)\eta^{-2}.$$

For small values of ξ and bounded values of ρ and γ we have

$$(3.43')\qquad \delta_A \sim -2,\ \delta_B \sim -4\xi^{-2},\ \delta_C \sim -4\rho^2\xi^{-2} \text{ and } \delta_D \sim -4(2\gamma-1-\rho^2)\xi^{-2}$$

as first approximations. To guarantee that the eigenvalues $\delta_{n,m}$ are negative we require that

$$(3.44a)\qquad \gamma \geq \frac{1}{2}(1+\rho^2).$$

Assuming that $\xi \geq \eta$, i.e. $\rho \geq 1$, we obtain

$$(3.45)\qquad \delta_0 \sim 2,\ \sigma(D) \sim 4\xi^{-2}\ \mathrm{Max}(\rho^2, 2\gamma-1-\rho^2).$$

Formulae (3.23) and (3.32) indicate that it is desirable to choose $\sigma(D)$ as small as possible. Therefore we shall require that

(3.44b) $$\gamma \leq \frac{1}{2}\,(1 + \rho^2) + \frac{1}{2}\,\rho^2,$$

which results in the approximation $\sigma(D) \sim 4\rho^2\xi^{-2} = 4\eta^{-2}$. Substitution of this value for b into (3.35) yields for small values of $n\eta$

(3.46) $$\tau_{eff} \sim \frac{1}{2}\, n\eta^2[1 - \frac{1}{3}\, an^2\eta^2 + O(a^2n^4\eta^4)],$$

where $0 \leq a \leq 2$.

It may be remarked that the usual five-point difference formula ($\alpha = \beta = \infty$ or $\gamma = 1 + \rho^2$) leads to $\sigma(D) \sim 8\eta^{-2}$ resulting in effective time steps which are twice as small as the time step (3.46). One may object that the values of γ satisfying (3.44a) and (3.44b) generally give rise to nine-point formulae, which are roughly twice as laborious as the five-point formula with $\gamma = 1 + \rho^2$. However, by choosing $\gamma = \rho^2 = 1$ we also obtain a five-point formula, while (3.44a) and (3.44b) are still satisfied.

3.4 The method of non-uniform complex time steps

A second important class of initial boundary value problems leads to operators of type $A = 1 + \tau D$ where D has imaginary eigenvalues $\delta_j = iy_j$, $j = 1, 2, \ldots, m$. Such difference schemes arise from transport problems (see Richtmyer [1957], chapter VII).

The eigenvalues α_j of A are given by $\alpha_j = 1 + \tau\delta_j = 1 + i\tau y_j$, so that

(3.47) $$|\alpha_j| = \sqrt{1 + \tau^2\, y_j^2}\,.$$

Clearly, there is no B-H-K stability.

Let δ_j satisfy the inequality

(3.48) $$0 < \delta_0 \leq |\delta_j| \leq \sigma(D),\ j = 1, 2, \ldots, m.$$

Then, von Neumann's condition is satisfied if, as $\tau \to 0$,

(3.49) $$\tau \leq \frac{c}{\sigma^2(D)}\,,$$

where c does not depend on τ. In most cases this is a very stringent

condition. For instance, let $1/\sigma(D)$ be proportional to the space step ξ (compare chapter III). Then, if we are not content with the accuracy obtained, we may decide to reduce ξ by a factor $\frac{1}{2}$, which involves a reduction of τ by a factor $\frac{1}{4}$. In such cases it may be desirable to soften the stability conditions.

As in the preceding subsection the operator A is replaced by a polynomial operator

$$P_n(\tau D) = 1 + \tau D + \beta_2 \tau^2 D^2 + \ldots + \beta_n \tau^n D^n .$$

The eigenvalues of this operator are given by the values of $P_n(\tau\delta_j) = P_n(i\tau y_j)$, where $j = 1, 2, \ldots, m$. Unlike to the situation in the preceding subsection, these values here are complex. We have the following theorem.

Theorem 3.8

Let $Q_n(z)$ be a polynomial of degree n in z which has the form

$$(3.50) \qquad Q_n(z) = (1 - \beta_2 z + \beta_4 z^2 + \ldots)^2 + z(1 - \beta_3 z + \beta_5 z^2 + \ldots)^2 ,$$

and let b(n) be a positive number independent of τ such that $Q_n(z)$ is less than 1 for $0 < z < b(n)$. Then, the scheme $u_{k+1} = P_n(\tau D)u_k$ is B-H-K stable and satisfies von Neumann's condition if

$$(3.51) \qquad \tau < \frac{\sqrt{b(n)}}{\sigma(D)} .$$

Proof

Let us define the numbers $z_j = \tau^2 y_j^2$, $j = 1, 2, \ldots, m$. Then it is easily verified that

$$|P_n(\tau\delta_j)| = |P_n(i\tau y_j)| = \sqrt{Q_n(z_j)} .$$

Thus, if (3.51) is satisfied we have

$$(3.52) \qquad \sigma(P_n(\tau D)) = \max_j \sqrt{Q_n(z_j)} < 1 .$$

This proves the theorem (compare theorem 3.2).

We shall now discuss the construction of $Q_n(z)$ for n = 2, 3 and 4, with the extra requirement that b is as large as possible.

The polynomial $Q_2(z)$

In the case n = 2, $Q_n(z)$ assumes the form

$$Q_2(z) = \beta_2^2 z^2 - (2\beta_2 - 1)z + 1. \tag{3.53}$$

It is easily verified that we may take

$$b = \frac{2\beta_2 - 1}{\beta_2^2},$$

provided that $\beta_2 > \frac{1}{2}$. The maximal value of b is assumed for $\beta_2 = 1$. Hence we obtain for the scheme

$$u_{k+1} = (1 + \tau D + \tau^2 D)u_k \tag{3.54}$$

the stability condition

$$\tau < \frac{1}{\sigma(D)}. \tag{3.55}$$

It is not possible to factorize the operator $P_2(\tau D)$ in real, linear factors. In fact, we have

$$u_{k+1} = (1 + \tau_1 D)(1 + \tau_2 D)u_k, \tag{3.54'}$$

where $\tau_{1,2} = \frac{1}{2}(1 \pm i\sqrt{3})\tau$. Scheme (3.54) may be interpreted as the original scheme with non-uniform, complex time steps.

The polynomial $Q_3(z)$

In the case n = 3 we have

$$Q_3(z) = \beta_3^2 z^3 + (\beta_2^2 - 2\beta_3)z^2 + (1 - 2\beta_2)z + 1. \tag{3.56}$$

The inequality $Q_3(z) < 1$ leads, for $0 < z < b$, to the inequality

$$\beta_3^2 z^2 + (\beta_2^2 - 2\beta_3)z + 1 - 2\beta_2 < 0. \tag{3.57}$$

From the conditions $Q_3(0) = 1$ and $Q_3'(0) \leq 0$ we deduce that

$$\beta_2 \geq \frac{1}{2}, \quad \beta_3 \geq \frac{1}{2}\beta_2^2. \tag{3.58}$$

Let us write (3.57) in the equivalent form

$$z^2(\beta_2 - \frac{1}{z})^2 + z^3(\beta_3 - \frac{1}{z})^2 < 1. \tag{3.57'}$$

In the $\beta_2\beta_3$-plane this inequality represents the interior of an ellipse with its centre in the point $(\frac{1}{z}, \frac{1}{z})$ and a horizontal axis of length $2/z$ (see figure 3.4). The shaded region in figure 3.4 consists of points satisfying (3.58) and (3.57').

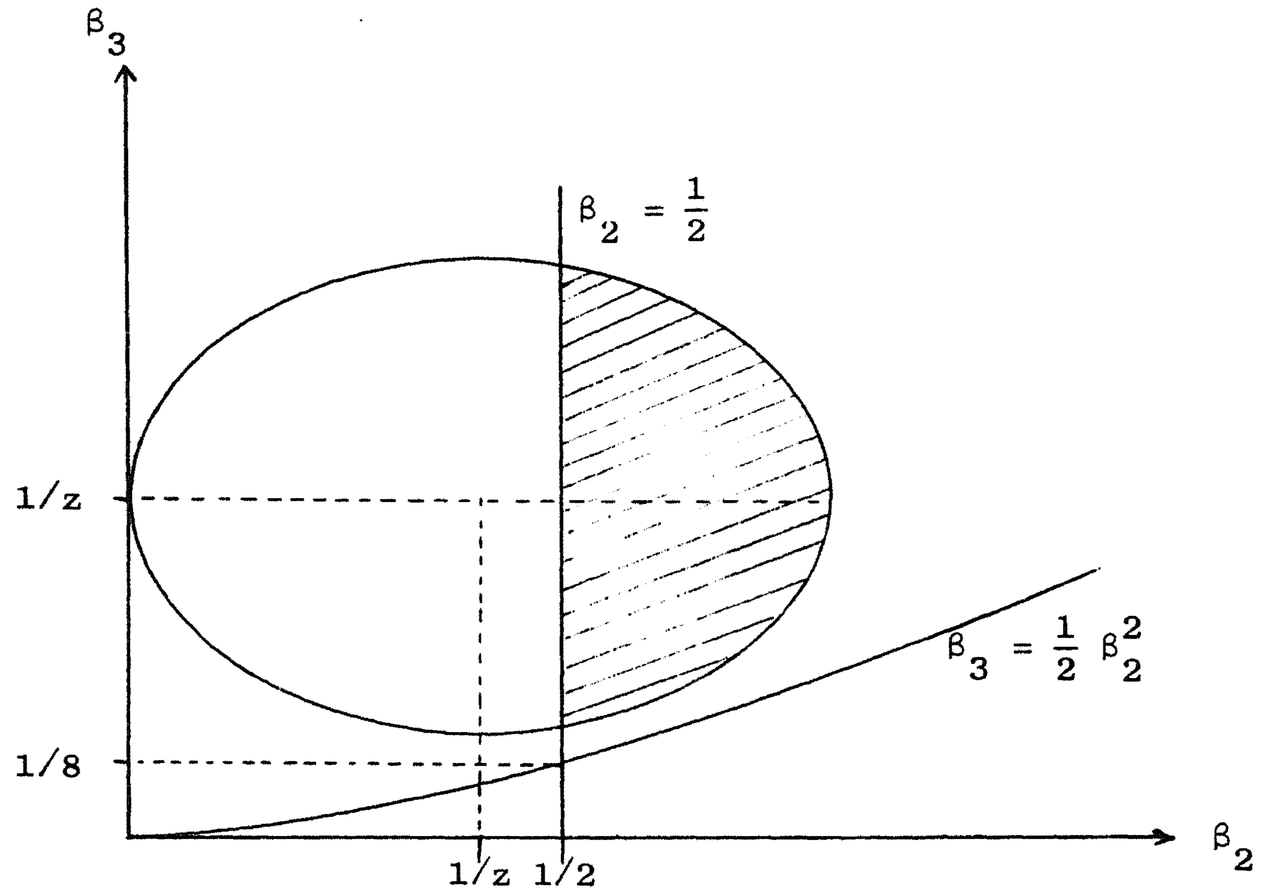

fig. 3.4

It is clear from the figure that b assumes its maximal value for $\beta_2 = \frac{1}{2}$ and $\beta_3 = \frac{1}{4}$, namely the value $b = 4$. We obtain the difference scheme

$$u_{k+1} = (1 + \tau D + \frac{1}{2}\tau^2 D^2 + \frac{1}{4}\tau^3 D^3)u_k \tag{3.59}$$

with the stability condition

(3.60) $$\tau < \frac{2}{\sigma(D)} .$$

In order to save storage room one may use the factorized form

(3.59') $$u_{k+1} = -(1 - \frac{\tau}{z_0} D)(1 - \frac{1}{4} z_0(2 + z_0)\tau D - \frac{1}{4} z_0 \tau^2 D^2)u_{k+1},$$

where z_0 is the real root of the equation $z^3 + 2z^2 + 4z + 4 = 0$ ($z_0 \sim -1.56$).

The polynomial $Q_4(z)$

Analogously to the considerations given for $Q_3(z)$ we obtain for $n = 4$ the scheme

(3.61) $$u_{k+1} = (1 + \tau D + \frac{1}{2} \tau^2 D^2 + \frac{1}{6} \tau^3 D^3 + \frac{1}{24} \tau^4 D^4)u_k$$

with the stability criterium

(3.62) $$\tau < \frac{2\sqrt{2}}{\sigma(D)} .$$

In practice one should use the factorized form

(3.61') $$u_{k+1} = (1 + p\tau D + q\tau^2 D^2)(1 + r\tau D + s\tau^2 D^2)u_k,$$

where p, q, r and s have the approximate values

(3.63) $$p = .9148 , \quad q = .2646 , \quad r = .0852 , \quad s = .1575.$$

From theorem 3.8 it follows that the difference schemes we have constructed are B-H-K stable and satisfy von Neumann's condition. A further advantage over the original scheme is the linear dependence of τ upon $1/\sigma(D)$ (compare condition 3.49)).

In order to compare the effectiveness of the schemes $u_{k+1} = P_n(\tau D)u_k$ we introduce the effective time step

(3.64) $$\tau_{eff} = \frac{\tau}{n} .$$

From the conditions (3.55), (3.60) and (3.62) we obtain respectively

$$(3.65) \qquad \tau_{eff} < \frac{0.5}{\sigma(D)}\ , \quad \tau_{eff} < \frac{0.66}{\sigma(D)}\ , \quad \tau_{eff} < \frac{0.71}{\sigma(D)}\ .$$

In the numerical treatment of the North Sea Problem (chapter III) we shall deal with a difference scheme which may be improved by the method of non-uniform complex time steps.

3.5 Implicit difference schemes

Instead of employing polynomial operators $P_n(\tau D)$ to soften the stability conditions, one may stabilize the scheme $u_{k+1} = (1 + \tau D)u_k$ by using implicit schemes, i.e.

$$(3.66) \qquad (1 - \tau C)u_{k+1} = (1 + \tau(D - C))u_k,$$

where C is a difference operator not depending on k. Such schemes are of type (2.1) with

$$(3.67) \qquad A = (1 - \tau C)^{-1}(1 + \tau(D - C)).$$

Theorem 3.9

Let the operators C and D have the same set of eigenfunctions with eigenvalues γ_j and δ_j respectively. Then scheme (3.66) is B-H-K stable and satisfies von Neumann's condition if τ satisfies the inequality

$$(3.68) \qquad \max_j \{\tau[|\delta_j|^2 - 2(\mathrm{Re}\gamma_j \mathrm{Re}\delta_j + \mathrm{Im}\gamma_j \mathrm{Im}\delta_j)] + 2\mathrm{Re}\delta_j\} < 0.$$

Proof

The eigenvalues α_j of the operator A are given by

$$\alpha_j = \frac{1 - \tau(\gamma_j - \delta_j)}{1 - \tau\gamma_j}\ .$$

The theorem is proved if $\sigma(A) < 1$. This condition reduces to (3.68) as may be verified by direct computation.

Theorem 3.10

Every difference scheme $u_{k+1} = (1 + \tau D)u_k$ with $Re\delta_j \leq 0$ and $\delta_j \neq 0$ for $j = 1, 2, \ldots, m$ can be transformed into an implicit scheme which is B-H-K stable and satisfies von Neumann's condition for unrestricted time intervals.

Proof

From theorem 3.9 it follows that, if

$$(3.69) \qquad Re\delta_j \leq 0, \ |\delta_j|^2 - 2(Re\gamma_j Re\delta_j + Im\gamma_j Im\delta_j) < 0, \ j = 1, 2, \ldots, m,$$

then scheme (3.66) is B-H-K stable and satisfies von Neumann's condition for unrestricted time intervals τ.

The first condition of (3.69) is satisfied by hypothesis.

The second condition may be satisfied by choosing

$$(3.70) \qquad C = qD, \ q > \frac{1}{2} .$$

It may be remarked that the implicit scheme we constructed above is unconditionally stable, at the cost of the solving of the equations

$$(3.71) \qquad (1 - \tau C)u_{k+1} = g_k, \ k = 0, 1, 2, \ldots, N-1,$$

where g_k is a known function. The numerical solution of such matrix equations is a large and widely studied subject (see for instance Varga [1962]). In chapter IV we shall study iterative methods for solving matrix equations.

Finally, we observe that the condition $Re\delta_j \leq 0$ of theorem 3.10 is related to a similar condition one has to impose upon ordinary differential equations in order to guarantee stability in the sense of Lyapunov (cf. Cesari [1959], p. 21).

3.6 Introduction of dissipative terms

In the subsections 3.3 and 3.4 we have stabilized a given difference scheme $u_{k+1} = (1 + \tau D)u_k$ by replacing the operator $1 + \tau D$ by a poly-

nomial operator $P_n(\tau D) = 1 + \tau D + \beta_2 \tau^2 D^2 + \ldots + \beta_n \tau^n D^n$. Physically, such polynomial operators may be interpreted as the operator $1 + \tau D$ to which viscosity terms of increasing order are added (cf. Saul'yev [1964], p. 41). In the case of implicit difference schemes we replaced $1 + \tau D$ by the operator A defined by formula (3.67). This operator may also be interpreted as the original operator plus viscosity terms. Expansion of A in a formal Taylor series yields

$$A = 1 + \tau D + \tau^2 DQD + \tau^3 DQDQD + \ldots, \tag{3.72}$$

where Q is defined by $DQ = C$. The viscosity terms are more general than the terms occurring in the polynomial $P_n(\tau D)$. In literature, Q is called the artificial viscosity (cf. Lax and Wendroff [1960]). In the following chapter we shall discuss a difference scheme which is stabilized considerably by introducing a viscosity term of second order, i.e. a term of the form $\tau^2 DQDu_k$.

Other types of dissipative terms are also possible. In chapter III, section 5 we shall need an artificial friction term. In this section the effect of an artificial intertia term will be discussed.

Let us change the scheme $u_{k+1} = (1 + \tau D)u_k$ to the three-level scheme

$$\frac{u_{k+1} - u_{k-1}}{2\tau} + c\tau \frac{u_{k+1} - 2u_k + u_{k-1}}{\tau^2} = Du_k, \tag{3.73}$$

where c is a parameter which is uniformly bounded as $\tau \to 0$. This scheme is still a consistent approximation of equation (2.7) (for simplicity the inhomogeneous term and the boundary conditions are neglected). We write (3.73) in the form

$$u_{k+1} = (\beta + \gamma D)u_k + (1 - \beta)u_{k-1}, \tag{3.73'}$$

where

$$\beta = \frac{4c}{1 + 2c} \quad \text{and} \quad \gamma = \frac{2}{1 + 2c}.$$

By introducing the vector v_k with components u_k and u_{k-1}, and the operator

$$A = \begin{pmatrix} \beta + \gamma D & 1 - \beta \\ 1 & 0 \end{pmatrix}, \tag{3.74}$$

the difference scheme reduces to the two-level scheme

$$\vec{v}_{k+1} = A\vec{v}_k. \tag{3.75}$$

The initial function $\vec{v}_1$ of this scheme is composed of the net functions u_0 and u_1, where u_1 may be found from the formula $u_1 = (1 + \tau D)u_0$.

The eigenvalues α of A satisfy the equation

$$\alpha^2 - (\beta + \gamma\delta)\alpha + \beta - 1 = 0, \tag{3.76}$$

where δ represents the eigenvalues of D.

We consider the two important cases of negative and imaginary eigenvalues δ .

Theorem 3.11

Let the eigenvalues of D be negative. Then, scheme (3.75) is B-H-K stable and satisfies von Neumann's condition if

$$\tau < \frac{4c}{\sigma(D)} . \tag{3.77}$$

Proof

If the coefficients of the quadratic equation $\alpha^2 - S\alpha + P = 0$ are real, and if they satisfy the inequalities

$$P < 1, \; 1 - S + P > 0, \; 1 + S + P > 0,$$

then the roots of the equation are within the unit circle. Application of this criterium to equation (3.76) yields the conditions

$$\beta < 2, \; -\gamma\delta > 0, \; 2\beta + \gamma\delta > 0. \tag{3.78}$$

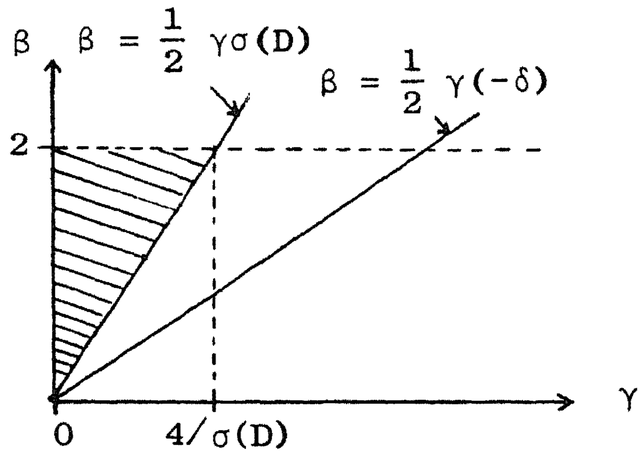

fig. 3.5

In figure 3.5 we have illustrated the region of points (γ,β) which satisfy the inequalities (3.78) (shaded part of the figure). From the definition of β and γ the theorem follows immediately.

Comparing this criterium with the criteria derived in section 3.3, we see that three-level scheme (3.73') admits arbitrary time steps τ. However, one needs twice as much storage room as the two-level schemes given in section 3.3 require. Further, when the net is refined we still have to choose τ proportional to $1/\sigma(D)$. Therefore, only when a rough knowledge of the analytical solution is needed, the three-level formula may be preferred over the two-level formula.

It may be remarked that the three-level scheme proposed by du Fort and Frankel [1953] for the one-dimensional diffusion equation as well as the generalization to two-dimensional diffusion equations by Saul'yev [1964], p. 157, are special cases of scheme (3.73').

We now consider the case where D has non-zero, imaginary eigenvalues.

Theorem 3.12

Let the eigenvalues of D be imaginary and let the eigenfunctions be linearly independent. Then, scheme (3.75) is B-H-K stable and satisfies von Neumann's condition if

$$\tau < \frac{1}{\sigma(D)} \quad . \tag{3.79}$$

Proof

We introduce the real variable $z = -i\gamma\delta$, where δ is an eigenvalue of D. The eigenvalues α of A are given by

$$(3.80) \qquad \alpha = \frac{1}{2}\beta + \frac{1}{2} iz \pm \sqrt{(1 - \frac{1}{4} z^2) + (\frac{1}{4}\beta^2 - \beta + i\beta z)}.$$

It is easily verified that for $z = 2$ and $\beta \neq 0$ the eigenvalues α are outside the unit circle. However, for $|z| \leq 2$ and $\beta = 0$ we have that

$$|\alpha| = |\frac{1}{2} iz \pm \sqrt{1 - \frac{1}{4} z^2}| = 1.$$

Hence we shall choose $\beta = 0$, i.e. $c = 0$ and $\gamma = 2\tau$.
From the definition of z we see that von Neumann's condition is satisfied for $\tau \leq 1/\sigma(D)$.
Further, using the condition imposed on the eigenfunctions of D, we see that for $|z| < 2$, the eigenfunctions of A are all linearly independent, i.e. $\mu(A) = 1$. From this and theorem 3.2 we deduce that the scheme is B-H-K stable if $\tau < 1/\sigma(D)$.

Comparing the stability criterium of theorem 3.12 with the criteria given in section 3.4 (see formula (3.65)), we see that the three-level scheme admits slightly larger time steps than the method of non-uniform complex time steps. However, the last method also guarantees B-H-K stability when D has dependent eigenfunctions. On the other hand, the three-level scheme has accuracy $O(\tau^2)$ in τ, where the two-level formulae have accuracy $O(\tau)$.

An example of a difference scheme which may be successfully transformed into a three-level formula is given in the following chapter.

3.7 Concluding remarks

In the preceding subsections we have given criteria which guarantee that the difference scheme is B-H-K stable (in the weak sense) and satisfies von Neumann's condition as well. In most cases these criteria only consist of an inequality for the time step τ. Our first remark is that the strict observance of these inequalities is not necessary in order to satisfy von Neumann's condition. For future reference we summarize the von Neumann conditions associated with the methods of

non-uniform time steps, implicit difference schemes, and of three-level schemes.

TABLE 3.1

Type of the method	Von Neumann's condition	
Non-uniform time steps	real eigenvalues	$\tau \leq \frac{2n^2}{\sigma(D)}$ as $\tau \to 0$
	imag.eigenvalues	$\tau \leq \frac{\sqrt{b(n)}}{\sigma(D)}$ as $\tau \to 0$
Implicit difference schemes	$\max_j \{\tau[\lvert\delta_j\rvert^2 - 2(\mathrm{Re}\gamma_j \mathrm{Re}\delta_j + \mathrm{Im}\gamma_j \mathrm{Im}\delta_j)] + 2\mathrm{Re}\delta_j\} \leq 0$ as $\tau \to 0$	
Three-level schemes	real eigenvalues	$\tau \leq \frac{4c}{\sigma(D)}$ as $\tau \to 0$
	imag.eigenvalues	$\tau \leq \frac{1}{\sigma(D)}$ as $\tau \to 0$

We recall that in these conditions it is assumed that terms of order τ are omitted in the operator A.

Secondly, we remark that these conditions are not changed when D is allowed to have an eigenvalue equal to zero, which was excluded in the preceding subsections.

Finally, when D has a zero eigenvalue with multiplicity 1, then theorems 3.7, 3.8, 3.9, 3.11 and 3.12 still guarantee B-H-K stability.

4. Stability with respect to the inhomogeneous terms and the boundary conditions

In the preceding section we discussed the stability with respect to the initial condition, which was determined by the operator P_0. We expressed a number of stability criteria in terms of the operators A_k.

In this section we shall consider the stability with respect to the inhomogeneous term and the boundary conditions. These kinds of stability are determined by the operators Q_k and S_k respectively (see formula (2.3)). Using the norm definition (2.4), we are only concerned with the behaviour of the quantities $|||Q||| = \max_k ||Q_k||$ and $|||S||| = \max_k ||S_k||$. We shall give estimates for $|||Q|||$ and $|||S|||$ in terms of the operators A_k, I_k and B_k.

In analogy to the definition of $|||A|||$ we define

$$|||I||| = \max_l ||I_l||, \quad |||B||| = \max_l ||B_l|| . \tag{4.1}$$

Theorem 4.1

If the operators A_k are non-stationary we have

$$|||Q||| \leq |||I||| \frac{1 - |||A|||^N}{1 - |||A|||}, \tag{4.2}$$

where $N = T/\tau$, and if they are stationary we have

$$|||Q||| \leq N |||I||| \max_k ||A^k|| = N |||I||| \; |||P_0||| \tag{4.3}$$

or

$$|||Q||| \leq C(\tau) |||I||| \sum_{l=0}^{N} l^{p-1} [\sigma(A)]^l, \tag{4.4}$$

where $C(\tau)$ is a function of τ (and not of T) and p is the largest order of all diagonal submatrices J_r of the Jordan normal form J of A with $\sigma(J_r) = \sigma(A)$.
Further, the same inequalities hold for $|||S|||$ when I is replaced by B.

Proof

From the definition of the operator Q_k we obtain

$$||Q_k|| = \sup_{|||f|||=1} \left\| \sum_{l=1}^{k+1} P_{lk} \; I_{l-1} \; f_{l-1} \right\| \leq$$

$$\leq \sup_{|||f|||=1} \sum_{l=1}^{k+1} \|P_{lk}\| \, \|I_{l-1}\| \, \|f_{l-1}\| \cdot$$

By using (2.4) and (4.1) this reduces to

$$\|Q_k\| \leq |||I||| \sum_{l=1}^{k+1} \|P_{lk}\| . \tag{4.5}$$

For non-stationary operators A_k we may write

$$\|Q_k\| \leq |||I|||(1 + |||A||| + |||A|||^2 + \ldots + |||A|||^k),$$

from which we derive the inequality (4.2).

For stationary operators A inequality (4.5) reduces to

$$\|Q_k\| \leq |||I||| \sum_{l=0}^{N-1} \|A^l\| . \tag{4.5'}$$

The estimate (4.3) follows immediately from (4.5').

In order to prove (4.4) we write

$$\sum_{l=0}^{N-1} \|A^l\| = \sum_{l=0}^{N-1} (l^{1-p}[\sigma(A)]^{-l}\|A^l\|) l^{p-1}[\sigma(A)]^l \leq$$

$$\leq \max_l (l^{1-p}[\sigma(A)]^{-l}\|A^l\|) \sum_{l=0}^{N-1} l^{p-1}[\sigma(A)]^l .$$

From formula (3.14) it follows that the first factor of the right member of this inequality is bounded by a function of τ which does not depend on T. This proves (4.4).

It is clear that the same inequalities hold for $|||S|||$ when we replace $|||I|||$ by $|||B|||$.

Theorem 4.2

If the boundary conditions are of the first kind, then scheme (2.1) is stable with respect to the boundary conditions.

Proof

It is easily seen from formula (2.3) that for first kind boundary conditions

$$S_k\phi = B_k\phi_k = \phi_{k+1},$$

so that

$$|||S||| = 1.$$

This proves the theorem.

Example 4.1

Using estimate (4.3) we see that difference schemes of type (2.8) are both F-W and R-F stable with respect to the inhomogeneous term whenever the schemes are F-W and R-F stable respectively with respect to the initial condition. Further, since the boundary conditions in (2.8) are of the first kind, we have F-W, R-F and B-H-K stability with respect to the boundary conditions. In order to have B-H-K stability with respect to the inhomogeneous term (strong stability) it is not sufficient to require merely stability with respect to the initial condition (weak stability). From theorem 4.1 we derive for non-stationary processes the condition $|||A||| < 1$ and for stationary processes the condition $\sigma(A) < 1$. Note that for $\sigma(A) = 1$ we have linear instability when $p = 1$ and instability when $p > 1$. In connection with this we mention the special case discussed in subsection 3.3, where the operator D had negative eigenvalues. Such schemes were stabilized by replacing the operator $A = 1 + \tau D$ by the Chebyshev polynomial operator $C_n(a,b,D)$ with $a \geq 0$ and $b \geq \sigma(D)$. This led to schemes of the form

$$u_{k+1} = C_n(a,b,D)u_k + If_k + \phi_{k+1}, \tag{4.6}$$

where $I = \tau$ and where τ is given by formula (3.32). In order to guarantee strong stability we require that $\sigma(C_n(a,b,D)) < 1$. From formula (3.32) it follows that we must choose $a > 0$ (see figure 3.3 in subsection 3.3).

In practice, one uses the fact that the polynomial $C_n(a,b,\delta)$ has real zeroes and one actually applies the following scheme (compare formula (3.37)).

$$
(4.6') \qquad \begin{cases} u_k^{(0)} = u_k, \\ u_k^{(r+1)} = (1 + \omega_r D)u_k^{(r)} + I_r f_k^{(r)} + \phi_k^{(r+1)}, \\ \qquad\qquad r = 0, 1, \ldots, n-1, \\ u_{k+1} = u_k^{(n)}, \end{cases}
$$

where $I_r = \omega_r$, and where $f_k^{(r)}$ and $\phi_k^{(r)}$ are discretizations of the inhomogeneous term and the boundary term at time $t_k^{(r)} = t_k + \sum_{r=0}^{r} \omega_r$.

These formulae are slightly more accurate than (4.6) and more convenient from the computational point of view. However, scheme (4.6') has the disadvantage that the numerical solution u_k is affected by systematic round-off errors rather than by random round-off errors. For large values of n this phenomenon may destroy the solution completely (numerical instability). In van der Houwen [1967 c] the problem of numerical stability is discussed and arrangements of the relaxation parameters ω_r are given, which keep the systematic error small. Nevertheless, it is desirable that scheme (4.6') is strongly stable, i.e. $a > 0$.

Example 4.2

Consider the difference scheme defined by formula (3.5) for the one-dimensional diffusion equation in the interval $0 \leq x \leq 1$. In this case we have

$$
I = \frac{\tau\sqrt{r}}{\sqrt{r} - b\sqrt{\tau}} X_+, \quad B = \frac{\sqrt{\tau}}{\sqrt{r} - b\sqrt{\tau}} E_+ \quad \text{in } x_j = 0,
$$

$$
I = \tau, \quad B = 0 \quad \text{in internal net points},
$$

$$
I = \tau X_-, \quad B = \sqrt{\frac{\tau}{r}}\, E_+ \quad \text{in } x_j = 1.
$$

In example 3.1 we have given conditions with respect to the maximum norm which guarantee R-F stability with respect to the initial function ($b \leq 0$ and $r \leq \frac{1}{2}$). Since $I = O(\tau)$ it is clear from the estimate (4.3) that these conditions also guarantee R-F stability with respect to the inhomogeneous term. However, since $B = O(\sqrt{\tau})$ we derive from (4.3) that $|||S||| \leq O(1/\sqrt{\tau})$, so that we only have F-W stability with respect to the boundary function ϕ.
Further, we see from (3.6) that the maximum norm of A is never less than 1, hence we have no strong stability with respect to the maximum norm.

Chapter III

THE NORTH SEA PROBLEM

1. Introduction

The analytical discussion of the non-stationary motion of a shallow sea subjected to a windfield meets with considerable difficulties. In a sequence of papers concerning the analytic computation of the water elevation of the North Sea (see van Dantzig and Lauwerier [1960 a], [1960 b] and Lauwerier[1960 a], [1960 b], [1960 c], [1961 a], [1961 b]), one had to simplify the mathematical model by neglecting the influence of irregularities of the coast and the influence of the Channel leak stream. In fact, the North Sea model considered in these papers was a rectangle bounded on three sides by coasts and bordering on an infinitely deep ocean on the remaining side. Further, the depth was assumed either to be uniform or to increase exponentially in the direction of the ocean. In Lauwerier and Damsté [1963] a difference scheme was constructed in order to deal with more realistic models for the North Sea Problem. This scheme, however, was subject to very stringent stability conditions.

In this chapter it will be shown that this scheme can be stabilized by applying the method of non-uniform complex time steps and the method of dissipative terms developed in the preceding chapter. This will result in three explicit difference schemes satisfying stability conditions which are acceptable from the computational point of view.

It will turn out that the friction due to the bottom stress plays an important rôle in the stability properties of the schemes. Since the friction will damp the perturbations of the data, we shall be interested in the stability of the schemes when friction is omitted. Scheme I, which is obtained by the method of non-uniform complex time steps and scheme III, a variant of the three-level scheme discussed

in the preceding chapter, remain stable for vanishing bottom stress. Scheme II, which is a variant of a scheme given by Fischer, becomes unstable. We compensate for this by introducing an artificial friction term in those cases where the bottom friction is small or absent.

2. The mathematical model

2.1 The partial differential equations

The mathematical model for the North Sea Problem considered in this chapter, is an initial boundary value problem for the following two equations (cf. van Dantzig and Lauwerier [1960 a] or Veltkamp [1960]).

$$(2.1)\qquad \begin{cases} \dfrac{\partial}{\partial t}\,\vec{W} = -\,\lambda\vec{W} - \Omega T\vec{W} - gh\ \mathrm{grad}\ Z + \vec{F}, \\[2ex] \dfrac{\partial}{\partial t}\,Z = -\,\mathrm{div}\ \vec{W}, \end{cases}$$

where

- t is the time coordinate,
- $\vec{W}$ is the horizontal component of the velocity of the water, averaged in vertical direction from bottom to surface,
- Z is the elevation of the water surface with respect to its equilibrium position,
- $\vec{F}$ is the surface stress due to the windfield,
- λ is a coefficient of friction,
- Ω is the coefficient of Coriolis,
- g is the constant of gravity,
- h is the depth function,
- T denotes a rotation through a rightangle in the horizontal plane in the positive sense,
- grad,div are defined in the horizontal plane.

The first of these equations constitutes the equation of the motion of the sea. The terms $-\lambda\vec{W}$, $-\Omega T\vec{W}$ and $-gh\ \mathrm{grad}\ Z$ represent the forces

caused by the friction at the bottom of the sea, the rotation of the earth and the deviation from the equilibrium position of the water surface.
The second equation is the equation of continuity.

2.2 The boundary conditions

The sea is bounded by oceans and coasts. The oceanic part of the boundary will be denoted by Γ_{oc} and the coastal part by Γ_c. The boundary conditions are as follows.
The elevation Z is prescribed along Γ_{oc}, i.e.

$$Z = Z_{oc} \tag{2.2}$$

(cf. Veltkamp [1960], p. 11) and the stream $\vec{W}$ is defined along Γ_{oc} by the equation of motion.
Along the coasts one has the condition

$$W_n = 0, \tag{2.3}$$

where W_n is the component of $\vec{W}$ in the direction normal to the coast.
We shall derive the equation for the stream along Γ_c. Then, the elevation follows from the equation of continuity.
Let $\vec{c}$ be the unit vector tangential (in the positive sense) to the coast. Forming the inner-product between $\frac{\partial}{\partial t}\vec{W}$ and $\vec{c}$ we obtain from (2.1)

$$\frac{\partial}{\partial t}(\vec{c}\cdot\vec{W}) = -\lambda(\vec{c}\cdot\vec{W}) - \Omega(\vec{c}\cdot T\vec{W}) - gh(\vec{c}\cdot\nabla Z) + (\vec{c}\cdot\vec{F}).$$

By (2.3) we find for the stream along Γ_c

$$\frac{\partial}{\partial t}(\vec{c}\cdot\vec{W}) = -\lambda(\vec{c}\cdot\vec{W}) - gh(\vec{c}\cdot\nabla Z) + (\vec{c}\cdot\vec{F}),$$

so that

$$\frac{\partial}{\partial t}\vec{W} = -\lambda\vec{W} - gh(\vec{c}\cdot\nabla Z)\vec{c} + (\vec{c}\cdot\vec{F})\vec{c}. \tag{2.3'}$$

When the initial state of the sea is given, equations (2.1), (2.2) and (2.3) completely determine the subsequent motion of the sea (Veltkamp [1960], p. 12).

Clearly the North Sea Problem is an initial boundary value problem in the sense of definition 2.1 (chapter I). $\{\vec{W};Z\}$ is the unknown function, $\{\vec{F};0\}$ is the inhomogeneous term, and $\{(\vec{c}\cdot\vec{F})\vec{c};0\}$ and $\{\vec{F};Z_{oc}\}$ represent the boundary function along the coast and the ocean respectively.

Introducing rectangular coordinates x and y in the horizontal plane and defining the operator $\tilde{D}$ by

$$\tilde{D} = \begin{pmatrix} -\lambda & \Omega & -\sqrt{gh}\,\frac{\partial}{\partial x} \\ -\Omega & -\lambda & -\sqrt{gh}\,\frac{\partial}{\partial y} \\ -\sqrt{gh}\,\frac{\partial}{\partial x} & -\sqrt{gh}\,\frac{\partial}{\partial y} & 0 \end{pmatrix} \tag{2.4}$$

in non-boundary points,

$$\tilde{D} = \begin{pmatrix} -\lambda & 0 & -\sqrt{gh}\; s(\vec{c}\cdot\nabla) \\ 0 & -\lambda & -\sqrt{gh}\; c(\vec{c}\cdot\nabla) \\ -\sqrt{gh}\,\frac{\partial}{\partial x} & -\sqrt{gh}\,\frac{\partial}{\partial y} & 0 \end{pmatrix} \tag{2.5}$$

along the coast and by

$$\tilde{D} = \begin{pmatrix} -\lambda & \Omega & -\sqrt{gh}\,\frac{\partial}{\partial x} \\ -\Omega & -\lambda & -\sqrt{gh}\,\frac{\partial}{\partial y} \\ 0 & 0 & 0 \end{pmatrix} \tag{2.6}$$

along the ocean,

we may express the North Sea Problem by the initial value problem

$$\frac{\partial}{\partial t}\vec{S} = \tilde{D}\,\vec{S} + \vec{G}, \tag{2.7}$$

where

$$\vec{S} = \begin{pmatrix} U \\ V \\ \sqrt{gh}\, Z \end{pmatrix}, \quad \vec{G} = \begin{pmatrix} F_1 \\ F_2 \\ 0 \end{pmatrix} \text{ in non-boundary points, } \vec{G} = \begin{pmatrix} s(sF_1+cF_2) \\ c(sF_1+cF_2) \\ 0 \end{pmatrix}$$

along Γ_c, and where $\vec{G} = \begin{pmatrix} F_1 \\ F_2 \\ \sqrt{gh}\, \frac{\partial}{\partial t} Z_{oc} \end{pmatrix}$ along Γ_{oc}.

Here, (U,V), (F_1,F_2) and (s,c) are the components in the x,y-plane of $\vec{W}$, $\vec{F}$ and $\vec{c}$ respectively. Note that we have inserted in $\vec{G}$ both the inhomogeneous and the boundary term.

2.3 The stationary solution

If the windfield $\vec{F}$ does not depend on time and if we take $Z_{oc} = 0$ (cf. Veltkamp [1960]), then the motion of the sea becomes stationary for $t \to \infty$. The stationary state is governed by the equation

$$\tilde{D}\,\vec{S} + \vec{G} = 0. \tag{2.8}$$

Introducing a streamfunction Φ by means of

$$U = -\frac{\partial \Phi}{\partial y}, \quad V = \frac{\partial \Phi}{\partial x} \tag{2.9}$$

(cf. van Dantzig and Lauwerier [1960 b]), we may reduce equation (2.8) to an elliptic equation with oblique boundary conditions along the ocean.

$$\begin{cases} \lambda \Delta \Phi + A \frac{\partial \Phi}{\partial x} + B \frac{\partial \Phi}{\partial y} = C, \\ \Phi = 0 \quad \text{along} \quad \Gamma_c, \quad \lambda \frac{\partial \Phi}{\partial y} + \Omega \frac{\partial \Phi}{\partial x} = -F_1 \text{ along } \Gamma_{oc}, \end{cases} \tag{2.8'}$$

where $$A = \frac{\partial \lambda}{\partial x} - h^{-1}(\lambda \frac{\partial h}{\partial x} + \Omega \frac{\partial h}{\partial y}), \quad B = \frac{\partial \lambda}{\partial y} - h^{-1}(\lambda \frac{\partial h}{\partial y} - \Omega \frac{\partial h}{\partial x}),$$

and where $$C = \nabla \times \vec{F} - h^{-1}(F_1 \frac{\partial h}{\partial x} - F_2 \frac{\partial h}{\partial y}).$$

In the following chapter, where elliptic boundary value problems are treated, we discuss numerical methods to solve problem (2.8').

3. The characteristic criterium

Before constructing difference approximations of the North Sea Problem (2.8) we derive from the characteristics of equation (2.8) a criterium for the time steps τ_k, which has to be satisfied by every difference scheme (cf. Forsythe and Wasow [1960], p. 18).

The characteristic equation of the North Sea Problem is given by

$$\det\left[pI + q\begin{pmatrix} 0 & 0 & \sqrt{gh} \\ 0 & 0 & 0 \\ \sqrt{gh} & 0 & 0 \end{pmatrix} + r\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \sqrt{gh} \\ 0 & \sqrt{gh} & 0 \end{pmatrix} \right] = 0,$$

where p, q and r are the direction cosines of the line elements (dt,dx,dy) perpendicular to the characteristic directions (cf. Forsythe and Wasow [1960], p. 384). This equation may be reduced to

$$p(p^2 - gh\, q^2 - gh\, r^2) = 0,$$

which is satisfied by line elements parallel to the x,y-plane and the line elements parallel to the generators of the cone

$$t^2 - gh\, x^2 - gh\, y^2 = 0.$$

Hence the characteristics are given by directions parallel to the t-axis and parallel to the generators of the cone

$$t^2 - (gh)^{-1}x^2 - (gh)^{-1}y^2 = 0. \tag{3.1}$$

From (3.1) it follows that a characteristic line element (dt,dx,dy) satisfies the relation

$$dt = \sqrt{\frac{(dx)^2 + (dy)^2}{gh}}\,. \tag{3.2}$$

We shall investigate the effect of this relation on difference approximations of the North Sea Problem. Let us assume that the difference scheme is of the form

$$\vec{s}_{k+1} = A_k\, \vec{s}_k, \tag{3.3}$$

where $\vec{s}_k = [\vec{S}_k]_d = [\vec{S}(x,y,t_k)]_d$ (cf. chapter II) and where inhomogeneous terms are neglected. Further, we assume that in the point $\vec{x}_0$ the vector function $\vec{s}_{k+1}$ is calculated from the vector function $\vec{s}_k$ in points $\vec{x}_j$ satisfying $|\vec{x}_j - \vec{x}_0| \leq \rho$. The netpoints $(\vec{x}_j, t_k)$ form the domain of dependence of the difference solution in the net point $(\vec{x}_0, t_{k+1})$. Therefore, we have from (3.2) the necessary condition

$$\tau_k = t_{k+1} - t_k \leq \frac{\rho}{\sqrt{gh}} \quad . \tag{3.4}$$

This criterium for the time step τ_k is called the characteristic criterium.

4. The use of non-uniform complex time steps (scheme I)

In this section we construct a difference scheme for the North Sea Problem using the polynomial method discussed in chapter II, section 3.4.

In the x,y-plane we define a rectangular net with spatial steps ξ and η, and with respect to this net we define difference operators D_x and D_y which are consistent approximations of the differential operators $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ respectively. We assume that the operators D_x and D_y make no use of net points outside the boundary $\Gamma_c \cup \Gamma_{oc}$. Hence D_x and D_y will depend on the net point $\vec{x}_j$ in which they are applied, accordingly whether $\vec{x}_j$ is a boundary point or not. Further, we define the operator D which arises from $\tilde{D}$ by replacing $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ by D_x and D_y. Clearly D is a consistent approximation of $\tilde{D}$. Let us consider the difference scheme

$$\vec{s}_{k+1} = (1 + \tau D)\vec{s}_k + \tau \vec{g}_k \tag{4.1}$$

(compare ex. 2.2 of chapter II). We shall now discuss the stability of this scheme.

The usual procedure in stability analysis is to find the eigenvalues of the operator D. In this case, however, the operator D is too complicated.

We shall follow another approach of the stability problem. Let us ex-

tend the net over the whole x,y-plane and let us define on this extended net a difference scheme with constant coefficients which equals the given difference scheme (4.1) at the net point $\vec{x}_j$ of the original net ($\vec{x}_j \in G_\tau \cup \Gamma_\tau$). We may add such a local difference scheme to each net point $\vec{x}_j$ of $G_\tau \cup \Gamma_\tau$. The difference scheme which is actually applied may be considered as a combination of these local schemes. It is assumed that a given difference scheme is stable when its local difference schemes are stable (O'Brien, Hyman and Kaplan [1951], p. 226, Rjabenki and Filippov [1960], p. 64 and Leendertse [1967]). In most stability investigations one neglects the influence of the boundary conditions and restricts the considerations to the local stability of the internal net points (cf. Fischer [1959], p. 62, Lauwerier and Damsté [1963], p. 172, and Harris and Jelesnianski [1964], p. 420).

The internal local stability of scheme (4.1) was investigated by Fischer [1959] for the central difference approximation of $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, and by Lauwerier [1963] for the average central difference approximation. These approximations are particular cases of the general difference approximation

$$(4.2)\quad D_x = \frac{aY_+ + b + aY_-}{2a + b}\,\frac{X_+ - X_-}{2\xi}\,, \quad D_y = \frac{aX_+ + b + aX_-}{2a + b}\,\frac{Y_+ - Y_-}{2\eta}\,,$$

where X_+ and Y_+ are translations defined in example 3.1 of chapter II, and where a and b are real parameters which are not both equal to zero. For a = 0 and b = 0 we have the central and the average central difference form, respectively. The eigenfunctions of the local operator $D(\vec{x}_j)$ with $\vec{x}_j \in G_\tau$ are given by

$$(4.3)\qquad \vec{d}_0(\vec{\omega})\ \exp[i(\vec{\omega}\cdot\vec{x})],$$

where $\vec{x}$ runs through all net points and $\vec{\omega}$ is a two-dimensional vector index with components ω_1 and ω_2. The vector $\vec{d}_0(\vec{\omega})$ depends on $\vec{x}_j$ and $\vec{\omega}$, and is an eigenvector of the matrix

$$(4.4)\qquad \hat{D}(\vec{x}_j,\vec{\omega}) = \begin{pmatrix} -\lambda & \Omega & -i\sqrt{gh}\,\delta_1 \\ -\Omega & -\lambda & -i\sqrt{gh}\,\delta_2 \\ -i\sqrt{gh}\,\delta_1 & -i\sqrt{gh}\,\delta_2 & 0 \end{pmatrix},$$

where $i\delta_1$ and $i\delta_2$ are eigenvalues of the operators D_x and D_y corresponding to the eigenfunctions $\exp i(\vec{\omega}\cdot\vec{x})$. A straightforward computation leads to the expressions

$$(4.5)\qquad \delta_1 = \frac{\sin \omega_1\xi(b + 2a\cos \omega_2\eta)}{(b + 2a)\xi}\ ,\quad \delta_2 = \frac{\sin \omega_2\eta(b + 2a\cos \omega_1\xi)}{(b + 2a)\eta}\ .$$

It is easily seen that the eigenvalues of $D(\vec{x}_j)$ are given by the eigenvalues of the matrix $\hat{D}(\vec{x}_j,\vec{\omega})$. These eigenvalues determine the stability of scheme (4.1) in the point $\vec{x}_j$. The analysis of Lauwerier [1963] resulted in the following conditions for the B-H-K stability

$$(4.6)\qquad \tau < \frac{\lambda\xi^2\eta^2}{gh(\xi^2 + \eta^2)} \quad \text{with} \quad \tau < \mathrm{Min}\left(\frac{2}{3\lambda}\ ,\ \frac{2\lambda}{\lambda^2 + \Omega^2}\right).$$

These conditions lead to unacceptably small time steps. This is due to the small North Sea value of λ. In a realistic model we have

$$(4.7)\qquad \lambda \sim 25\ 10^{-6}\ \mathrm{sec}^{-1},\ \Omega \sim 125\ 10^{-6}\ \mathrm{sec}^{-1},\ g \sim 10\ \mathrm{m\ sec}^{-2},$$

$$h_{max} = 200\ \mathrm{m} \quad \text{and} \quad \xi = \eta = 2\ 10^4\ \mathrm{m},$$

which lead to $\tau \sim 5$ sec. For calculations where T is 20 to 60 hours, this time step will require a large amount of computation time.
We note that the scheme is R-F stable under less stringent conditions. To show this we neglect the friction and Coriolis terms in $D(\vec{x}_j)$ (according to lemma 3.2, chapter II). The eigenvalues $\delta(\vec{\omega})$ of the reduced operator $D_0(x_j)$ are identical to the eigenvalues of the corresponding amplification matrix $\hat{D}_0(\vec{x}_j,\vec{\omega})$, i.e.

$$(4.8)\qquad \delta_1(\vec{\omega}) = 0,\ \delta_{2,3}(\vec{\omega}) = \pm\, i\sqrt{gh(\delta_1^2 + \delta_2^2)}.$$

Applying formula (3.49) of chapter II we find that von Neumann's necessary condition is satisfied when

$$\tau \leq \frac{c}{\sigma^2(D_0(\vec{x}_j))}, \tag{4.9}$$

where

$$\sigma(D_0(\vec{x}_j)) = \max_{\vec{\omega}} \sqrt{gh(\delta_1^2 + \delta_2^2)} \tag{4.10}$$

and where c is a constant which does not depend on τ. It is easily verified that $\hat{D}_0(\vec{x}_j,\vec{\omega})$ is a normal matrix, hence condition (4.9) is also sufficient for R-F stability in the net point $\vec{x}_j$.
The value of $\sigma(D_0(\vec{x}_j))$ depends on ξ, η, a and b, and will be discussed at the end of this section. For the moment we only remark that τ depends quadratically on ξ and η, which means that halving the values of ξ and η implies a four times smaller value of the time step.
The considerations above indicate that it is desirable to look for difference schemes which are more appropriate in actual computation.

Formula (4.8) shows that the eigenvalues of $D_0(\vec{x}_j)$ are imaginary. This immediately suggests the application of the method of non-uniform complex time steps discussed in chapter II, section 3.4. For instance, we may define the scheme (scheme I)

$$\vec{s}_{k+1} = (1 + \tau D + \frac{1}{2}\tau^2 D^2 + \frac{1}{6}\tau^3 D^3 + \frac{1}{24}\tau^4 D^4)\vec{s}_k + \tau \vec{g}_k \tag{4.11}$$

(compare formula (3.61) of chapter II). According to theorem 3.8 and formula (3.62) of chapter II, the necessary and sufficient condition for R-F stability in the internal net point $\vec{x}_j$ is

$$\tau \leq \frac{2\sqrt{2}}{\sigma(D_0(x_j))}. \tag{4.12}$$

Note that τ now depends linearly on ξ and η.

Next we discuss the B-H-K stability of scheme I.
The eigenvalues δ of the matrix $\hat{D}(\vec{x}_j,\vec{\omega})$ satisfy the equation

$$(\delta + \lambda)(\delta^2 + \lambda\delta + gh(\delta_1^2 + \delta_2^2) + \Omega^2) - \lambda\Omega^2 = 0 \tag{4.13}$$

and the eigenvalues α of scheme I are given by

$$\alpha = 1 + \tau\delta + \frac{1}{2}\tau^2\delta^2 + \frac{1}{6}\tau^3\delta^3 + \frac{1}{24}\tau^4\delta^4. \tag{4.14}$$

Assuming that both $\lambda\tau$ and $\Omega\tau$ are small with respect to 1 we may approximate in (4.14) the eigenvalues δ by (compare Fischer [1959])

$$\delta_1(\vec{\omega}) \sim -\lambda, \quad \delta_{2,3}(\vec{\omega}) \sim -\frac{1}{2}\lambda \pm i\sqrt{gh(\delta_1^2 + \delta_2^2) + \Omega^2 - \frac{1}{4}\lambda^2}. \tag{4.15}$$

We remark that for $\lambda = 0$ and $\Omega = 0$ these expressions represent the exact solution of equation (4.13).
Let us consider the case where $\lambda = 0$ or $\lambda << \Omega$. Then we may neglect the real parts in the expressions for $\delta(\vec{\omega})$ and we obtain the condition

$$\tau < \frac{2\sqrt{2}}{\sqrt{\sigma^2(D_0(\vec{x}_j)) + \Omega^2}} \sim \frac{2\sqrt{2}}{\sigma(D_0(\vec{x}_j))}. \tag{4.16}$$

If λ cannot be neglected with respect to Ω, we divide the eigenvalues into two classes for which $\tau^2 gh(\delta_1^2 + \delta_2^2) < \lambda\tau$ and $\tau^2 gh(\delta_1^2 + \delta_2^2) \geq \lambda\tau$ respectively. In the first case we see from (4.15) that $|\tau\delta| << 1$, so that $\alpha \sim 1 + \tau\delta$. These eigenvalues are the eigenvalues of scheme (4.1). An analysis along the lines of Lauwerier [1963] yields the conditions

$$\tau < \frac{\lambda}{gh(\delta_1^2 + \delta_2^2)}, \tag{4.17}$$

$$\tau < \text{Min}\left(\frac{2}{3\lambda}, \frac{2\lambda}{\lambda^2 + \Omega^2}\right). \tag{4.18}$$

The first condition is satisfied by the assumption above, the second condition is identical to the second condition of (4.6).
In the second case we have

$$\sqrt{\tau^2 gh(\delta_1^2 + \delta_2^2) + \Omega^2\tau^2 - \frac{1}{4}\lambda^2\tau^2} \geq \sqrt{\lambda\tau + \Omega^2\tau^2 - \frac{1}{4}\lambda^2\tau^2} \sim \sqrt{\lambda\tau}.$$

Since $\lambda\tau << \sqrt{\lambda\tau} << 1$ we see from (4.15) that the real parts of $\tau\delta$ may be neglected, which yields condition (4.16).
It is easily verified that the eigenfunctions corresponding to the eigenvalues α are linearly independent, hence the conditions (4.16) and (4.18) are sufficient conditions.

We now discuss the spectral radius of the matrix $D_0(\vec{x}_j)$ as a function of ξ, η, a and b, i.e. the function $\sigma(D_0(\vec{x}_j)) = \sigma(\xi,\eta,a,b)$.

Theorem 4.1

The function $\sigma(\xi,\eta,a,b)$ satisfies the relations

(4.19a) $$\sigma(\xi,\eta,a,0) \leq \sigma(\xi,\eta,a,b),$$

(4.19b) $$\sigma(\xi,\eta,a,b) \leq \sigma(\xi,\eta,0,b),$$

(4.20) $$\sigma(\xi,\eta,a,0) = \sqrt{gh}/\mathrm{Min}(\xi,\eta),$$

(4.21) $$\sigma(\xi,\eta,0,b) = \sqrt{(1/\xi^2 + 1/\eta^2)gh}.$$

Proof

Inequality (4.19b) and equation (4.21) are obvious from (4.5) and (4.10).

Inequality (4.19a) follows from the fact that the function

$$\delta_1^2 + \delta_2^2 = \sin^2\omega_1\xi\left(1 - \frac{2a}{b+2a}(1 - \cos\omega_2\eta)\right)^2\xi^{-2} +$$

$$+ \sin^2\omega_2\eta\left(1 - \frac{2a}{b+2a}(1 - \cos\omega_1\xi)\right)^2\eta^{-2}$$

decreases for every set of fixed values of ω_1, ω_2 and a, when b decreases.

Finally, for b = 0 we have

$$\delta_1^2 + \delta_2^2 = \frac{\sin^2\omega_1\xi(1 - \sin^2\omega_2\eta)}{\xi^2} + \frac{\sin^2\omega_2\eta(1 - \sin^2\omega_1\xi)}{\eta^2} ,$$

which is a harmonic function of $\sin^2\omega_1\xi$ and $\sin^2\omega_2\eta$. Therefore the maximum value is reached at the boundary of the $(\sin^2\omega_1\xi, \sin^2\omega_2\eta)$-domain. This leads to equation (4.20).

From this theorem we see that condition (4.16) is most restrictive for the central difference form (a = 0), namely

$$\tau < \frac{2\sqrt{2}\ \xi\ \eta}{\sqrt{gh(\xi^2 + \eta^2)}}, \tag{4.16'}$$

and least restrictive for the average central difference form (b = 0), namely

$$\tau < \frac{2\sqrt{2}\ \mathrm{Min}(\xi, \eta)}{\sqrt{gh}} . \tag{4.16''}$$

The central difference form is most widely used in practice, as the computational labour is minimal. The average central difference form is roughly twice as laborious and was first used by Lauwerier and Damsté [1963]. For both approximations scheme I has the property that in net points where the stream is evaluated, the values of the elevation are not needed. Further, the average central difference form has the additional advantage that in half of all the net points neither stream nor elevation values are needed, so that the computational labour per time step is roughly equal for the two approximations and the total amount of labour will be a factor $\sqrt{1 + \mathrm{Min}(\xi^2/\eta^2, \eta^2/\xi^2)}$ in favour of the average central difference form.

Finally, for the characteristic criterium of scheme I we refer to section 7.

5. Introduction of dissipative terms (scheme II)

By introducing an artificial viscosity and friction term into scheme (4.1) we shall construct a difference scheme (scheme II) which satisfies still weaker stability conditions.

Omitting the terms of order τ we define scheme II by the formula

$$\vec{s}_{k+1} = (1 + \tau D_0 + \tau^2 D_0 Q D_0)\vec{s}_k + \tau \vec{g}_k, \tag{5.1}$$

where Q is a 3×3 matrix, i.e.

$$Q = \begin{pmatrix} q_1 & q_2 & q_3 \\ r_1 & r_2 & r_3 \\ s_1 & s_2 & s_3 \end{pmatrix}, \tag{5.2}$$

with real entries to be defined later. The term $\tau^2 D_0 Q D_0 \vec{s}_k$ represents an artificial viscosity term (compare Lax and Wendroff [1960] and chapter II, section 3.6 of the present paper). Scheme (5.1) clearly is consistent with (2.7) for $\lambda = \Omega = 0$.

For practical reasons we require that (5.1) can be written as

$$(1 - \tau C)\vec{s}_{k+1} = (1 + \tau(D_0 - C))\vec{s}_k + \tau(1 - \tau C)\vec{g}_k, \tag{5.1'}$$

where the upper non-diagonal matrix elements of the matrix operator C are zero. From the relation

$$(1 - \tau C) = (1 + \tau D_0 Q)^{-1}$$

we derive that

$$q_3 = r_3 = s_1 = s_2 = s_3 = 0$$

which leads to the operator

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -gh(q_1 D_x + r_1 D_y) & -gh(q_2 D_x + r_2 D_y) & 0 \end{pmatrix}. \tag{5.3}$$

We investigate the R-F stability of scheme (5.1) in the internal net point $\vec{x}_j$. We have

$$A_0(\vec{x}_j) = 1 + \tau D_0(\vec{x}_j) + \tau^2 D_0(\vec{x}_j) Q D_0(\vec{x}_j) = 1 + \tau(1-\tau C)^{-1} D_0(\vec{x}_j).$$

The eigenvalues α of $A_0(\vec{x}_j)$ satisfy the equation

$$\det[\alpha(1 - \tau\hat{C}(\vec{\omega})) - 1 - \tau(\hat{D}_0(\vec{x}_j,\vec{\omega}) - \hat{C}(\vec{\omega}))] = 0, \tag{5.4}$$

where $\hat{C}(\vec{\omega})$ is the amplification matrix corresponding to C.
A direct computation leads to the following expressions for the eigenvalues α.

$$\alpha_1 = 1, \ \alpha_{2,3} = \frac{1}{2} S \pm \sqrt{S^2 - 4P}, \tag{5.5}$$

where

$$S = 2 - \tau^2 gh(q_1\delta_1^2 + (r_1 + q_2)\delta_1\delta_2 + r_2\delta_2^2),$$

$$P = S - 1 + \tau^2 gh(\delta_1^2 + \delta_2^2).$$

The eigenvalues α satisfy the inequality $|\alpha| \leq 1$ (von Neumann's condition) when

$$P \leq 1, \ 1 - S + P \geq 0, \ 1 + S + P \geq 0.$$

This results in the criterium

$$0 \leq \tau^2 gh(\delta_1^2 + \delta_2^2) \leq \tau^2 gh(q_1\delta_1^2 + (r_1 + q_2)\delta_1\delta_2 + r_2\delta_2^2) \leq 2gh +$$

$$+ \frac{1}{2}\tau^2 gh(\delta_1^2 + \delta_2^2).$$

Choosing $q_1 = r_2 = 1$ and $r_1 = -q_2$ we obtain the least restrictive condition for τ, i.e.

$$\tau \leq \frac{2}{\sigma(D_0(\vec{x}_j))}. \tag{5.6}$$

It is easily verified that the eigenfunctions of $A_0(\vec{x}_j)$ are linearly independent, hence condition (6.5) also is sufficient for R-F stability

in the point $\vec{x}_j$.

In order to compare condition (5.6) with the R-F stability condition (4.12) for scheme I we introduce the effective time step τ_{eff} (cf. formula (3.65) of chapter II).
For scheme I we find $\tau_{eff} = \tau/4$ and for scheme II, owing to the triangular form of the operator C, $\tau_{eff} = \tau$. From this it follows that we have gained a factor $2\sqrt{2} \sim 3$ in computation time.

Next we define the complete scheme II by the formula

$$(5.1') \qquad (1 - \tau C)\vec{s}_{k+1} = (1 + \tau(D - C))\vec{s}_k + \tau(1 - \tau C)\vec{g}_k.$$

This scheme is a consistent approximation of equation (2.7) and is R-F stable in the point $\vec{x}_j$ when τ satisfies condition (5.6).
We shall discuss the B-H-K stability of scheme II.
First we note that for $q_2 = 0$ and $a = 0$, Fischer [1959] obtained the following conditions for B-H-K stability.

$$(5.7) \qquad \tau < \xi\sqrt{\frac{1 - \frac{1}{2}\lambda\tau + \sqrt{1 - \lambda\tau + \Omega^2\tau^2}}{gh}}, \quad \tau < \frac{\lambda}{\Omega^2}, \quad \Omega \neq 0,$$

where $\xi = \eta$. For all practical purposes the first condition of (5.7) reduces to

$$(5.7') \qquad \tau < \frac{\sqrt{2}\,\xi}{\sqrt{gh}}.$$

We shall show, however, that it has certain advantages to include additional viscosity terms $\tau q_2 D_y u_{k+1}$ and $-\tau q_2 D_x v_{k+1}$ in the formula for the elevation z_{k+1}, and to use average central differences instead of central differences.
The method of analysis will be that of Lauwerier [1963], which uses the Hurwitz-criterium to guarantee that the eigenvalues α of the matrix $\hat{A}(\vec{x}_j, \vec{\omega})$ are within the unit circle. The eigenvalues α satisfy the equation

$$(5.8) \qquad \det[(\alpha - 1)(1 - \tau\hat{C}(\vec{\omega})) - \tau\hat{D}(\vec{x}_j, \vec{\omega})] = 0.$$

This equation may be reduced to

$$(5.8') \qquad (\alpha - 1)^3 + b_1(\alpha - 1)^2 + b_2(\alpha - 1) + b_3 = 0,$$

where

$$b_1 = \tau^2 gh(\delta_1^2 + \delta_2^2) + 2\lambda\tau,$$

$$b_2 = \tau^2 gh(\delta_1^2 + \delta_2^2)(1 + \lambda\tau - q_2\Omega\tau) + \lambda^2\tau^2 + \Omega^2\tau^2,$$

$$b_3 = \tau^2 gh(\delta_1^2 + \delta_2^2)\lambda\tau.$$

From the Routh-Hurwitz criteria we derive that the roots α of (5.8') are within the unit circle when

$$(5.9) \qquad \begin{cases} b_3 > 0, \\ 8 - 4b_1 + 2b_2 - b_3 > 0, \\ 2b_2 - 3b_3 > 0, \\ (b_2 - b_3)(b_1 - b_2 + b_3) > 0, \end{cases}$$

(compare Lauwerier and Damsté [1963] and Leendertse [1967]).
For $\delta_1^2 + \delta_2^2 \neq 0$ these conditions reduce to

$$(5.10) \qquad \tau < \frac{2}{\sigma(D_0(\vec{x}_j))} \sqrt{1 - \frac{1}{2}\lambda\tau - \frac{1}{2}q_2\Omega\tau}, \quad \tau < \frac{\lambda}{\lambda^2 + \Omega^2}$$

with

$$(5.11) \qquad -\frac{2 - \lambda\tau}{\Omega\tau} < q_2 < \frac{2 - \lambda\tau}{2\Omega\tau}.$$

For a detailed analysis of the inequalities (5.9) we refer to v.d. Houwen [1966].
If $\delta_1^2 + \delta_2^2 = 0$ we have for α the equation

$$(5.8'') \qquad (\alpha - 1)(\alpha^2 - 2(2 - \lambda\tau)\alpha + \tau^2(\lambda^2 + \Omega^2) - 2\lambda\tau + 1) = 0.$$

We have $\alpha_1 = 1$ and $|\alpha_{2,3}| < 1$ provided that $\tau < 2\lambda/(\lambda^2 + \Omega^2)$. This involves no further conditions (see (5.10)).

Since we have now proved that, if (5.10) and (5.11) are satisfied, all eigenvalues α are within the unit circle with the exception of one which is on the unit circle, it follows from theorem 3.2, chapter II, that scheme II is B-H-K stable in the point $\vec{x}_j$.

In practice the first condition of (5.10) reduces to the R-F stability condition (5.6) and is identical to Fischer's condition (5.7') for the central difference form. As was already observed in the preceding section we gain a factor $\sqrt{1 + \mathrm{Min}(\xi^2/\eta^2, \eta^2/\xi^2)} = \sqrt{2}$ by using the average central difference form.

The second condition of (5.10) corresponds to Fischer's second condition. If λ is comparable with Ω neither condition is a restriction of the time step τ. If $\lambda << \Omega$ the conditions are identical. Presently, we shall discuss this case further.

Fischer's third condition excludes calculations at the equator. From our analysis it follows, however, that this condition is not necessary to guarantee B-H-K stability. Finally, we remark that condition (5.11) hardly is a restriction of the value of q_2.

Comparing scheme II with scheme I we conclude that scheme I is to be preferred when $\lambda << \Omega$, and scheme II is to be preferred when λ and Ω are of the same order of magnitude. This suggests the introduction of an artificial friction term into scheme II for models where $\lambda << \Omega$. Let us replace λ by $\bar{\lambda} = \lambda + r\tau$ where r is a constant not depending on τ. Since the additional friction term vanishes for $\tau \to 0$, the difference scheme remains a consistent approximation of the analytical problem. We shall choose $\bar{\lambda}$ in such a way that condition (5.6) is not less restrictive than the second condition of (5.10), i.e.

$$\frac{\bar{\lambda}}{\bar{\lambda}^2 + \Omega^2} \geq \frac{2}{\sigma(D_0(\vec{x}_j))} .$$

For small values of $\Omega/\sigma(D_0(x_j))$ this leads to

$$\bar{\lambda} \geq \frac{2\Omega^2}{\sigma(D_0(\vec{x}_j))} . \tag{5.12}$$

Assuming that we use time steps near to their upper bound we obtain for r the value (compare van der Houwen [1967 a])

$$r = \mathrm{Max}(0,\ \Omega^2 - \frac{1}{2}\lambda\sigma(D_0(\vec{x}_j))). \tag{5.13}$$

In connection with this it is interesting to investigate a difference scheme proposed by Miss Sielecki [1967]. The scheme may be interpreted as Fischer's scheme in which for both the elevation and the stream field the most recent values are used. In the same manner we can modify scheme II which results in scheme (5.1') where C is now given by

$$C = \begin{pmatrix} 0 & 0 & 0 \\ \Omega & 0 & 0 \\ -\sqrt{gh}(D_x - q_2 D_y) & -\sqrt{gh}(q_2 D_x + D_y) & 0 \end{pmatrix}. \tag{5.14}$$

This transformation has the main effect of increasing the bottom stress and decreasing the Coriolis force in the y direction. To see this we express the new operator $\bar{A}(\vec{x}_j)$ in terms of the operator $A(\vec{x}_j)$ corresponding to scheme II. We find for $q_2 = 0$

$$\bar{A}(\vec{x}_j) = A(\vec{x}_j) + \Omega\tau^2 \begin{pmatrix} 0 & 0 & 0 \\ \lambda & -\Omega & i\sqrt{gh}\, D_x \\ -i\Omega\lambda\tau\sqrt{gh}\, D_y & i\Omega\tau\sqrt{gh}\, D_y & \tau^2 ghD_xD_y \end{pmatrix}. \tag{5.15}$$

Thus in the y direction the friction λ transforms to $\bar{\lambda} = \lambda - \Omega^2\tau$ and the coefficient of Coriolis Ω transforms to $\bar{\Omega} = \Omega - \Omega\lambda\tau$. We may expect, therefore, that for $\lambda = 0$ this scheme satisfies weaker stability conditions than scheme II. In fact, we have for $\lambda = q_2 = 0$ the eigenvalue equation

$$(\bar{\alpha} - 1)(\bar{\alpha}^2 + (\bar{b}_1 - 2)\bar{\alpha} + 1) = 0, \tag{5.16}$$

where

$$\bar{b}_1 = \tau^2 gh(\delta_1^2+\delta_2^2) + \tau^2\Omega^2 - \tau^3\Omega gh\delta_1\delta_2 \sim \tau^2(gh(\delta_1^2+\delta_2^2) + \Omega^2).$$

The eigenvalues are given by

$$\bar{\alpha}_1 = 1,\ \bar{\alpha}_{2,3} = 1 - \frac{1}{2}\bar{b}_1 \pm \sqrt{\frac{1}{4}\bar{b}_1^2 - \bar{b}_1}. \tag{5.17}$$

They are all on the unit circle if $0 \leq \bar{b}_1 < 4$, or equivalently if

$$\tau < \frac{2}{\sqrt{\sigma^2(D_0(\vec{x}_j))+\Omega^2}} \sim \frac{2}{\sigma(D_0(\vec{x}_j))}. \tag{5.18}$$

This condition is sufficient for B-H-K stability in the net point $\vec{x}_j$.

We concludè this section with a discussion of the damping effect of the operators $A(\vec{x}_j)$ when the dissipative friction terms are omitted. Let $\lambda = 0$ or $\bar{\lambda} = 0$, then we obtain from (5.8') the equation

$$(\alpha - 1)(\alpha^2 + (b_1 - 2)\alpha + 1 - b_1 + b_2) = 0 \tag{5.8'''}$$

with the solutions

$$\alpha_1 = 1,\ \alpha_{2,3} = 1 - \frac{1}{2}b_1 \pm \sqrt{\frac{1}{4}b_1^2 - b_2}. \tag{5.19}$$

From (5.10) it follows that $\frac{1}{4}b_1^2 < b_2$ so that

$$|\alpha_{2,3}|^2 = 1 - q_2\tau^2 gh(\delta_1^2 + \delta_2^2)\Omega\tau + \Omega^2\tau^2. \tag{5.20}$$

In figure 5.1 we have illustrated the behaviour of $|\alpha|^2$ as a function of $gh(\delta_1^2 + \delta_2^2)$ for $q_2 > 0$ and $q_2 = 0$. The time step τ is given by $\tau = \tau_{II} \sim 2/\sigma(D_0(\vec{x}_j))$.

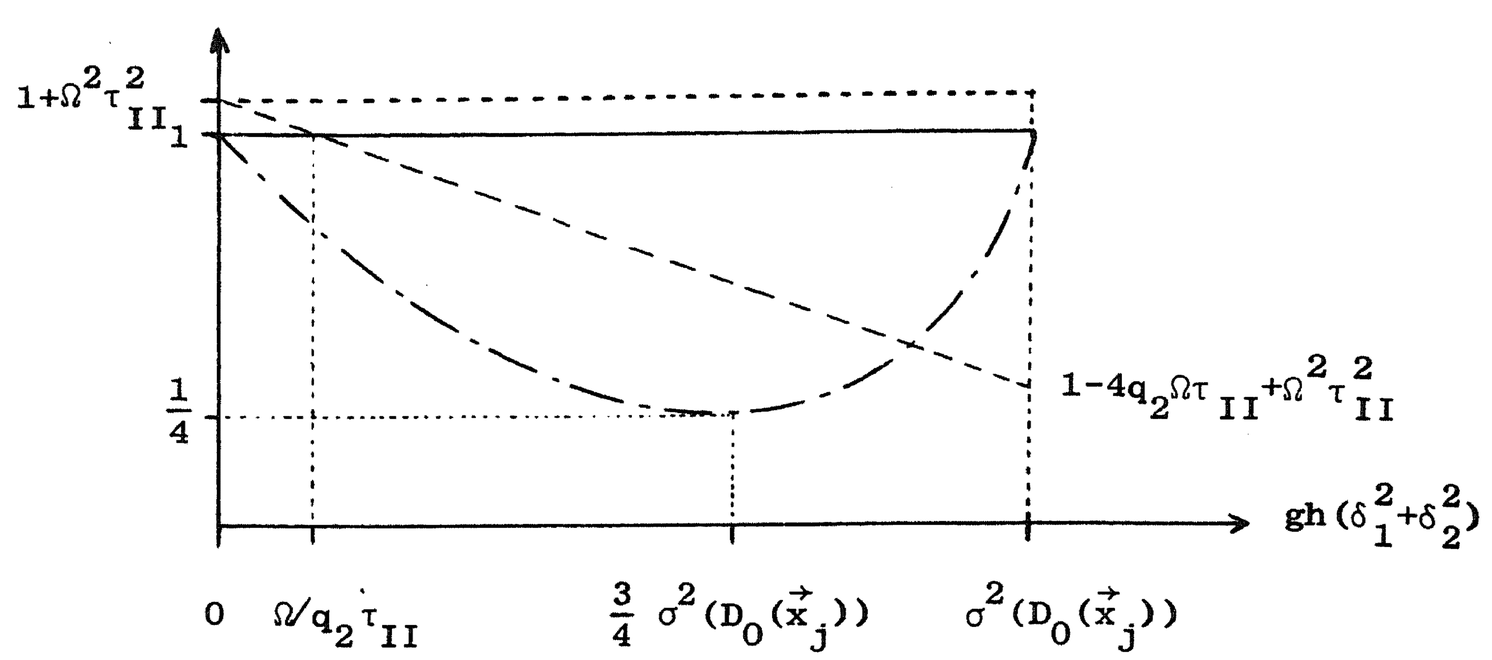

fig. 5.1 —.—.— scheme I

– – – – – scheme II with $q_2 > 0$

..... scheme II with $q_2 = 0$ (Fischer's scheme)

_____ common spectrum: $|\alpha_1|^2 = 1$

From the figure we see that for $q_2 = 0$, i.e. for Fischer's scheme, all eigenfunctions will increase with time, while for $q_2 > 0$ the greater part of the eigenfunctions is damped. In connection with this it is interesting to consider the spectrum of scheme I. From the theory given in chapter II, section 3.4 we derive that

$$(5.21) \qquad |\alpha_1| = 1, \quad |\alpha_{2,3}|^2 = 1 - \frac{1}{72}\tau^6(gh(\delta_1^2+\delta_2^2)+\Omega^2)^3 + \frac{1}{576}\tau^8(gh(\delta_1^2+\delta_2^2)+\Omega^2)^4,$$

hence no eigenfunction can increase with time (see figure 5.1).

6. Three-level schemes (scheme III)

Harris and Jelesnianski [1964] considered the following difference scheme

$$(6.1) \qquad \vec{s}_{k+1} = \vec{s}_{k-1} + 2\tau\, D\, \vec{s}_k + 2\tau\, \vec{g}_k.$$

This scheme is identical to the scheme which arises from (4.1) by applying the method of chapter II, section 3.6. From theorem 3.12 of chapter II we derive that von Neumann's condition is satisfied when

$$\tau \leq \frac{1}{\sigma(D_0(\vec{x}_j))} . \tag{6.2}$$

Since $\hat{D}_0(\vec{x}_j,\vec{\omega})$ is normal this condition also is sufficient for R-F stability in the point $\vec{x}_j$.

In order to derive conditions for the B-H-K stability we consider the eigenvalue equation of scheme (6.1), i.e.

$$\det[\alpha^2 - 2\tau\, \hat{D}(\vec{x}_j,\vec{\omega})\alpha - 1] = 0. \tag{6.3}$$

For $\lambda = 0$ this equation reduces to

$$(\alpha^2 - 1)(\alpha^4 - 2(1 - 2\tau^2 gh(\delta_1^2+\delta_2^2) - 2\tau^2\Omega^2)\alpha^2 + 1) = 0. \tag{6.3'}$$

If

$$\tau < \frac{1}{\sqrt{\sigma^2(D_0(\vec{x}_j) + \Omega^2}} \sim \frac{1}{\sigma(D_0(\vec{x}_j)} , \tag{6.4}$$

then we have three different roots on the unit circle. This proves the B-H-K stability of scheme (6.1) (compare condition (4.16) for scheme I). We may expect that for $\lambda \neq 0$ the eigenvalues α will lie within the unit circle which will improve the stability.

Next we consider the following three-level scheme (scheme III)

$$\vec{s}_{k+1} = 2\,\frac{2c + \tau(1 - \tau C)^{-1}D}{2c + 1}\,\vec{s}_k - \frac{2c - 1}{2c + 1}\,\vec{s}_{k-1} + \frac{2\tau}{2c + 1}\,\vec{g}_k, \tag{6.5}$$

where C is defined by (5.14) and c is a parameter independent of τ. This scheme arises from scheme (5.1'), (5.14) by adding an inertia term (see subsection 3.6 of chapter II). The eigenvalues δ of the operator $(1 - \tau C)^{-1}D$ may be derived from (5.17) for $\lambda = q_2 = 0$.
We find

$$(6.6)\qquad \delta_1 = 0,\ \delta_{2,3} = -\frac{\bar{b}_1}{2\tau} \pm \frac{i}{\tau}\sqrt{\bar{b}_1(1 - \frac{1}{4}\bar{b}_1)}.$$

Since these eigenvalues are not imaginary theorem 3.12 of chapter cannot be applied. We shall investigate the eigenvalues α of the amplification matrix of scheme (6.5). From equation (3.76) of chapter II we derive

$$(6.7)\qquad \alpha_j = \frac{2c + \tau\delta_j}{2c + 1} \pm \sqrt{\left(\frac{2c + \tau\delta_j}{2c + 1}\right)^2 - \frac{2c - 1}{2c + 1}},\quad j = 1, 2, 3.$$

From this relation it follows that

$$(6.7')\qquad \alpha_1 = 1,\ \frac{2c - 1}{2c + 1}.$$

Hence we must choose $c \geq 0$.

In figure 6.1 we have plotted $\mathrm{Max}(|\alpha_2|, |\alpha_3|)$ as a function of $\bar{b}_1$ for some values of c. Note that scheme (6.5) reduces to scheme (5.1'), (5.14) if c = .5 is substituted.

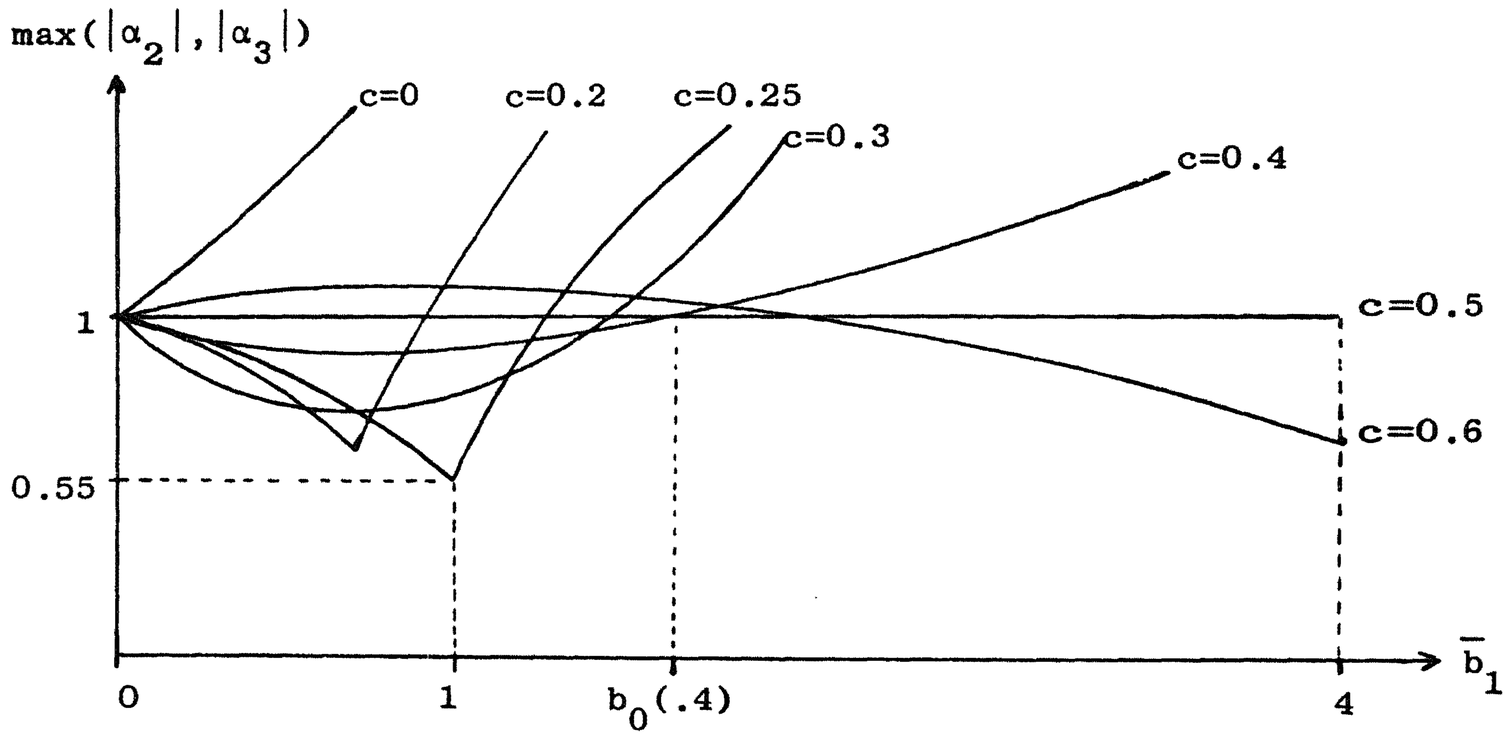

fig. 6.1

The scheme is stable for $0 < c \leq .5$ provided that τ is sufficiently small, i.e.

$$\tau < \frac{\sqrt{b_0(c)}}{\sqrt{\sigma^2(D_0(\vec{x}_j)) + \Omega^2}} \sim \frac{\sqrt{b_0(c)}}{\sigma(D_0(\vec{x}_j))} \quad , \tag{6.8}$$

where $b_0(c)$ is the non-zero value of $\bar{b}_1$ for which $\text{Max}(|\alpha_2|, |\alpha_3|) = 1$. For $c = .5$ the admissable time step τ becomes maximal. However, in this case none of the eigenfunction components is damped in actual computation. For $0 < c < .5$ the greater part of the eigenfunctions will decrease with time (see figure 6.1).

We conclude this section with a figure which illustrates the behaviour of the function $\sqrt{b_0(c)}$.

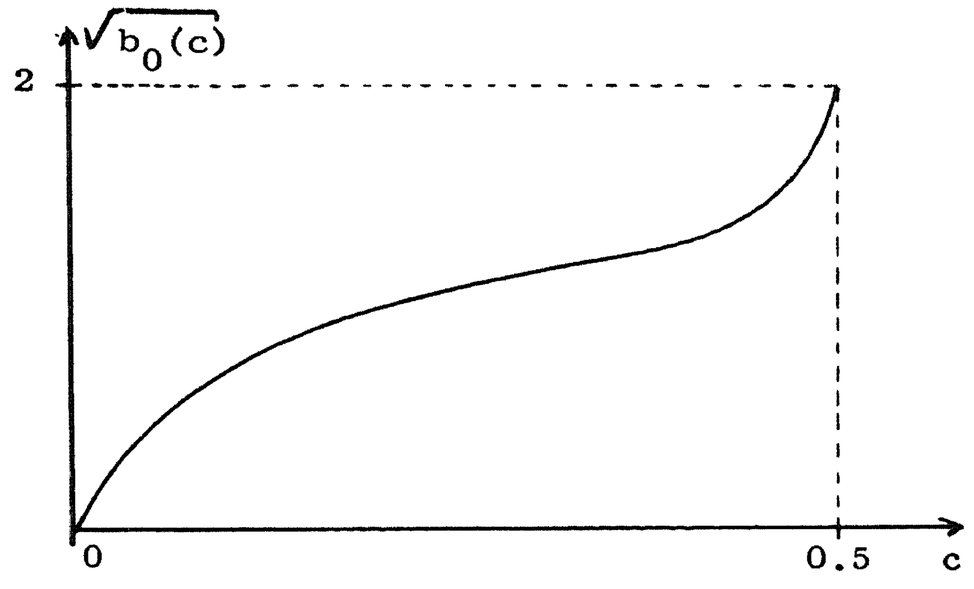

fig. 6.2

7. A survey of the stability properties of scheme I, II and III

In this section the results derived in the preceding sections are summarized. In table 7.1 we have listed the conditions for the local stability with respect to the initial condition in non-boundary points.

The R-F stability conditions do not depend on λ and Ω. From these conditions one may infer to R-F stability with respect to the inhomogeneous term $\vec{g}$.

The B-H-K stability conditions are weakened by an increase of λ, but they are strengthened by an increase of Ω. In general, if $\lambda \neq 0$ we may expect that weak stability implies strong stability.

TABLE 7.1

Stability conditions for the average central difference form with $\xi=\eta$

	Scheme I	Scheme II	Scheme III
Effective time step	$\tau_{eff} \sim \tau/4$	$\tau_{eff} \sim \tau$	$\tau_{eff} \sim \tau$
Characteristic criterium	$\tau \leq 4\sqrt{2}\xi/\sqrt{gh}$	$\tau \leq 2\sqrt{2}\xi/\sqrt{gh}$	$\tau \leq 2\sqrt{2}\xi/\sqrt{gh}$
R-F stability condition	$\tau < 2\sqrt{2}\xi/\sqrt{gh}$	$\tau < 2\xi/\sqrt{gh}$	$\tau < \sqrt{b_0(c)}\,\xi/\sqrt{gh}$ $0 < c \leq .5$
Additional B-H-K stability condition for $\bar{\lambda} \sim \Omega$	$\tau < 2/3\bar{\lambda}$ $\tau < 2\bar{\lambda}/(\bar{\lambda}^2+\Omega^2)$	$\tau < \bar{\lambda}/(\bar{\lambda}^2+\Omega^2)$	none
Additional B-H-K stability condition for $\bar{\lambda} = 0$	none	unstable	none
Damping effect for $\bar{\lambda} = 0$	linearly unstable	unstable	linearly unstable
Storage room	6 components	3 components	6 components

Chapter IV

ELLIPTIC DIFFERENTIAL EQUATIONS

1. Introduction

In the previous chapters we have discussed finite difference methods to solve linear initial boundary value problems. This enables us to find the numerical solution of important classes of hyperbolic and parabolic differential equations. Matters are different for elliptic differential equations. Such equations lead to pure boundary value problems to which the preceding theory cannot directly be applied. However, if the solution of an elliptic boundary value problem is interpreted as the stationary solution of an appropriate initial boundary value problem, then such problems can be treated by the methods of chapter II.

Here we shall consider the method of non-uniform real time steps to solve elliptic boundary value problems. In literature, this method is known as Richardson's method (cf. Forsythe and Wasow [1960], p. 226) or the Chebyshev iterative method (cf. Varga [1962], p. 138). Several authors have proposed accelerating procedures for Richardson's method to improve the convergence of the level functions u_k to the stationary solution u_∞. These accelerating procedures are based on the elimination of the fundamental modes of the error $u_k - u_\infty$. The elimination method given by Shortley [1953] eliminates in succession the fundamental modes by means of operators which are linear in the operator D. As will be shown in subsection 3.2 of this chapter, Shortley's method may be improved considerably by replacing the linear operators by Chebyshev operators of well-defined degree. We shall prove that of all polynomial elimination operators, these Chebyshev operators are optimal with regard to the rate of convergence of the scheme. Further, we shall investigate non-polynomial elimination operators, which have

certain interesting properties not possessed by polynomial operators. Basic for the elimination methods mentioned above, is the knowledge of the eigenvalues of the fundamental modes. We shall give formulae in order to calculate these fundamental (or first) eigenvalues. These formulae are less laborious than the method given by Flanders and Shortley [1950].

It may be remarked that Stiefel has given an elimination method in which the knowledge of the first eigenvalues is not necessary. However, experiments reported by Frank [1960] turned out to be unsatisfactory.

2. Definition of iterative processes

In this section the theory of the well-known iterative method called Richardson's method is reviewed. In order to make this method compatible with the theory given in chapter I and II, the solution of an elliptic boundary value problem will be interpreted as the stationary solution of an initial boundary value problem. Then, Richardson's method of first degree is identical to the method of non-uniform real time steps discussed in section 3.3 of chapter II, and Richardson's method of second degree is a special, non-stationary version of scheme (3.73'), chapter II.

2.1 Richardson's method of first degree

Let us assume that the difference solution of the discrete elliptic boundary value problem is given by the stationary solution of the scheme

$$u_{k+1} = (1 + \tau D)u_k + \tau g_0, \tag{2.1}$$

where g_0 does not depend on k. This scheme is of the type discussed in section 2 of chapter II with τg_0 representing the terms $I_k f_k$ and $B_k \phi_k$. The stationary solution of (2.1) satisfies the equation

$$Du_\infty + g_0 = 0. \tag{2.2}$$

Without loss of generality it can be assumed that D has real eigenvalues δ_j, $j = 1, 2, \ldots, m$ only. For, otherwise we can always replace D by D^*D and g_0 by D^*g_0, where D^* is the adjoint operator of D. Throughout this section we shall make the additional assumption that

$$-\sigma(D) = \delta_m \leq \delta_{m-1} \leq \ldots \leq \delta_1 = -\delta_0 < 0. \tag{2.3}$$

The application of the method of non-uniform real time steps (see chapter II, section 3.3 and example 4.1) results in the following iteration method.

$$\begin{cases} u_k^{(0)} = u_k, \\ u_k^{(r+1)} = (1 + \omega_r D)u_k^{(r)} + \omega_r g_0, \quad r = 0, 1, \ldots, n-1, \\ u_{k+1} = u_k^{(n)}, \end{cases} \tag{2.4}$$

where $k = 0, 1, \ldots, N-1$ and where the relaxation parameters ω_r are defined by formula (3.36) of chapter II. Iteration method (2.4) may be written more compactly as

$$u_{k+1} = C_n(a,b,D)u_k + h_0^{(n)} \tag{2.4'}$$

with

$$h_0^{(n)} = \sum_{l=1}^{n-1} \prod_{r=1}^{n-1} (1 + \omega_r D)\omega_{r-1} g_0 + \omega_{n-1} g_0. \tag{2.5}$$

For considerations on convergence it is convenient to introduce the error functions v_k and $v_k^{(r)}$ defined by

$$v_k = u_k - u_\infty, \quad v_k^{(r)} = u_k^{(r)} - u_\infty, \quad r = 0, 1, \ldots, n, \quad k = 0, 1, \ldots, N. \tag{2.6}$$

From (2.2) and (2.4) it follows that v_k satisfies the homogeneous scheme

$$v_{k+1} = C_n(a,b,D)v_k. \tag{2.7}$$

Assuming that n does not depend on k we obtain for the initial error v_N the estimate

$$\|v_N\| \leq \|C_n^N(a,b,D)\| \, \|v_0\| . \tag{2.8}$$

For $a/b \ll 1$, $0 \leq a \leq \delta_0$ and $b \geq \sigma(D)$ we derive from lemma 3.1 and theorem 3.7 of chapter II

$$\|C_n^N(a,b,D)\| \sim \nu N^{p-1}[\cosh(2n\sqrt{\frac{a}{b}})]^{-N}, \tag{2.9}$$

where ν is a constant and p is the largest order of all diagonal submatrices J_r of the Jordan normal form J of $C_n(a,b,D)$ with $\sigma(J_r) = \sigma(C_n(a,b,D))$. From (2.8) and (2.9) we see that $\|v_N\|$ is minimized by

$$a = \delta_0 , \quad b = \sigma(D) \tag{2.10}$$

for given values of n and N. Compare figure 3.3 of chapter II. Formulae (2.4) and (2.10) define Richardson's method of first degree (or order).

The greater part of the literature about Richardson's method deals with the case where D is a definite symmetric matrix (cf. Young [1953], Forsythe and Wasow [1960] and Varga [1962]). In that case, or more generally when D is a normal matrix with negative eigenvalues, approximation (2.9) is valid for N = 1, 2, ... with $p = \nu = 1$. For non-normal matrices (2.9) is valid for sufficiently large values of N.

Finally, note that the iterants $u_k^{(r)}$ $(r \neq 0, n)$ do not necessarily approximate u_∞. For large values of n this may be disadvantageous from a practical point of view.

2.2 Richardson's method of second degree

Let us consider the three-level scheme defined by

$$\begin{cases} u_k^{(0)} = u_k, \\ u_k^{(1)} = (1 + \gamma_0 D)u_k^{(0)} + \gamma_0 g_0, \\ u_k^{(r+1)} = (\beta_r + \gamma_r D)u_k^{(r)} + (1 - \beta_r)u_k^{(r-1)} + \gamma_r g_0, \\ \qquad\qquad r = 1, 2, \ldots, n-1, \\ u_{k+1} = u_k^{(n)}, \end{cases} \tag{2.11}$$

where $k = 0, 1, \ldots, N-1$ and where the parameters β_r and γ_r still have to be determined. If $N = 1$ and if β_r and γ_r do not depend on r, this scheme is equivalent to the three-level scheme discussed in chapter II, section 3.6. It has been shown there, that such stationary schemes are stable when

$$\beta < 2, \quad \gamma > 0, \quad 2\beta - \gamma\sigma(D) > 0, \tag{2.12}$$

and it is easily seen that these conditions also guarantee convergence when the scheme is used as an iteration process (compare figure 3.5 in chapter II). However, one may construct a non-stationary iteration method which converges faster than any stationary method.

We express the error $v_k^{(r)}$ of the iterant $u_k^{(r)}$ in terms of the error $v_k^{(0)}$, i.e.

$$v_k^{(r)} = Q_r(D)v_k^{(0)}, \tag{2.13}$$

where $Q_r(D)$ is a polynomial operator of degree r in D satisfying the relations

$$\begin{cases} Q_0(D) = 1, \\ Q_1(D) = 1 + \gamma_0 D, \\ Q_{r+1}(D) = (\beta_r + \gamma_r D)Q_r(D) + (1 - \beta_r)Q_{r-1}(D). \end{cases} \tag{2.14}$$

If $Q_n(D) = C_n(a,b,D)$, then (2.13) reduces to (2.7) for $r = n$, so that the iteration methods (2.4) and (2.11) will result in the same sequence of iterants u_k. The operators $C_r(a,b,D)$ satisfy the relations

$$\begin{cases} C_0(a,b,D) = 1, \\ C_1(a,b,D) = 1 + \dfrac{2}{b-a} D, \\ C_{r+1}(a,b,D) = (2w_0 + \dfrac{4}{b-a} D) \dfrac{T_r(w_0)}{T_{r+1}(w_0)} C_r(a,b,D) \\ \qquad - \dfrac{T_{r-1}(w_0)}{T_{r+1}(w_0)} C_{r-1}(a,b,D), \end{cases} \tag{2.15}$$

where $w_0 = (b+a)/(b-a)$. Thus, if

$$(2.16)\quad \gamma_0 = \frac{2}{b-a}\ ,\ \beta_r = 2w_0\,\frac{T_r(w_0)}{T_{r+1}(w_0)}\ ,\ \gamma_r = \frac{4}{b-a}\,\frac{T_r(w_0)}{T_{r+1}(w_0)}\ ,$$

$$r = 1,\ 2,\ \ldots,\ n-1,$$

then the operators $Q_r(D)$ are identical to the Chebyshev operators $C_r(a,b,D)$ for $r = 0,\ 1,\ \ldots,\ n$.
As in the first order Richardson process we set $a = \delta_0$ and $b = \sigma(D)$ obtaining Richardson's method of second degree (compare Frank [1960] and Varga [1962], p. 137).

Note that in the second order process the intermediate results $u_k^{(r)}$ may also be used as approximations to the limit function u_∞, contrary to the first order process. On the other hand, the second order process requires twice as much storage as needed for the first order process.

2.3 The rate of convergence

In order to compare the convergence properties of different iterative processes, Young [1954] introduced the average rate of convergence of an iterative method. According to Young the average rate of convergence of the methods (2.4) and (2.11), (2.16) is defined by

$$(2.17)\qquad R(n,N) = -\frac{1}{nN}\ln\left\|C_n^N(a,b,D)\right\|,$$

provided that $0 \leq a \leq \delta_0$ and $b \geq \sigma(D)$.
We shall call an iterative method convergent when the average rate of convergence for nN iterations is positive and divergent when it is negative.

Theorem 2.1

Richardson's method is convergent if D has negative eigenvalues and if n and N are sufficiently large.

Proof

To simplify the formulae it will be assumed that

$$a << b,\ \exp(-2n\sqrt{\frac{a}{b}}) << \exp(2n\sqrt{\frac{a}{b}}). \tag{2.18}$$

Then, we derive from (2.9) and (2.17)

$$R(n,N) \sim \frac{1}{n} \ln[\cosh(2n\sqrt{\frac{a}{b}})] - \frac{1}{nN} \ln[\nu N^{p-1}] \tag{2.19}$$

$$\sim 2\sqrt{\frac{a}{b}} - \frac{1}{nN} \ln[\nu 2^N N^{p-1}].$$

If n and N are chosen sufficiently large the method becomes convergent. Since Richardson's method arises for $a = \delta_0$ and $b = \sigma(D)$, we have proved the convergence of Richardson's method.

We obtain for normal matrices D ($p = \nu = 1$) the familiar formula

$$R(n,N) \sim 2\sqrt{\frac{a}{b}} - \frac{1}{n} \ln 2. \tag{2.20}$$

Compare Forsythe and Wasow [1960], p. 231.

Since (2.20) holds for any positive integer N, R(n,N) may be maximized by choosing N = 1. For a given value of the total number of iterations nN this may lead to large values of n. Unfortunately, the numerical stability of the first order Richardson method depends strongly on the distribution of the relaxation parameters ω_r, particularly when n is large. Therefore, if one decides to apply the linear form of Richardson's method, one has to order the relaxation parameters very carefully. This problem was discussed by Young (cf. Young [1953] or Forsythe and Wasow [1960], p. 234). For the special arrangement of the relaxation parameters Young recommended, he got convergence for a case where $\sigma(D)/\delta_0 \sim 162$ and n = 40. The iteration process was repeated (N > 1) until the desired accuracy was obtained. However, this reduces the average rate of convergence, so that it should be desirable to use an arrangement which allows us to take N = 1.

In connection with this we remark that in van der Houwen [1967 c] an arrangement of ω_r may be found which proved to be numerically stable

for a case where $\sigma(D)/\delta_0 \sim 1296$ and $n = 80$.
Note that the second order Richardson method is numerically stable. This follows from the fact that the coefficients β_r and γ_r defined by (2.16) satisfy the stability conditions (2.12).

If D is a non-normal matrix, formula (2.19) holds for sufficiently large values of N. However, if the eigenfunctions of D are linearly independent, we may derive a lower bound for the average rate of convergence for nN iterations which holds for N = 1, 2, ... Let

$$v_0 = \sum_{j=1}^{m} c_j e_j, \tag{2.21}$$

where e_j is an eigenfunction corresponding to δ_j and c_j is a scalar. Further, let $\|v_0\| = 1$. Then we have

$$\|C_n^N(a,b,D)\| = \sup_{\|v_0\|=1} \|v_N\| \leq \sigma(C_n^N(a,b,D) \sup_{\|v_0\|=1} \sum_{j=1}^{m} |c_j|. \tag{2.22}$$

From (2.9), (2.17), (2.18) and (2.22) we find the inequality

$$R(n,N) \geq 2\sqrt{\frac{a}{b}} - \frac{1}{nN} \ln[2^N \sup_{\|v_0\|=1} \sum_{j=1}^{m} |c_j|], \quad N = 1, 2, \ldots . \tag{2.23}$$

This result proves that Richardson's method is also convergent when N = 1 and when n is sufficiently large, provided that the eigenfunctions of D are linearly independent.
Note that for relatively small values of nN the average rate of convergence for nN iterations depends strongly on the values of ν or $\sum_{j=1}^{m} |c_j|$. These values are related to the conditioning of the eigenfunctions of the operator D (cf. Varga [1962], p. 65). Therefore, if one is faced with an ill-conditioned set of eigenfunctions, it is recommended to remove the ill-conditioned components from the starting error v_0 (compare Coolen and van der Houwen [1968]).

Finally, we remark that, for a fixed value of N, R(n,N) is an increasing function of n bounded by

$$R(\infty,N) \sim 2\sqrt{\frac{a}{b}}. \tag{2.24}$$

R(∞,N) will be called the asymptotic rate of convergence.

3. Accelerating procedures

From (2.19) we see that Richardson's method converges very slowly for large values of the so-called P-condition number $P = \sigma(D)/\delta_0$ (see Forsythe and Wasow [1960], p. 227). Unfortunately, in iterative solution methods of elliptic boundary value problems one is nearly always faced with ill-conditioned matrices D, i.e. $P >> 1$. Therefore, it is desirable to construct accelerating procedures.

3.1 The reduction-elimination method

The essence of Richardson's method was the reduction of all eigenfunction components e_j of the initial error v_0 by applying N times the operator $C_n(a,b,D) = C_n(\delta_0,\sigma(D),D)$. This means that the eigenfunction components of v_0 are simultaneously reduced.
One point of departure in accelerating Richardson's method is to reduce the eigenfunction components in two phases. In the first phase the eigenfunctions e_j with $-b \leq \delta_j \leq -a$ and $a > \delta_0$ are reduced (reduction phase) and in the second phase the remaining components are eliminated (elimination phase). If D is not a normal matrix and has no independent set of eigenfunctions, this reduction-elimination method may be repeated N times where N is sufficiently large.

Let us represent the effect of the reduction-elimination method upon the error v_k by the formula

$$v_{k+1} = A(n,n^*)v_k, \tag{3.1}$$

where n and n^* are the numbers of iterations of the reduction and the elimination phase respectively. Then the average rate of convergence for $(n + n^*)N$ iterations is given by

$$R(n,n^*,N) = -\frac{1}{(n + n^*)N} \ln\|A^N(n,n^*)\|. \tag{3.2}$$

Further, let us assume that the reduction is achieved by the Chebyshev operator $C_n(a,b,D)$ and let

$$\sigma(A(n,n^*)) = \varepsilon_0\alpha_0, \tag{3.3}$$

where ε_0 is the contribution of the elimination operator and α_0 is the maximum norm of $C_n(a,b,\delta)$ over the interval $[-b,-a]$ (compare formula (3.33) of chapter II). Applying lemma 3.1 of chapter II we obtain

$$(3.2') \qquad R(n,n^*,N) \sim 2\sqrt{\frac{a}{b}} - \frac{1}{(n+n^*)N} \ln[\nu 2^N N^{p-1}] - \frac{2n^*\sqrt{\frac{a}{b}} + \ln \varepsilon_0}{n+n^*},$$

where $a > \delta_0$, $b \geq \sigma(D)$ and where it is assumed that (2.18) is satisfied. For the asymptotic rate of convergence $R(\infty,n^*,N)$ we again find $2\sqrt{a/b}$, but as $a > \delta_0$ this value is larger than the value obtained for Richardson's method. In practice, however, the effect of a finite value of n is to reduce the average rate of convergence below its asymptotic value. Therefore, it is important to construct for a given value of a/b elimination methods for which the value of $2n^*\sqrt{a/b} + \ln \varepsilon_0$ is as small as possible.

Shortley [1953] suggested an elimination method for the case where $v_k^{(n)}$ is known to be a linear combination of just the first two eigenfunctions of D. As will be shown in subsection 3.2, however, this method leads to large values of ε_0 which may reduce the asymptotic rate of convergence considerably in actual computation.

Two other methods were proposed by Stiefel. However, these methods turned out to be unsatisfactory when tried on a computer (cf. Frank [1960]).

In the following subsections we propose two elimination methods which were used successfully on a computer.

Throughout this section we shall assume that (2.18) is satisfied and that the eigenvalues of the eigenfunctions to be eliminated are known (methods to calculate these eigenvalues are given in section 4). Further, we drop the condition that all eigenvalues of D are negative, but we shall not allow more than a few positive eigenvalues.

3.2 Elimination methods of first degree

In our analysis of the rate of convergence of the reduction-elimination method we shall introduce a function $s(\delta)$ such that the values of $s(\delta)$ in $\delta = \delta_j$, $j = 1, 2, \ldots, m$ are the eigenvalues of $A(n,n^*)$. This function will be called the spectrum function of the reduction-elimination method. It turns out that $s(\delta)$ is a rapidly oscillating function of δ. For this reason it is convenient to use the envelope of $s(\delta)$ rather than $s(\delta)$ itself. We introduce a function $\varepsilon(\delta)$ such that

$$|s(\delta)| \leq \varepsilon(\delta)\alpha_0. \tag{3.4}$$

For $\varepsilon(\delta)\alpha_0$ the envelope of $|s(\delta)|$ may be chosen.

We now define the reduction-elimination method of first degree by the scheme

$$\begin{cases} u_k^{(n)} = C_n(a,b,D)u_k + h_0^{(n)}, \\ u_k^{(n+j)} = (1 - \delta_j^{-1}D)u_k^{(n+j-1)} - \delta_j^{-1}g_0, \quad j = 1,2,\ldots,m_1, \\ u_{k+1} = u_k^{(n+m_1)}, \end{cases} \tag{3.5}$$

where $u_k^{(n)}$ is constructed by the first order scheme (2.4) and where the eigenvalues δ_j, $j = 1, 2, \ldots, m_1$ are outside the interval $[-b,-a]$.

The operator $A(n,n^*)$ is here given by

$$A(n,n^*) = A(n,m_1) = \prod_{j=1}^{m_1} (1 - \delta_j^{-1}D)C_n(a,b,D) \tag{3.6}$$

and the eigenvalues of $A(n,m_1)$ are given by the values of $s(\delta_j)$, where

$$s(\delta) = \prod_{j=1}^{m_1} (1 - \delta_j^{-1}\delta)C_n(a,b,\delta). \tag{3.7}$$

Since $s(\delta_j) = 0$ for $j = 1, 2, \ldots, m_1$, the corresponding eigenfunctions are eliminated from the error v_{k+1}. Assuming that $|\delta_j| \ll \sigma(D)$ for $j = 1, 2, \ldots, m_1$, we obtain from formula (3.2') the result

$$(3.8)\quad R(n_1,m_1,N) \sim 2\sqrt{\frac{a}{b}} - \frac{1}{(n+m_1)N}\ln[\nu 2^N N^{p-1}] - \frac{2m_1\sqrt{\frac{a}{b}} + \ln[(\sigma(D))^{m_1}/\prod|\delta_j|]}{n+m_1}.$$

The elimination method given by Shortley [1953] may be reduced to that represented by (3.5) when $m_1 = 2$ is substituted.

The operator $A(n,m_1)$ has the property that the eigenfunction components of $v_k^{(n)}$ corresponding to large negative eigenvalues of D are strongly amplified. If these components are weakly represented in the initial error v_0 and if they are not introduced by round-off errors, then this selective amplification is rather convenient and the actual rate of convergence of scheme (3.5) will be larger than predicted by formula (3.8). If, however, these components are strongly represented in v_0, or if they are introduced during the iteration process, which is very likely in first order methods considered here, then it is desirable to smooth the effect of the elimination operators upon the spectrum of $C_n(a,b,D)$. For that purpose we consider the more general scheme

$$(3.5')\quad \begin{cases} u_k^{(n)} = C_n(a,b,D)u_k + h_0^{(n)}, \\ u_{k+1} = E_n^*(D)u_k^{(n)} + h_0^{(n^*)}, \end{cases}$$

where $h_0^{(n^*)}$ is defined in the same manner as $h_0^{(n)}$ and where $E_n^*(D)$ is a polynomial operator of degree n^* in D which satisfies the conditions

$$(3.9)\qquad E_n^*(0) = 1,\ E_n^*(\delta_j) = 0,\ j = 1,2,\ldots,m_1.$$

Theorem 3.1

Let

$$a_j^* = \frac{b\left(\cos\left(\frac{\pi}{2n_j^*}\right) - 1\right) - 2\delta_j}{\cos\left(\frac{\pi}{2n_j^*}\right) + 1} \qquad (3.10)$$

Then the polynomial $C_{n_j}^*(a_j^*,b,\delta)$ satisfies the conditions

$$C_{n_j}^*(a_j^*,b,0) = 1, \quad C_{n_j}^*(a_j^*,b,\delta_j) = 0. \qquad (3.11)$$

Further, of all polynomials $Q_{n_j}^*(\delta)$ of degree n_j^* satisfying

$$Q_{n_j}^*(0) = 1, \quad Q_{n_j}^*(\delta_j) = 0, \qquad (3.12)$$

the polynomial $C_{n_j}^*(a_j^*,b,\delta)$ has the smallest maximum norm over the interval $[-b,-a]$, provided that

$$n_j^* \leq \frac{1}{4}\pi\sqrt{\frac{3b}{a+\delta_j}}. \qquad (3.13)$$

Proof

It is clear that $C_{n_j}^*(a_j,b,0) = 1$. Further, by substituting $n = n_j^*$, $a = a_j^*$ and $1 = 0$ into the formula for the zeroes of $C_n(a,b,\delta)$, which was given by formula (3.36) of chapter II, we find $z_0 = \delta_j$, i.e. δ_j is the first zero of $C_{n_j}^*(a_j^*,b,\delta)$.

In order to prove the minimax property, we assume the existence of a polynomial $Q_{n_j}^*(\delta)$ of degree n_j^* in δ satisfying (3.12) and the inequality

$$\|Q_{n_j}^*(\delta)\| < \|C_{n_j}^*(a_j^*,b,\delta)\|, \qquad (3.14)$$

where $\|\ \|$ denotes the maximum norm over the interval $[-b,-a]$.

Let us define the polynomial

$$V(\delta) = Q_{n_j}^*(\delta) - C_{n_j}^*(a_j^*, b, \delta).$$

$V(\delta)$ has positive values in those points of the interval $[-b,-a]$ where $C_{n_j}^*(a_j^*, b, \delta)$ assumes the value $-\|C_{n_j}^*(a_j^*, b, \delta)\|$, and negative values in the points where it assumes the value $+\|C_{n_j}^*(a_j^*, b, \delta)\|$.

Moving from $\delta = \delta_1$ in the negative direction along the δ-axis, the first extreme value $-\|C_{n_j}^*(a_j^*, b, \delta)\|$ is reached at the point

$$\delta = -\bar{\delta} = \frac{1}{2}(b - a_j^*)\cos\left(\frac{\pi}{n_j^*}\right) - \frac{1}{2}(b + a_j^*). \tag{3.15}$$

For $a \leq \bar{\delta}$ the polynomial $V(\delta)$ assumes n_j extreme values $\pm\|C_{n_j}^*(a_j^*, b, \delta)\|$. Since these values are alternatively negative and positive, $V(\delta)$ has at least $n_j^* - 1$ zeroes in $[-b,-a]$. In addition, $V(\delta)$ has two other zeroes in the points $\delta = 0$ and $\delta = \delta_j$. Hence $V(\delta)$ has at least $n_j^* + 1$ zeroes. On the other hand $V(\delta)$ is at most of degree n_j^*, implying at most n_j^* zeroes. This contradiction eliminates the existence of a polynomial $Q_{n_j}^*(\delta)$ satisfying (3.12) and (3.14) and therefore proves the minimax property of $C_{n_j}^*(a_j^*, b, \delta)$.

We still have to consider the condition $a \leq \bar{\delta}$. Substituting (3.10) into (3.15) we derive the inequality

$$2(b + \delta_j)\cos^2\left(\frac{\pi}{2n_j^*}\right) - (b - a)\cos\left(\frac{\pi}{2n_j^*}\right) - (b - a) \leq 0,$$

which is satisfied for

$$\frac{1 - \sqrt{1 + 8c_j}}{4c_j} \leq \cos\left(\frac{\pi}{2n_j^*}\right) \leq \frac{1 + \sqrt{1 + 8c_j}}{4c_j},$$

where

$$c_j = \frac{b + \delta_j}{b - a} = 1 + \frac{a + \delta_j}{b - a} \sim 1 + \frac{a + \delta_j}{b}.$$

This last inequality is approximately satisfied when

$$-\frac{1}{2} \leq \cos\left(\frac{\pi}{2n_j^*}\right) \leq 1 - \frac{2}{3}\frac{a+\delta_j}{b},$$

or equivalently when

$$1 \leq n_j^* \leq \frac{1}{4}\pi\sqrt{\frac{3b}{a+\delta_j}}.$$

For a given value of n_j^* satisfying (3.13), the operator $C_{n_j}*(a_j^*,b,D)$ defined by theorem 3.1 is the "best" operator to eliminate the eigenfunction corresponding to δ_j. If n_j^* does not satisfy (3.13), the operator $C_{n_j}*(a_j^*,b,D)$ still eliminates this eigenfunction, but the theorem does not indicate whether that operator is the "best" operator which could be used.

Let us define the operator $A(n,n^*)$ by the formulae

$$(3.16)\quad \begin{cases} A(n,n^*) = E_n*(D)C_n(a,b,D), \\ E_n*(D) = \prod_{j=1}^{m_1} C_{n_j}*(a_j^*,b,D), \\ n^* = \sum_{j=1}^{m_1} n_j^*. \end{cases}$$

Obviously the polynomial $E_n*(\delta)$ satisfies condition (3.9).
Note that for $n_j^* = 1$, $j = 1, 2, \ldots, m_1$, scheme (3.5'), (3.16) reduces to scheme (3.5).

Theorem 3.2

Let x_j, $j = 1, 2, \ldots, m_1$ be the solution of the equations

$$(3.17)\quad 2\sqrt{\frac{a}{b}} + \left[\arccos w_j - \frac{b\pi \sin\left(\frac{\pi}{2x_j}\right)}{2x_j(b+\delta_j)\sqrt{1-w_j^2}}\right] \operatorname{tg}(x_j \arccos w_j) = 0,$$

$$j = 1, 2, \ldots, m_1,$$

where

$$(3.18)\qquad w_j = \frac{b\cos(\frac{\pi}{2x_j}) - \delta_j}{b + \delta_j}\ .$$

Then, for fixed values of N and $n+n^*$ the average rate of convergence for $N(n + n^*)$ iterations of method (3.5'), (3.16) is maximized by $n_j^* = \text{entier}\,(x_j + \frac{1}{2})$, $j = 1, 2, \ldots, m_1$.

Proof

The rate of convergence of method (3.5'), (3.16) may be approximated by the formula (compare (3.8))

$$(3.8')\quad R(n,n^*,N) \sim 2\sqrt{\frac{a}{b}} - \frac{1}{(n+n^*)N}\ln[\nu 2^N N^{p-1}] - \frac{2n^*\sqrt{\frac{a}{b}} + \sum_{j=1}^{m_1}\ln\sigma(C_{n_j^*}(a_j^*,b,D))}{n + n^*}\ .$$

For constant values of N and $n + n^*$ this expression depends on the last term. $R(n,n^*,N)$ is maximized when this term is minimized. Let us consider the stationary values of $R(n,n^*,N)$. These are reached for those values of n_j^* which satisfy the equations

$$(3.19)\qquad \begin{cases} 2\sqrt{\frac{a}{b}} + \frac{d}{dx_j}\ln\sigma\,(C_{x_j}(a_j^*,b,D)) = 0,\ j = 1,2,\ldots,m_1, \\ n_j^* = \text{entier}(x_j + \frac{1}{2}). \end{cases}$$

From the definition of $C_{x_j}(a_j^*,b,D)$ we derive

$$(3.20)\quad \frac{d}{dx_j} \ln \sigma\,(C_{x_j}(a_j^*,b,D)) = -\frac{d}{dx_j} \ln \left| T_{x_j}\left(\frac{b + a_j^*}{b - a_j^*}\right)\right| =$$

$$= -\frac{d}{dx_j} \ln \left| T_{x_j}\left(\frac{b\cos\left(\frac{\pi}{2x_j}\right) - \delta_j}{b + \delta_j}\right)\right| =$$

$$= \left[\arccos w_j - \frac{b\pi \sin\left(\frac{\pi}{2x_j}\right)}{2x_j(b+\delta_j)\sqrt{1-w_j^2}}\right]$$

$$\operatorname{tg}(x_j \arccos w_j),$$

where w_j is defined by (3.18).

By means of this formula we have calculated the function $-\frac{d}{dx_j} \ln \sigma\,(C_{x_j}(a_j^*,b,D))$ in a set of points $(x_j, b/\delta_j)$. The results are listed in table 3.1.

For a constant value of b/δ_j this function decreases when x_j increases, so that $R(n,n^*,N)$ has negative derivatives with respect to the variables n_j^*, $j = 1, 2, \ldots, m_1$ in its stationary points. Therefore, the rate of convergence is maximized by the values of n_j^* determined by (3.19), i.e. determined by equation (3.17).

In applications, the optimal value of n_j^* can be derived from table 3.1 as soon as b/δ_j is known. Then, the value of $\sigma(C_{n_j^*}(a_j^*,b,D))$ follows from table 3.2 (compare section 5).

TABLE 3.1

Values of the function $-\frac{d}{dx_j} \ln \sigma \left(C_{x_j}(a_j^*, b, D)\right)$

b/δ_j \ x_j	1	2	3	4	5	6	7	8	9	10
-100	2.6	1.1	0.71	0.54	0.44	0.38	0.34	0.31	0.29	0.27
- 95	2.6	1.1	0.71	0.54	0.45	0.38	0.34	0.31	0.29	0.28
- 90	2.6	1.1	0.71	0.54	0.45	0.39	0.35	0.32	0.30	0.28
- 85	2.6	1.1	0.71	0.54	0.45	0.39	0.35	0.32	0.30	0.29
- 80	2.6	1.1	0.71	0.55	0.45	0.39	0.35	0.33	0.31	0.29
- 75	2.6	1.1	0.71	0.55	0.46	0.40	0.36	0.33	0.31	0.30
- 70	2.6	1.1	0.72	0.55	0.46	0.40	0.36	0.34	0.32	0.30
- 65	2.6	1.1	0.72	0.56	0.46	0.41	0.37	0.34	0.32	0.31
- 60	2.6	1.1	0.72	0.56	0.47	0.41	0.37	0.35	0.33	0.32
- 55	2.6	1.1	0.73	0.56	0.47	0.42	0.38	0.36	0.34	0.33
- 50	2.6	1.1	0.73	0.57	0.48	0.43	0.39	0.37	0.35	0.34
- 45	2.6	1.1	0.74	0.58	0.49	0.44	0.40	0.38	0.36	0.35
- 40	2.6	1.1	0.74	0.59	0.50	0.45	0.41	0.39	0.38	0.36
- 35	2.6	1.1	0.75	0.60	0.51	0.46	0.43	0.41	0.39	0.38
- 30	2.6	1.1	0.76	0.61	0.53	0.48	0.45	0.43	0.42	0.41
- 25	2.6	1.1	0.78	0.63	0.55	0.51	0.48	0.46	0.45	0.44
- 20	2.6	1.1	0.80	0.66	0.59	0.55	0.52	0.50	0.49	0.49
- 15	2.6	1.2	0.84	0.71	0.64	0.61	0.58	0.57	0.56	0.55
- 10	2.7	1.2	0.92	0.80	0.75	0.72	0.70	0.69	0.68	0.68
10	- 32	0.90	0.43	0.13	-0.16	-0.64	-3.7	1.3	0.37	-0.02
20	- 63	0.98	0.56	0.33	0.17	0.03	-0.13	-0.34	-0.83	-9.2
30	- 95	1.0	0.60	0.39	0.26	0.15	0.05	-0.05	-0.16	-0.34
40	-130	1.0	0.62	0.42	0.30	0.20	0.12	0.05	-0.02	-0.11
50	-160	1.0	0.63	0.44	0.32	0.23	0.16	0.10	0.04	-0.02
60	-190	1.0	0.64	0.45	0.33	0.25	0.18	0.13	0.08	0.03
70	-220	1.0	0.65	0.46	0.34	0.26	0.20	0.15	0.10	0.06
80	-250	1.0	0.65	0.47	0.35	0.27	0.21	0.16	0.12	0.08
90	-280	1.0	0.66	0.47	0.36	0.28	0.22	0.17	0.13	0.10
100	-310	1.0	0.66	0.47	0.36	0.28	0.23	0.18	0.14	0.11

TABLE 3.2

Values of the function $\sigma(C_{n_j}*(a_j^*, b, D))$

b/δ_j \ n_j	1	2	3	4	5	6	7	8	9	10
-100	99	20	8.6	4.6	2.8	1.9	1.3	0.95	0.70	0.53
- 95	94	19	8.1	4.4	2.7	1.8	1.2	0.89	0.66	0.50
- 90	89	18	7.7	4.1	2.5	1.7	1.2	0.83	0.61	0.46
- 85	84	17	7.2	3.9	2.4	1.6	1.1	0.77	0.57	0.42
- 80	79	16	6.8	3.6	2.2	1.5	1.0	0.72	0.52	0.39
- 75	74	15	6.3	3.4	2.1	1.3	0.93	0.66	0.48	0.35
- 70	69	14	5.9	3.1	1.9	1.2	0.85	0.60	0.43	0.32
- 65	64	13	5.4	2.9	1.7	1.1	0.77	0.54	0.39	0.28
- 60	59	12	5.0	2.7	1.6	1.0	0.69	0.48	0.35	0.25
- 55	54	11	4.5	2.4	1.4	0.92	0.62	0.43	0.30	0.22
- 50	49	9.9	4.1	2.2	1.3	0.82	0.54	0.37	0.26	0.19
- 45	44	8.9	3.7	1.9	1.1	0.71	0.47	0.32	0.22	0.15
- 40	39	7.8	3.2	1.7	0.97	0.61	0.40	0.26	0.18	0.12
- 35	34	6.8	2.8	1.4	0.82	0.51	0.32	0.21	0.14	0.10
- 30	29	5.8	2.3	1.2	0.67	0.40	0.25	0.16	0.11	0.07
- 25	24	4.7	1.9	0.94	0.52	0.31	0.19	0.12	0.08	0.05
- 20	19	3.7	1.4	0.70	0.38	0.21	0.13	0.08	0.05	0.03
- 15	14	2.7	1.0	0.47	0.24	0.13	0.07	0.04	0.02	0.01
- 10	9	1.6	0.58	0.25	0.11	0.06	0.03	0.01	0.01	0.00
10	11	2.6	1.4	1.0	1.0	1.5	6.5	2.4	1.2	1.0
20	21	4.6	2.2	1.4	1.1	1.0	1.1	1.3	2.3	22
30	31	6.7	3.1	1.9	1.4	1.1	1.0	1.0	1.1	1.4
40	41	8.7	4.0	2.4	1.7	1.3	1.1	1.0	1.0	1.1
50	51	11	4.9	2.9	2.0	1.5	1.2	1.1	1.0	1.0
60	61	13	5.7	3.4	2.3	1.7	1.4	1.2	1.1	1.0
70	71	15	6.6	3.9	2.6	1.9	1.5	1.3	1.1	1.0
80	81	17	7.5	4.3	2.9	2.1	1.7	1.4	1.2	1.1
90	91	19	8.4	4.8	3.2	2.3	1.8	1.5	1.3	1.2
100	100	21	9.3	5.3	3.5	2.6	2.0	1.6	1.4	1.2

3.3 Elimination methods of second degree

Since operators of type $C_n(a,b,D)$ can be generated by second order formulae, we may formulate scheme (3.5') in terms of second order equations. The average rate of convergence is not changed, but the method is less sensitive to round-off errors (see section 2.3).

In this subsection another second order elimination method will be discussed, which essentially uses the fact that the reduction process is of second degree.

Consider the scheme

$$(3.21)\qquad \begin{cases} u_k^{(n)} = C_n(a,b,D)u_k + h_0^{(n)}, \\ u_{k+1} = (\beta + \gamma D)u_k^{(n)} + (1-\beta)u_k^{(n-1)} + \gamma g_0, \end{cases}$$

where $u_k^{(n)}$ and $u_k^{(n-1)}$ are constructed with the second order scheme defined by equations (2.11) and (2.16). We will choose the parameters β and γ such that one or two eigenfunctions are eliminated from $v_k^{(n)}$.

Theorem 3.3

Let $q(\delta)$ be defined by

$$(3.22)\qquad q(\delta) = \frac{b + a + 2\delta + 2\sqrt{ab + \delta(b+a) + \delta^2}}{b + a + 2\sqrt{ab}}.$$

Then the first eigenfunction component e_1 is approximately eliminated from the error $v_k^{(n)}$ if

$$(3.23)\qquad \beta \sim \frac{\gamma\delta_1 q(\delta_1) + 1}{1 - q(\delta_1)},$$

and both the first and the second eigenfunctions are eliminated from $v_k^{(n)}$ if the additional relation

$$(3.24)\qquad \gamma \sim \frac{q(\delta_2) - q(\delta_1)}{q_1(\delta_1)q(\delta_2)(\delta_2 - \delta_1) + \delta_1 q(\delta_1) - \delta_2 q(\delta_2)}$$

is satisfied.

Proof

The operator associated with scheme (3.21) is given by

$$A(n,n^*) = (\beta + \gamma D)C_n(a,b,D) + (1 - \beta)C_{n-1}(a,b,D). \tag{3.25}$$

If $\gamma \neq 0$ we set $n^* = 1$ and if $\gamma = 0$ we set $n^* = 0$.
The eigenvalues of $A(n,n^*)$ are defined by $s(\delta_j)$ where

$$s(\delta) = (\beta + \gamma\delta)C_n(a,b,\delta) + (1 - \beta)C_{n-1}(a,b,\delta). \tag{3.26}$$

Let

$$w(\delta) = \frac{b + a + 2\delta}{b - a} \tag{3.27}$$

and let $\delta > -a$. Then it can be derived from the definition of $C_n(a,b,\delta)$ that

$$C_n(a,b,\delta) \sim \left[\frac{w(\delta) + \sqrt{w^2(\delta) - 1}}{w(0) + \sqrt{w^2(0) - 1}}\right]^n \sim [q(\delta)]^n. \tag{3.28}$$

It is easily verified that (3.23), (3.26) and (3.28) result in $s(\delta_1) \sim 0$ and that (3.23), (3.24), (3.26) and (3.28) result in $s(\delta_1) \sim s(\delta_2) \sim 0$.
This proves the theorem.

Theorem 3.4

Let the function $\varepsilon(\delta)$ be defined by

$$\varepsilon(\delta) = \sqrt{c_1\delta^2 + c_2\delta + c_3}, \tag{3.29}$$

where

$$c_1 = \gamma(\gamma + 4(1 - \beta)\frac{\exp(2\sqrt{\frac{a}{b}})}{b - a}),$$

$$c_2 = \gamma(2(1-\beta)(b+a)\frac{\exp(2\sqrt{\frac{a}{b}})}{b - a} + 2\beta) + 4\beta(1-\beta)\frac{\exp(2\sqrt{\frac{a}{b}})}{b - a},$$

$$c_3 = \beta^2 + 2\beta(1-\beta)(b+a)\,\frac{\exp(2\sqrt{\frac{a}{b}})}{b-a} + (1-\beta)^2\exp(4\sqrt{\frac{a}{b}}).$$

Then, the spectrum function $s(\delta)$ of scheme (3.20) satisfies the inequality

(3.4) $$|s(\delta)| \leq \varepsilon(\delta)\alpha_0.$$

Proof

From (3.26) we have for $-b \leq \delta \leq -a$

(3.26') $$s(\delta) \sim [(\beta+\gamma\delta)\cos(n \arccos w(\delta)) + (1-\beta)\exp(-2\sqrt{\frac{a}{b}})\cos((n-1)\arccos w(\delta))]\alpha_0.$$

This expression may be written as

(3.26'') $$s(\delta) \sim [A\cos(n \arccos w(\delta)) + B \sin(n \arccos w(\delta))]\alpha_0,$$

where

$$A = \beta + \gamma\delta + (1-\beta)\exp(2\sqrt{\frac{a}{b}})w(\delta),$$

$$B = (1-\beta)\exp(2\sqrt{\frac{a}{b}})\sin(\arccos w(\delta)).$$

From this last relation we obtain the inequality

$$|s(\delta)| \leq \sqrt{A^2 + B^2}\cdot\alpha_0.$$

Substitution of A and B yields inequality (3.4) where $\varepsilon(\delta)$ is defined by (3.29).

From theorem 3.4 it follows that the quantity ε_0 defined by (3.3) satisfies the inequality

(3.30) $$\varepsilon_0 \leq \max_{-\sigma(D)\leq\delta\leq -a} \varepsilon(\delta) = \max_{-\sigma(D)\leq\delta\leq -a} \sqrt{c_1\delta_1^2 + c_2\delta + c_3}.$$

Therefore, as soon as the coefficients c_1, c_2 and c_3 are determined, the average rate of convergence may be evaluated by means of formula (3.2') and (3.30). An example of this method is given in section 5 of this chapter.

3.4 Extrapolation formula of Ljusternik

In this subsection we consider a special case of scheme (3.21), where $\gamma = 0$ and β is determined by (3.23). The elimination formula may be interpreted as an extrapolation formula and is related to the extrapolation formula used by Ljusternik to accelerate stationary processes (cf. Forsythe and Wasow [1960], p. 219).

From theorem 3.4 and formula (3.30) it follows that

$$\varepsilon_0 \leq \sqrt{-c_2\sigma(D) + c_3}, \tag{3.31}$$

provided that $\beta < 0$ or $\beta > 1$. For all practical purposes we have by (3.23) that $|\beta| >> 1$, so that neglecting the computational labour of the extrapolation process ($n^* = 0$) the following inequality is obtained

$$R(n,0,N) \geq 2\sqrt{\frac{a}{b}} - \frac{1}{nN}\ln[\nu 2^N N^{p-1}] - \frac{\ln[-c_2\sigma(D) + c_3]}{2n}. \tag{3.32}$$

As was already remarked in subsection 3.2, the rate of convergence may be improved by applying some smoothing operator.
Here, the effect of the operator $1 + \omega D$ will be considered. The function $\varepsilon(\delta)$ transforms into

$$\varepsilon(\delta) = |(1 + \omega\delta)|\sqrt{c_2\delta + c_3}. \tag{3.33}$$

The stationary values of $\varepsilon(\delta)$ are reached in the points $\delta = -1/\omega$ and $\delta = -1/3\omega - 2c_3/3c_2$ (see figure 3.1).

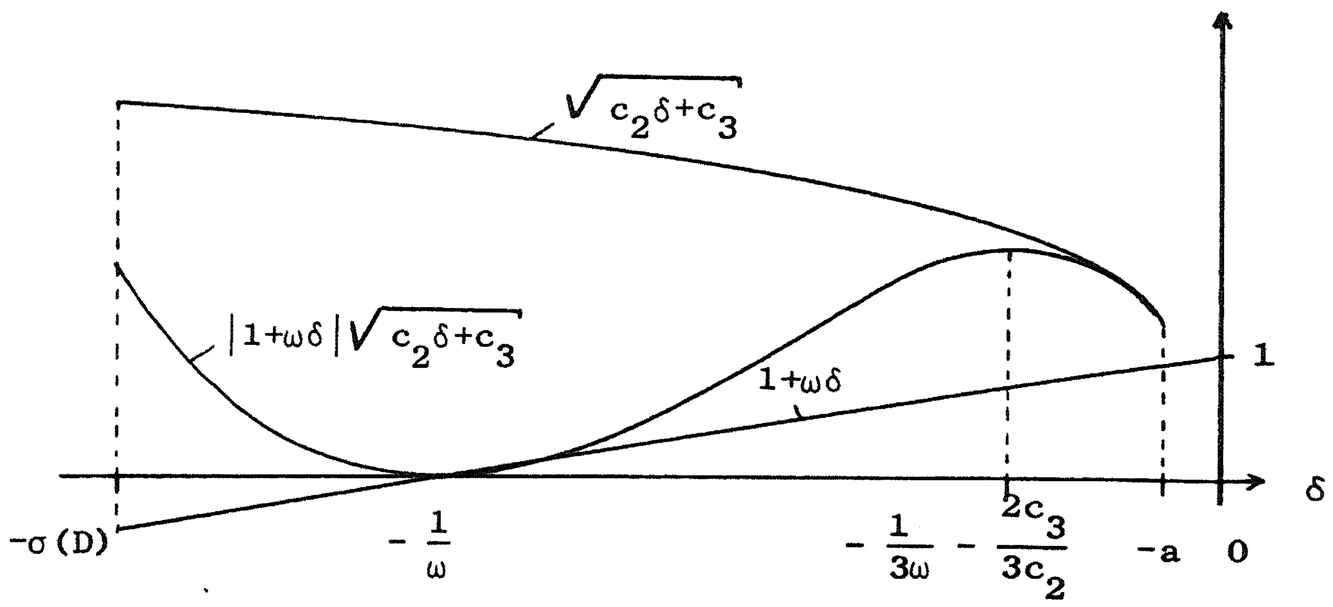

fig. 3.1

The optimal value of ω is determined by the equation

$$\varepsilon(-\sigma(D)) = \varepsilon(-\frac{1}{3\omega} - \frac{2c_3}{3c_2}). \tag{3.34}$$

Introducing the variables $x = \omega c_3/c_2$ and $c = c_2\sigma(D)/c_3$ this equation may be written as

$$(4 - 27(1-c)c^2x^3 + (54(1-c)c - 12)x^2 + (3 - 27(1-c))x - 4 = 0. \tag{3.34'}$$

For all practical purposes we have $c >> 1$, so that (3.34') reduces to

$$27(cx)^3 - 54(cx)^2 + 27cx - 4 = 0. \tag{3.34''}$$

This equation has one real root $cx = 4/3$, thus

$$\omega = \frac{cx}{\sigma(D)} = \frac{4}{3\sigma(D)} . \tag{3.35}$$

For ε_0 we find

$$\varepsilon_0 \leq \frac{1}{3}\sqrt{-c_2\sigma(D) + c_3}, \tag{3.36}$$

and for the rate of convergence we find

$$R(n,1,N) \geq 2\sqrt{\frac{a}{b}} - \frac{1}{(n+1)N}\ln[\nu 2^N N^{p-1}] - \frac{\ln[-c_2\sigma(D)+c_3]+4\sqrt{\frac{a}{b}} - \ln 9}{2(n+1)} . \tag{3.37}$$

4. Evaluation of the first eigenvalues of the operator D

In the preceding section we have assumed that the eigenvalues of D outside the interval [-b,-a] are known. This could be a disadvantage of the reduction-elimination method. However, it is possible to evaluate these eigenvalues during the iteration process. We shall derive formulae which yield approximate values of the eigenvalues corresponding to the dominant eigenfunctions without much computational labour. These formulae are applied at the end of the reduction phase of the first cycle. Throughout this section it will be assumed that D has a complete set of eigenfunctions.

4.1 General method

If n is sufficiently large, then the error $v_0^{(n)}$ is projected into the subspace spanned by the eigenfunctions $\{e_j\}_{j=1}^{m_1}$, i.e.

$$(4.1)\quad v_0^{(n)} = C_n(a,b,D)v_0 = C_n(a,b,D)\sum_{j=1}^{m} c_j e_j \sim \sum_{j=1}^{m_1} C_n(a,b,\delta_j)c_j e_j,$$

where c_j, $j = 1, 2, \ldots, m$ are the coefficients of the eigenfunction expansion of the initial error v_0.
From the definition of $C_n(a,b,D)$ we see that $C_n(a,b,\delta_j) >> C_n(a,b,\delta_{j+1})$ for $j = 1, 2, \ldots, m_1-1$. Hence it is expected that

$$(4.2)\qquad C_n(a,b,\delta_j)|c_j| >> C_n(a,b,\delta_{j+1})|c_{j+1}|,\ j = 1, 2, \ldots, m_1-1.$$

We shall assume that (4.2) is satisfied and that we have constructed a class of known functions w_1 which have the same property as $v_0^{(n)}$. Thus

$$(4.3)\qquad w_1 \sim \sum_{j=1}^{m_1} \Gamma_1(\delta_j)c_j e_j,$$

where $\Gamma_1(\delta)$ is a known function of δ such that

$$(4.4)\qquad \Gamma_1(\delta_j)|c_j| >> \Gamma_1(\delta_{j+1})|c_{j+1}|,\ j = 1, 2, \ldots, m_1-1.$$

For instance, we may define the class $\{w_1\}_{1=1}^{4}$, where

$$
(4.5)\qquad \begin{cases} w_1 = Du_0^{(n)} + g_0, \\ w_2 = Du_0^{(n-1)} + g_0, \\ w_3 = u_0^{(n)} - u_0^{(n-1)}, \\ w_4 = u_0^{(n-1)} - u_0^{(n-2)}. \end{cases}
$$

In order to calculate the eigenvalue δ_1 we form the two equations

$$
(4.6)\qquad \begin{cases} \|w_1\|^2 \sim \Gamma_1^2(\delta_1)c_1^2, \\ \|w_1'\|^2 \sim \Gamma_1'^2(\delta_1)c_1^2, \end{cases}
$$

where we have assumed that $\|e_1\| = 1$ and where $\|\ \|$ may be an arbitrary norm. Elimination of c_1^2 leads to the following equation for δ_1:

$$
(4.6')\qquad \det\begin{bmatrix} \|w_1\|^2 & \Gamma_1^2(\delta_1) \\ \|w_1'\|^2 & \Gamma_1'^2(\delta_1) \end{bmatrix} = 0.
$$

In those cases where the eigenfunctions e_j, $j = 1, 2, \ldots, m_1$ are successively eliminated, we repeat the preceding argument as soon as e_1 is eliminated from $v_0^{(n)}$. Then the eigenvalue δ_2 may be calculated, which enables us to remove the eigenfunction e_2, etc.

In scheme (3.21) - (3.24), however, it is necessary to know the values of δ_1 and δ_2 simultaneously. Let e_1 and e_2 be orthonormal. Then by eliminating c_1^2 and c_2^2 from the equations

$$
(4.7)\qquad \begin{cases} \|w_1\|^2 \sim \Gamma_1^2(\delta_1)c_1^2 + \Gamma_1^2(\delta_2)c_2^2, \\ \|w_1'\|^2 \sim \Gamma_1'^2(\delta_1)c_1^2 + \Gamma_1'^2(\delta_2)c_2^2, \\ \|w_1''\|^2 \sim \Gamma_1''^2(\delta_1)c_1^2 + \Gamma_1''^2(\delta_2)c_2^2, \end{cases}
$$

we obtain a second equation, namely

$$\text{(4.7')} \qquad \det \begin{bmatrix} \|w_1\|^2 & \Gamma_1^2(\delta_1) & \Gamma_1^2(\delta_2) \\ \|w_1'\|^2 & \Gamma_1^2{}'(\delta_1) & \Gamma_1^2{}'(\delta_2) \\ \|w_1''\|^2 & \Gamma_1^2{}''(\delta_1) & \Gamma_1^2{}''(\delta_2) \end{bmatrix} = 0.$$

Here, $\| \ \|$ denotes an inner product norm.
Equation (4.7') together with equation (4.6') determines the eigenvalues δ_2.

In the following subsections, formulae will be given which yield approximate values for the first eigenvalue δ_1.

4.2 The first order scheme

For the first order process the functions w_1 defined by (4.5) may be written as

$$\text{(4.8)} \quad \begin{cases} w_1 = Dv_0^{(n)} \sim \delta_1 C_n(a,b,\delta_1) c_1 e_1, \\ w_2 = Dv_0^{(n-1)} \sim \dfrac{\delta_1 C_n(a,b,\delta_1)}{1+\omega_{n-1}\delta_1} c_1 e_1, \\ w_3 = v_0^{(n)} - v_0^{(n-1)} = \omega_{n-1} Dv_0^{(n-1)} = \omega_{n-1} w_2, \\ w_4 = v_0^{(n-1)} - v_0^{(n-2)} = \omega_{n-2} Dv_0^{(n-2)} = \\ \qquad = \dfrac{\omega_{n-2}\delta_1 C_n(a,b,\delta_1)}{(1+\omega_{n-1}\delta_1)(1+\omega_{n-2}\delta_1)} c_1 e_1. \end{cases}$$

As an example, we apply formula (4.6') to the functions w_1 and w_2. We obtain for δ_1 the simple expression

$$\text{(4.9)} \qquad \delta_1 \sim \frac{1}{\omega_{n-1}} \left(\frac{\|w_1\|}{\|w_2\|} - 1 \right).$$

In the same manner other expressions may be obtained.
Note that the functions w_2 and w_3 are linearly dependent. Hence w_3 may be dropped in (4.8).

4.3 The second order scheme

In this case the functions w_l are approximated by

$$(4.10)\quad \begin{cases} w_1 \sim \delta_1 C_n(a,b,\delta_1)c_1e_1 \sim \delta_1[q(\delta_1)]^n c_1e_1, \\ w_2 \sim \delta_1 C_{n-1}(a,b,\delta_1)c_1e_1 \sim \delta_1[q(\delta_1)]^{n-1} c_1e_1, \\ w_3 \sim (C_n(a,b,\delta_1) - C_{n-1}(a,b,\delta_1))c_1e_1 \sim (q(\delta_1)-1)[q(\delta_1)]^{n-1} c_1e_1, \\ w_4 \sim (C_{n-1}(a,b,\delta_1) - C_{n-2}(a,b,\delta_1))c_1e_1 \sim (q(\delta_1)-1)[q(\delta_1)]^{n-2} c_1e_1, \end{cases}$$

where the function $q(\delta)$ is defined by (3.22).

Here, the functions w_l are linearly independent.

Let us apply formula (4.6') to the functions w_1 and w_2. Then we find

$$(4.11)\qquad q(\delta_1) \sim \frac{\|w_1\|}{\|w_2\|} .$$

This may be written as

$$(4.12)\qquad \delta_1 \sim \frac{1}{4}\left[(\sqrt{a}+\sqrt{b})^2 \frac{\|w_1\|}{\|w_2\|} - 2(a+b) + (\sqrt{a}-\sqrt{b})^2 \frac{\|w_2\|}{\|w_1\|}\right].$$

The same result is obtained when w_1 and w_2 are replaced by w_3 and w_4, respectively. For other expressions for δ_1 we refer to van der Houwen [1967 b].

Finally, we note that the formulae derived from (4.8) or (4.10) are valid for negative as well as positive eigenvalues δ_1.

5. The Dirichlet problem

In this section we shall apply the methods proposed in the preceding sections to a well-known problem in the theory of elliptic boundary value problems, namely the Dirichlet problem for the Poisson equation defined in a square of side π. This problem may be solved numerically

by the difference scheme discussed in example 3.3 of chapter II. We have shown that the difference scheme may be represented by a five-point formula with

$$\delta_0 \sim 2 \ , \quad \sigma(D) \sim 4\xi^{-2}, \tag{5.1}$$

where ξ is the mesh length along both the x and the y axis. The P-condition number of this scheme is given by $2\xi^{-2}$. We recall that the P-condition number of the scheme commonly used to solve this problem is twice as large.

Further, it may be deduced from formula (3.41) of chapter II that

$$\delta_1 \sim -2 \ , \quad \delta_2 \sim -5 \ , \quad \delta_3 \sim -8 \ , \quad \delta_4 \sim -10 \tag{5.2}$$

for small values of ξ.

We shall calculate the average rate of convergence for $\xi = \pi/20$.

First we consider Richardson's method. The rate of convergence is given by formula (2.19) with $a = -\delta_1 \sim 2$ and $b = \sigma(D) \sim 4\xi^{-2} \sim 162$. Since the operator D is represented by a symmetric matrix D, we have $p = \nu = 1$ and we may choose $N = 1$. This results in

$$R(n,1) \sim 0.222 - \frac{0.693}{n} \ . \tag{5.3}$$

Next we apply the reduction-elimination method with $a = 5$ and $b = 162$, where the remaining eigenfunction e_1 is eliminated by a linear operator, i.e. we apply scheme (3.5) with $m_1 = 1$. From formula (3.8) it follows that

$$R(n,1,1) \sim 0.351 - \frac{5.439}{n+1} \ . \tag{5.4}$$

Hence, after about 37 iterations this last method becomes faster than Richardson's method.

The average rate of convergence can be improved still more by replacing the linear elimination operator by the Chebyshev operator $C_{n_1}^{*}(a_1^{*},b,D)$ defined by the theorems 3.1 and 3.2. From table 3.1 we see that $n_1^{*} = 7$ is the optimal value of n_1^{*}. Further, from table 3.2 we have $\sigma(C_7(a_1^{*},b,D)) \sim 1$, so that applying formula (3.8') for $m_1 = 1$ we find

$$R(n,7,1) \sim 0.351 - \frac{3.150}{n+7} \qquad (5.5)$$

.

Note that the value $n_1 = 7$ satisfies the condition $n_1^* \leq \pi\sqrt{3b}/\sqrt{16(a+\delta_1)} \sim$ ~ 9.8. Thus the operator $C_7(a_1^*,b,D)$ is the "best" polynomial operator which can be applied (see formula (3.13)).

Finally, we consider the second order scheme (3.21) where β is defined by (3.23), and where we have substituted $a = 5$, $b = 162$ and $\delta_1 \sim -2$. For each value of δ the function $\varepsilon(\delta)$ defined by theorem 3.4 is a function of the parameter γ. Further, we have for large values of n

$$\varepsilon_0 \sim \max_{-162 \leq \delta \leq -5} \varepsilon(\delta).$$

In figure 5.1 we have plotted $10\varepsilon(-5)$, $\varepsilon(-162)$ and ε_0 as functions of γ for $-1/10 \leq \gamma \leq 1$.

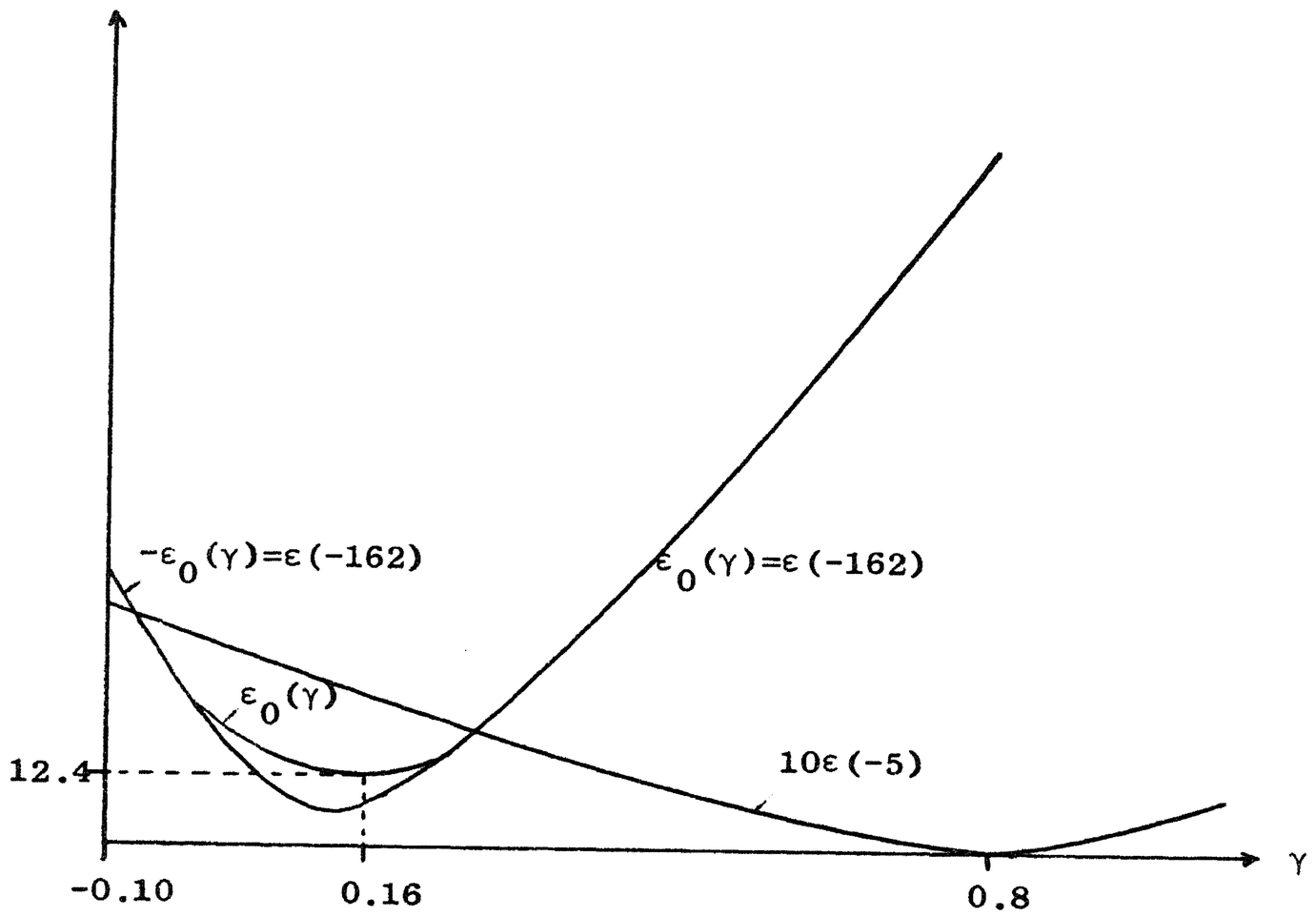

fig. 5.1

For $\gamma \sim .16$ the function $\varepsilon_0(\gamma)$ assumes a minimum value

$$\varepsilon_0(.16) \sim 12.4. \tag{5.6}$$

Substitution of (5.6) into (3.2') leads to

$$R(n,1,1) \sim 0.351 - \frac{3.562}{n+1}. \tag{5.7}$$

A slightly better rate of convergence is obtained when elimination is achieved by the smoothed extrapolation method discussed in subsection 3.4. The extrapolation formula contributes the number $\varepsilon_0(0) \sim 30.3$ which is reduced to $\varepsilon_0(0) \sim 10.1$ when the smoothing operator $1-4D/3\sigma(D)$ is applied. The average rate of convergence becomes

$$R(n,1,1) \sim 0.351 - \frac{3.357}{n+1}. \tag{5.8}$$

The accelerating methods discussed in this chapter were all tested on the EL X8 computer of the Mathematical Centre of Amsterdam. The numerical results were in good agreement with the theory.

In van der Houwen [1967 d] the polynomial elimination method was applied to the problem discussed in this section. The elimination of one and two eigenfunctions was treated. The corresponding eigenvalues were calculated by formulae indicated in section 4.

In Coolen and van der Houwen [1968] the same method was applied to a non-symmetrical matrix equation. A comparison was made with the successive overrelaxation method of Young.

Second order elimination methods were tested in van der Houwen [1968] for both the symmetrical and the non-symmetrical problem.

REFERENCES

O'Brien, G.G., Hyman, M.A., Kaplan, S. [1951] — A study of the numerical solution of partial differential equations. J. Math. Phys. 29, 223-251.

Cesari, L. [1959] — Asymptotic behaviour and stability problems in ordinary differential equations. Springer Verlag, Berlin.

Coolen, T.M.T., Houwen, P.J. van der [1968] — On the acceleration of Richardson's method IV. A non-symmetrical case. Report TW, Math. Centre, Amsterdam (to appear).

Courant, R., Friedrichs, K.O., Lewy, H. [1928] — Über die partiellen Differenzengleichungen der mathematischen Physik. Math. Ann. 100, 32-74.

Dantzig, D. van, Lauwerier, H.A. [1960 a] — General considerations concerning the hydrodynamical problem of the motion of the North Sea. The North Sea Problem I. Proc. Kon. Ned. Ak. v. Wetensch. A63, 170-180.

Dantzig, D. van, Lauwerier, H.A. [1960 b] — Free oscillations of a rotating rectangular sea. The North Sea Problem IV. Proc. Kon. Ned. Ak. v. Wetensch. A63, 339-354.

Du Fort, E.C., Frankel, S.P. [1953] — Stability conditions in the numerical treatment of parabolic differential equations. Math. Tables Aids Comput. 7, 135-152.

Esch, R.E. [1960] — A necessary and sufficient condition for stability of partial difference problems. J. Assoc. Comp. Mach. 7, 163-175.

Fischer, G. [1959] — Ein numerisches Verfahren zur Errechnung von Windstau und Gezeiten in Randmeeren. Tellus, 11, 60-76.

Flanders, D.A., Shortley, G. [1950] — Numerical determination of fundamental modes. J. Appl. Phys. 21, 1326-1332.

Forsythe, G.E., Wasow, W.R. [1960] — Finite difference methods for partial differential equations. John Wiley & Sons, Inc., New York.

Fox, L. [1962] — Numerical solutions of ordinary and partial differential equations. Pergamon Press, Oxford.

Frank, W. [1960] — Solution of linear systems by Richardson's method. J. Assoc. Comput. Mach. 7, 274-286.

Franklin, J.N. [1959] — Numerical stability in digital and analogue computation for diffusion problems. J. Math. Phys. 37, 305-315.

Godunov, J.K., Rjabenki, V.S. [1964] — Theory of difference schemes - an introduction. North-Holland Publishing Company, Amsterdam.

Hadamard, J. [1923] — Lectures on Cauchy's problem in linear partial differential equations. University Press, New Haven.

Harris, D.L., Jelesnianski, C.P. [1964] Some problems involved in the numerical solutions of tidal hydraulics equations. Monthly Weather Review, 92.

Houwen, P.J. van der [1966] On the stability of a difference scheme for the North Sea Problem. Report TW 100, Math. Centre, Amsterdam.

Houwen, P.J. van der [1967 a] Numerical treatment of the North Sea Problem without friction. Technical Note TN 47, Math. Centre, Amsterdam.

Houwen, P.J. van der [1967 b] On the acceleration of Richardson's method I. Theoretical part. Report TW 104, Math. Centre, Amsterdam.

Houwen, P.J. van der [1967 c] On the acceleration of Richardson's method II. Numerical aspects. Report TW 107, Math. Centre, Amsterdam.

Houwen, P.J. van der [1967 d] On the acceleration of Richardson's method III. Applications. Report TW 108, Math. Centre, Amsterdam.

Houwen, P.J. van der [1968] On the acceleration of Richardson's method V. A non-polynomial elimination method. Report TW, Math. Centre, Amsterdam (to appear).

John, F. [1952] On integration of parabolic equations by difference methods. Comm. Pure Appl. Math. 5, 155-211.

Kreiss, H.O. [1962] Über die Stabilitäts Definitionen für Differenzengleichungen die partielle Differentialgleichungen approximieren. Nordisk Tidsk. Informations-Behandling, 2, 153-181.

Lauwerier, H.A. [1960 a] Influence of a stationary windfield upon a bay with a uniform depth. The North Sea Problem II. Proc. Kon. Ned. Ak. v. Wetensch. A63, 266-278.

Lauwerier, H.A. [1960 b] Influence of a stationary windfield upon a bay with an exponentially increasing depth. The North Sea Problem III. Proc. Kon. Ned. Ak. v. Wetensch. A63, 279-290.

Lauwerier, H.A. [1960 c] Free motions of a rotating rectangular bay. The North Sea Problem V. Proc. Kon. Ned. Ak. v. Wetensch. A63, 423-438.

Lauwerier, H.A. [1961 a] Non-stationary windeffects in a rectangular bay. Theoretical part. The North Sea Problem VI. Proc. Kon. Ned. Ak. v. Wetensch. A64, 104-122.

Lauwerier, H.A. [1961 b] Non-stationary windeffects in a rectangular bay. Numerical part. The North Sea Problem VII. Proc. Kon. Ned. Ak. v. Wetensch. A64, 418-431.

Lauwerier, H.A., Damsté, B.R. [1963] — A numerical treatment. The North Sea Problem VIII. Proc. Kon. Ned. Ak. v. Wetensch. A66, 167-184.

Lavrientiev, M.M. [1967] — Some improperly posed problems of mathematical physics. Springer tracts in natural philosophy, vol. 11.

Lax, P.D., Richtmyer, R.D. [1956] — Survey of stability of linear finite difference equations. Comm. Pure Appl. Math. 9, 267-293.

Lax, P.D., Wendroff, B. [1960] — Systems of conservation laws. Comm. Pure Appl. Math. 13, 217-237.

Leendertse, J.J. [1967] — Aspects of a computational model for long-period waterwave propagation. Rand Memorandum RM-5294-PR. The Rand Corporation, Santa Monica, California.

Richtmyer, R.D. [1957] — Difference methods for initial value problems. Interscience Publishers Inc., New York.

Rjabenki, V.S., Filippov, A.F. [1960] — Über die Stabilität von Differenzen-gleichungen. Deutscher Verlag der Wissenschaften, Berlin.

Saul'yev, V.K. [1964] — Integration of equations of parabolic type by the method of nets. Pergamon Press, New York.

Shortley, G. [1953] — Use of Tschebyscheff-polynomial operators in the numerical solution of boundary value problems. J. Appl. Phys. 24, 392-396.

Strang, G. [1964] Accurate partial difference methods II. Non-linear problems. Numer. Math. 6, 37-46.

Todd, J. [1956] A direct approach to the problem of stability in the numerical solution of partial differential equations. Comm. Pure Appl. Math. 9, 597-619.

Varga. R.S. [1962] Matrix iterative analysis. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Veltkamp, G.W. [1960] Spectral properties of Hilbert space operators associated with tidal motions. Thesis, Rijksuniversiteit te Utrecht.

Young, D. [1953] On Richardson's method for solving linear systems with positive definite matrices. J. Math. Phys. 32, 243-255.

Young, D. [1954] Iterative methods for solving partial difference equations of elliptic type. Trans.Amer. Soc. 76, 92-111.

Yuan'Chzhao-Din [1958] Some difference schemes for the solution of the first boundary value problem for linear differential equations with partial derivatives. Thesis, Moscow State University.

INDEX