MC SYLLABUS 44

COLLOQUIUM NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

J.G. VERWER (ed.)

MATHEMATISCH CENTRUM AMSTERDAM 1980

1980 Mathematics subject classification: 65M, 65N, 65P05

ACM-Computing Reviews-category: 5.17

ISBN 90 6196 205 6

PREFACE

This syllabus comprises the lectures presented at a Colloquium on the Numerical Solution of Partial Differential Equations held in 1980 and organized by the Numerical Mathematics Departments of the Catholic University of Nijmegen, The Mathematical Centre of Amsterdam, and the Technical University of Delft. All contributors are connected to one of these three departments. The Colloquium consisted of 6 monthly meetings, starting in January, of one day and was alternatively held in Delft, Nijmegen and Amsterdam. The Colloquium was attended by 55 people, all active within The Netherlands at universities or in industry. This great number of attendants has certainly contributed to the success of the Colloquium.

The main purpose of the Colloquium was to report on current research and to stimulate contacts and co-operation between those who are involved in the numerical solution of partial differential equations. The topics cover a fairly wide range: multi-grid methods, incomplete factorization and preconditioned iterative methods, various aspects of splitting methods for time dependent problems, shallow-water equations, Navier-Stokes equations, free boundary problems, and, finally, various topics in the numerical integration of parabolic equations.

Sincere thanks are due to all contributors for their efforts in preparing their manuscripts. Further I would like to thank all those who took part in the organization of the Colloquium and the realization of this volume.

Amsterdam, November 1980

J.G. Verwer

CONTENTS

1. O. Axelsson 1
Computational aspects in the numerical solution of parabolic problems by finite element methods.

2. C. Cuvelier 10
Survey of some methods for solving free boundary problems governed by partial differential equations.

3. I. Gustafsson 41
On modified incomplete factorization.

4. P.W. Hemker 59
Introduction to multigrid methods.
Bibliography.

5. P.J. van der Houwen 98
Multistep splitting methods for non-linear initial value problems.

6. P.J. van der Houwen*) 109
On the treatment of time-dependent boundary conditions in splitting methods for parabolic differential equations.

7. R.M.M. Mattheij 114
On the stability of a class of difference methods for boundary value problems of the heat equation with time dependent coefficients.

8. N. Praagman*) 128
A comparison of discretization methods that are used to solve the shallow water equations.

9. A. Segal 133
Discretization of the continuity equation for the solutions of the Navier-Stokes equations using the finite element method.

10. A. Segal 152
On upwind discretizations of the convection diffusion equation.

11. J.G. Verwer 175
On the iterated defect correction and the LOD-method for parabolic equations.

12. J.G. Verwer*) 192
A class of stabilized Runge-Kutta methods for semi-discrete parabolic equations.

---

*) Only a summary of the lecture is given.

CONTRIBUTORS

A.O.H. AXELSSON
I. GUSTAFSSON
R.M.M. MATTHEIJ

Vakgroep Numerieke Wiskunde
Mathematisch Instituut
Katholieke Universiteit Nijmegen
Toernooiveld
6525 ED Nijmegen
The Netherlands

P.J. VAN DER HOUWEN
P.W. HEMKER
J.G. VERWER

Afdeling Numerieke Wiskunde
Stichting Mathematisch Centrum
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

C. CUVELIER
N. PRAAGMAN
A. SEGAL

Vakgroep Numerieke Wiskunde
Onderafdeling der Wiskunde
Afdeling der Algemene Wetenschappen
Technische Hogeschool Delft
Julianalaan 132
2628 BL Delft
The Netherlands

COMPUTATIONAL ASPECTS IN THE NUMERICAL SOLUTION OF PARABOLIC PROBLEMS BY FINITE ELEMENT METHODS

O. AXELSSON

1. INTRODUCTION

Three questions related to the numerical solution of parabolic equations have been studied: i) the initial phase problem, ii) the number of iterations in the solution of algebraic equations with matrix of the form $M+k(1-\theta)K$, M mass matrix, K stiffness matrix and iii) the efficiency of different degrees of approximations.

Consider the nonlinear heat conduction problem

$$c(u)\frac{\partial u}{\partial t} = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(k(u)\frac{\partial u}{\partial x_i}) + q(u,x,t) \quad \forall x \in \Omega \subset \mathbb{R}^n, \quad t > 0 \tag{1.1}$$

where u may be the temperature, c heat capacity, k thermal conductivity, and q is the rate of internal heat generation. Let $u(x,0) = u_0(x) \forall x \in \Omega$ be the initial condition and let

$$u = \gamma, \quad \forall x \in \Gamma_1, \quad k(u)\frac{\partial u}{\partial \nu} + g(u,x) = 0, \quad \forall x \in \Gamma_2$$

where $\Gamma_1 \cap \Gamma_2 = \partial\Omega$, $(\Gamma_1 \cap \Gamma_2 = \emptyset)$ be boundary conditions. ν is the outward pointing normal to Γ_2. A typical example of g is the "fourth order power law"

$$g(u) = \sigma(u^4 - u_m^4)$$

where u_m is the temperature of the ambient medium (i.e. a gas) and σ is an emissivity constant.

Let (see e.g. [ČERMÁK, ZLÁMAL (1980)]

$$H = \phi(u) = \int_{u_R}^{u} c(s)ds$$

where u_R is an arbitrarily chosen reference temperature, denote the enthalphy. Let ψ denote the inverse function to ϕ, i.e. $u = \psi(H)$. Then

$$\frac{\partial H}{\partial t} = c(u)\frac{\partial u}{\partial t}, \qquad \nabla u = \frac{\partial \psi}{\partial H}\nabla H = \frac{1}{c(u)}\nabla H.$$

The variational formulation of (1.1) over $V \times \mathring{V}$ where

$$V = \{H \in H'(\Omega);\quad H = \int_{u_R}^{\gamma} c(s)ds \quad \forall x \in \Gamma_1\}$$

$$\mathring{V} = \{v \in H'(\Omega);\quad v = 0 \quad \forall x \in \Gamma_1\}$$

now takes the form

$$(1.2) \qquad (\frac{\partial H}{\partial t}, v) + (a(H)\nabla H, \nabla v) + \oint_{\Gamma_2} g(\psi(H), x)dx = (q(\psi(H), x, t) \quad \forall x \in \Omega,$$

$$t > 0,\ \forall H \in V,\ v \in \mathring{V}.$$

and $H = \int_{u_r}^{u_0} c(s)ds$ for $t = 0$. Here

$$(u,v) = \int_{\Omega} uvdx, \qquad a(H) = \frac{k(u)}{c(u)}, \qquad u = \psi(H).$$

Hence, we have a variational formulation of a problem on divergence form. For discretization error estimates of the Galerkin method applied on such problems, see [LUSKIN, (1979)] and [AXELSSON, (1977)].

Typically one gets

$$h|H-H_h|_1 + |H-H|_0 \le Ch^{r+1}, \qquad h \to 0,\ t \ge t_0 > 0$$

where r is the degree of the polynomials in the finite element basis functions, under the condition that u is smooth enough. In general, C depends on t_0. h is a stepsize parameter.

In this talk we shall concentrate on three aspects of (1.2), namely

a) Initial (or transient) phase problem:
The solution is not smooth initially because the initial function does not usually satisfy boundary conditions. For $t > 0$ however, there is an exponential smoothing. Can one construct a numerical method for which there is also some kind of smoothing without having to choose excessively small timesteps?

b) Algebraic problem:

Stability can only be present for large timesteps if we choose implicit methods. Can one solve the implicit, possibly nonlinear, systems at each timestep to a reasonable cost in comparison with the explicit systems?

c) Discretization error problem:

Increasing the order of approximation increases in particular the complexity of numerical integration. High order approximations are effective only as long as the solution is smooth enough. The condition number of the algebraic systems depends on the ratio timestep to spacestep and hence on the degree of approximation in time and in space. Which combinations seem to be most effective?

In order to simplify the study, we analyze in particular the application of the θ-method for discretization in time.

2. INITIAL PHASE PROBLEM: ERROR ESTIMATES FOR SMALL t

Consider for simplicity the model problem,

$$u_t = u_{xx}, \quad 0 < x < 1, \; t > 0$$

$$u(t,0) = u(t,1) = 0, \quad t > 0$$

$$u(0,x) = u^0(x), \quad 0 < x < 1.$$

The solution can be given as a Fourierseries,

$$u(t,x) = \sum_{j=1}^{\infty} e^{-\lambda_j t} u_j^0 \, v_j(x),$$

where

$$\lambda_j = (j\pi)^2, \quad v_j(x) = \sqrt{2} \sin \pi_j x, \quad j = 1,2,\ldots$$

are eigensolutions of

$$-u_{xx} = \lambda u, \quad 0 < x < 1$$

$$u(0) = u(1) = 0$$

and

$$u_j^0 = (u^0, v_j) = \int_0^1 u^0(x) v_j(x) \, dx$$

are the Fourier coefficients.

Note that fast eigenmodes are damped out very fast. In fact, for every $t > 0$, u is infinitely smooth.

Furthermore, by Parsevals formula,

$$\|u(t)\| \leq e^{-\lambda_1 t}\|u^0\|, \quad t > 0,$$

that is, $u(t) \to 0$ as $t \to \infty$.

By a f.e.m. discretization over a finite dimensional space $V_N = \mathrm{SPAN}\{\phi_1,\phi_2,\ldots,\phi_N\}$, $\phi_i \in \overset{\circ}{H}{}^1(\Omega)$, corresponding to a mesh parameter h we get a system of ordinary differential equations

$$M_h \frac{du_h}{dt} + K_h u_h = 0, \quad t > 0, \quad u_h(0) = u_h^0$$

where

$$(M_h)_{ij} = \int_\Omega \phi_i(x)\phi_j(x)dx, \quad (K_h)_{ij} = \int_\Omega \phi_j'\phi_j'dx$$

and u_h^0 is an approximation of u^0, for instance the L_2-projection of u^0 onto V_N or the interpolant of u^0.

In order to solve this system of differential equations we discretize in time using the θ-method (For notational simplicity we now drop the subscript h):

$$[M+k(1-\theta)K]U(t+k) = [M-k\theta K]U(t), \quad t = 0,k,2k,\ldots,U(0) = U_h^0$$

where θ is a real-valued parameter and

$$U = U(t) = U(t,\theta)$$

is the corresponding approximation.

In order to study the error $u(t) - U(t)$ due to the time-discretization, we expand $u(t)$ and $U(t)$ is discrete Fourierseries about the eigenvectors z_q of $\tilde{K} = M^{-\frac{1}{2}}KM^{-\frac{1}{2}}$ i.e. $\tilde{K}z_q = \lambda_q z_q$. Let $\rho = \rho(\theta,\lambda) = \frac{1-\theta\lambda}{1+(1-\theta)\lambda}$. One gets

$$U_0 = \sum_q c_q z_q, \quad c_q = u_0^t z_q$$

$$u(t) = \sum_{q=1}^{N} c_q e^{-\lambda_q t} z_q, \quad U(t) = \sum c_q \rho_q^{t/k} z_q, \quad \rho_q = \rho(\theta,k\lambda_q)$$

and the discretization error

$$E(jk) = u(jk) - U(jk) = \sum_q c_q[e^{-k\lambda_q} - \rho_q]c_{qj}z_q,$$

where

$$c_{qj} = \sum_{\ell=0}^{j-1} \exp[-\lambda_q k(j-1-\ell)]\rho_q^{\ell} = O(1), \quad k \to 0.$$

Results: (see [AXELSSON, STEIHAUG (1978) and AXELSSON (1978)].

In order to have ρ_q decrease fast enough θ should be well below $\frac{1}{2}$, say $\frac{1}{4}$, but when $t = jk$ is large, c_{qj} gets relatively smaller for larger values of q, that is the higher modes in the expansion of the error are damped relatively faster.

Note: For $\theta = \frac{1}{2}$ we get the trapezoidal or Crank-Nicolson method. In this case $\rho_q \to -1$ as $q \to \infty$, for all fixed $k \to 0$. Hence we cannot bound the error, nor the discrete solution by an expression like

$$\|U(t)\| \leq Ce^{-\alpha t}\|U(0)\|, \tag{2.1}$$

unless we let α *depend on* k *(or* h*)*, i.e. we have only conditional asymptotic stability.

But for $\theta \leq \frac{1}{2} - \zeta k$, $\zeta > 0$, (2.1) is valid with α independent on k. This is true also for more general nonlinear problems (see AXELSSON (1977)]).

For error estimates for general evolution equations, see e.g. [HELFRICH (1976)], one gets typically

$$\|u(t)-U(t)\| \leq Ch^2[\frac{1}{t}\|u_0\| + M + \sup_{0\leq s\leq t}\|f(s)\|]$$

where C depends on λ_1, the smallest eigenvalue of the evolution operator (in our problem L in $\frac{\partial u}{\partial t} = Lu$).

3. ON THE NUMERICAL SOLUTION OF THE NONLINEAR ALGEBRAIC SYSTEMS OF EQUATIONS

At each time step we get a nonlinear system on the form

$$\underset{\sim}{A}(\underset{\sim}{H})\underset{\sim}{H} = [M+(1-\theta)kK(\underset{\sim}{H}(t+k))]\underset{\sim}{H}(t+k) = \underset{\sim}{F}(t,\underset{\sim}{H}(t))$$

where $\underset{\sim}{H}$ is the vector of unknown nodel values of $\underset{\sim}{H}$ and $\underset{\sim}{F}$ denotes a vectorial function of already calculated (or known) values at previous times.

For mildly nonlinear problems, $\underset{\sim}{H}(t,k)$ in the matrix K can be extrapolated from calculated values of $\underset{\sim}{H}$ at t,t-k,..., or from a predictor-corrector method (see e.g. DOUGLAS, DUPONT [1970]). Čermák, Zlámal proposes the use of a Gauss-Seidel method and proves convergence thereof. This method seems to be attractive in particular when a lumping has been used in order to get a diagonal matrix M. Van der Houwen (these proceedings) proposes Newton and splitting methods.

In the case of a problem where a = a(|∇u|) (like magnetic flux problems) the variational method is equivalent to a functional minimization problem, and then the Ritz-Galerkin method can be applied. Newton-Kantorovich method is then equivalent to the solution of a sequence of minimization problems for quadratic functionals. These latter can effectively be solved by a preconditioned conjugate gradient method (see [AXELSSON, NAVERT (1977)] and [AXELSSON, STEIHAUG (1978)]). The matrices $K(\underset{\sim}{H}(t+h))$ are only needed in the calculation of residuals, and hence do not have to be assembled.

This means that the updating of K can in practice be made pointwise, multiplying fixed element matrices with an updated factor a(u). The preconditioning matrix can for instance be constructed from the Hessian of the functional, and does not have to be updated at every iteration step, or not even at every time-step.

Consider now the solution of linear systems of equations on the form

$$\tilde{M}\underset{\sim}{H} = \underset{\sim}{F}$$

where $\tilde{M} = M+k(1-\theta)K$.

As already mentioned this can be solved by the conjugate gradient method, and each iteration costs only one multiplication by $\tilde{M}$ and three inner products. The number of iterations needed to reach a relative accuracy of ε is bounded above by the smallest natural number ℓ such that

$$2\frac{\sigma^{\ell}}{1+\sigma^{2\ell}} \leq \varepsilon,$$

where

$$\sigma = [1-\sqrt{\mu_1/\mu_0}]/[1+\sqrt{\mu_1/\mu_0}].$$

μ_1,μ_0 are the extreme eigenvalues of $\tilde{M}$.

One finds easily that the spectral condition number μ_1/μ_0 of $\tilde{M}$ behaves like $0(1+kh^{-2})$, $k,h \to 0$. Let $r = (1-\theta)k/h^2$. Then there is a clustering of eigenvalues as $r \to 0$ but already when $r = 0(1)$, we have $\mu_1/\mu_0 = 0(1)$. Hence, for the choice $k = 0(h^2)$, $h \to 0$, the conjugate gradient method makes an implicit methods, like the θ-method for $\theta < 1$, effectively an *explicit* time-integration method, in the respect that the number of operations per time-step is only $0(N)$, where N is the number of unknowns.

Of course, an explicit method (i.e. $\theta = 1$) could be used for such small time steps (typically of $k \le \frac{1}{n} h^2$) without causing instabilities in linear problems, but it is not clear if this will also be the case for non-linear problems.

If $k = 0(h^{-1})$, the condition number is $0(h^{-1})$, $h \to 0$, but with some preconditionings one can easily get an effective spectral condition number of $\sqrt{0(k/h^{-2})} = 0(h^{-\frac{1}{2}})$, and hence a number of iterations $0(h^{-\frac{1}{4}})$, that is, a very slow growth, $h \to 0$.

4. TOTAL DISCRETIZATION ERROR

One can prove (see e.g. [AXELSSON (1977)] that for $t \ge t_0 > 0$ the total L_2-norm error of the discretization by the θ-method in time and the f.e.m. in space is

$$0(k^2) \int_\Omega u_{ttt}^2 d\Omega + 0(h^{r+1}) |u|_{r+1}^2$$

where r is the degree of the piecewise polynomial splines.

This assumes that $\theta = \frac{1}{2} - \zeta k$, $\zeta > 0$ a parameter. It is reasonable to choose k such that

$$k^2 = 0(h^{r+1})$$

i.e.

$$k = 0(h^{\frac{r+1}{2}}) = \begin{cases} 0(h), & r = 1 \\ 0(h^{3/2}), & r = 2 \\ 0(h^2), & r = 3 \end{cases}$$

because then both of the error terms decrease asymptotically with the same rate.

Since for $t \geq t_0 > 0$, we have smoothness of "arbitrary degree", at least in the linear problem with smooth source terms, we get full accuracy for any r.

Consider now the number of operations needed to get a total error $0(\varepsilon)$ at a fixed time $t = T$. Since the condition number of $M+(1-\theta)kK$ is $0[\min(1,k/h^2)]$, $h \to 0$, the number of iterations needed at each time step, using a c-g method with MIC-preconditioning, is

$$0(\min(1,k/h^2))^{\frac{1}{2}} = 0(h^{\min(0,\frac{\frac{r+1}{2}-2}{4})}) = 0(h^{\min(0,\frac{r-3}{8})}).$$

The number of time steps is $0(k^{-1}) = 0(h^{-\frac{r+1}{2}})$ and the number of operations per iteration is $0(h^{-n})$, where n is the space-dimension. Hence, the total number of operations to reach an approximation at $t = T$ with error $0(\varepsilon)$, is

$$0(h^{\min(0,\frac{r-3}{8})-\frac{r+1}{2}-n}) = 0(h^{\min(-\frac{r+1}{2},-\frac{3r+7}{8})-n}), \quad h \to 0.$$

$$= \begin{cases} 0(h^{-\frac{3r+7}{8}-n}), & 1 \leq r \leq 3 \\ 0(h^{-\frac{r+1}{2}-n}), & r \geq 4 \end{cases}$$

Since $h^{r+1} = \varepsilon$ we finally get the total number of operations

$$\begin{cases} 0(\varepsilon^{-\frac{1}{2}+\frac{r-3}{8(r+1)}-\frac{n}{r+1}}), & 1 \leq r \leq 3 \\ 0(\varepsilon^{-\frac{1}{2}-\frac{n}{r+1}}) & r \geq 4. \end{cases}$$

The factor $0(\varepsilon^{-\frac{1}{2}})$ is the number of timesteps, $0(h^{\min(0,\frac{r-3}{8(r+1)})})$ the number of iterations per timesteps and $0(h^{-\frac{n}{r+1}})$ the number of operations per iteration.

We realize that it pays off to use as high order approximations as possible (from a practical point of view) since we can then use much larger stepsizes h. The increased number of iterations when $r = 1$ (or 2) instead of $r \geq 3$ is minor however.

The conclusion is that, as long as the solution is smooth enough and

the shape of the boundary is such to allow us to easily implement higher order approximations, we shall use them in order to decrease the work per time step. This is at least the case away from the transient region. Within the transient region it may not payoff to use too high order of approximations, however.

REFERENCES

O. AXELSSON, *Error estimates for Galerkin methods for quasilinear parabolic and elliptic differential equations in divergence form.* Numer. Math. 28 (1977), 1-14.

O. AXELSSON, U. NÄVERT, *On a graphical package for nonlinear partial differential equation problems,* Proc. IFIP Congress 77 (Ed. B. Gilchrist), IFIP, North-Holland, 1977.

O. AXELSSON, *On some experiments with time discretizations,* Proceedings COMPUMAG Conference on the Computation of Magnetic Fields, Grenoble, 4-6 September, 1978, to appear.

O. AXELSSON, T. STEIHAUG, *Some computational aspects in the numerical solution of parabolic equations,* J. Comp. Appl. Math. 4 (1978), 129-142.

L. ČERMÁK, M. ZLÁMAL, *Transformation of dependent variables and the finite element solution of nonlinear evolution equations,* Int. J. Num. Math. in Engin. 15 (1980), 31-40.

J. DOUGLAS, T. DUPONT, *Galerkin methods for parabolic equations,* SIAM J. Numer. Anal. 7 (1970), 575-626.

J. DOUGLAS, T. DUPONT & R.E. EWING, *Incomplete iteration for time-stepping a Galerkin method for a quasilinear parabolic problem,* SIAM J. Numer. Anal. 16 (1979), 503-522.

H.P. HELFRICH, *Lokale Fehlerabschatzungen für das Galerkin verfahren zur Lösung von Evolutionsgleichungen,* Bonner Math. Schriften, 1976.

M. LUSKIN, *A Galerkin method for nonlinear parabolic equations with nonlinear boundary conditions,* SIAM J. Numer. Anal. 16 (1979), 284-299.

SURVEY OF SOME METHODS FOR SOLVING FREE BOUNDARY PROBLEMS GOVERNED BY PARTIAL DIFFERENTIAL EQUATIONS

C. CUVELIER

INTRODUCTION

In many cases problems in mechanics and physics can be described by partial differential equations for the unknown functions. In hydrodynamics these unknowns are, for instance, the velocity vector and the pressure distribution in the fluid. When there are additional geometrical unknowns we speak of free boundary problems.

We can distinguish between two types of free boundary problems, depending on whether the free boundary depends on time or not. Generally we define (cf. CRYER [10],[11]) free boundary problems to be boundary value problems involving (partial) differential equations on domains, parts of whose boundaries, the free boundaries, are unknown and must be determined as part of the solution. A moving boundary problem is defined to be an initial-value problem for (partial)-differential equations involving unknown free boundaries.

For an exhaustive study of practically all possible known situations where free boundaries do arise, and a thorough survey of all available methods, both in hydrodynamics and other fields of physics, we refer to CRYER [11]. A survey of free boundary problems in hydrodynamics is given in CUVELIER, PRAAGMAN, SEGAL [17].

A general method which can be applied to almost all free and moving boundary problems is the so-called trial free boundary method. This is a numerical method for solving free boundary problems in which the boundary is found by guessing the position of the boundary, solving a boundary value problem in the resulting region, and using this solution to find a new guess for the position of the boundary, the procedure being repeated until the desired accuracy has been attained. This trial free boundary method will be applied to the problem of fluid flow through a porous medium (see section I).

A powerful method in the study of free and moving boundary problems is

the theory of variational inequalities, which was mathematically founded by LIONS, STAMPACCHIA [36] and further developed by BREZIS [4], BREZIS, STAMPACCHIA [5], STAMPACCHIA [48] and LIONS [33],[35]. A formulation in terms of a variational inequality has the advantage that the free boundary is no longer an unknown of the problem. We shall explain this technique in the problem of hydrodynamic partial lubrication in a journal bearing which will be studied in section II.

A thorough exposé on applications of the theory to a variety of (non) linear problems in mechanics and physics is contained in the monograph of DUVAUT, LIONS [22].

An other way of eliminating the free boundary from the problem will be discussed by considering the problem of a plasma flow in a Tokomak machine (see section III). This leads to a non-linear optimization formulation of the problem which can be solved using techniques of mathematical programming.

Some free and moving boundary problems can be formulated directly in terms of variational inequalities. Others admit a variational inequality formulation after a non-trivial change of the unknown functions (cf. BAIOCCHI, COMINCIOLI, MAGENES, POZZI [2], DUVAUT [20],[21], FREMOND [27]). Unfortunately not all free and moving boundary problems can be formulated as variational inequalities, and other techniques, like for instance the trial free boundary method, should be used.

We will discuss two trial-type methods for solving free boundary problems of viscous fluid flow which is governed by the (Navier-) Stokes equations and where the surface tension on the free boundary has to be taken into account. The first method (section IV) is based on techniques of optimal control theory (or optimum design); the second method (section V) is based on a constructive existence and uniqueness result in weighted Hölder spaces. Both methods are based on the same principle. A shape of the free boundary is assigned and the flow field within that shape is calculated after disregarding one of the boundary conditions on the free boundary. Next a new meniscus shape is computed which satisfies as closely as possible the boundary condition that was relaxed. This procedure is iterated until convergence is attained.

We restric ourselves to stationary (i.e. time independent) free boundary problems, and we shall accentuate more the mathematical formulation of the free boundary problems than its numerical solution.

Our program is as follows:

I A free boundary problem in porous media flows - Trial free boundary method.

II A free boundary problem in hydrodynamic partial lubrication - variational inequality

III A free boundary problem connected with a plasma flow in a Tokomak machine - Variational formulation.

IV A capillary free boundary problem governed by the Stokes equations - An optimal control approach.

V A capillary free boundary problem governed by the Navier-Stokes equations - Existence and uniqueness.

I. A FREE BOUNDARY PROBLEM IN POROUS MEDIA FLOWS - TRIAL FREE BOUNDARY METHOD

In this section we will study a free boundary problem in porous media flow using the trial free boundary method.

The flow of a liquid or gas through a porous medium is a phenomenon important in soil science, geotechnical engineering, hydraulics, oil reservoir engineering, environmental engineering and other fields.

Porous media flows are subdivided into steady and transient flows and these again into confined or unconfined flows. When the flow takes place between impervious boundaries it is said to be confined. If, on the other hand, some boundaries are free boundaries it is unconfined. A further subdivision is sometimes necessary, according to whether the medium is saturated or unsaturated.

The starting point for most formulations is a relationship between the velocity and the hydraulic gradient. The following are commonly used:
Darcy's law

$$u = -k \operatorname{grad} \phi \tag{1.1}$$

where u = average seepage velocity, ϕ = piezometric (or hydraulic) head and k = coefficient of permeability. Darcy's law can in general be considered to be valid for creeping flows with very low values (nearly equal to zero) of the Reynolds number Re (pre-laminar regime). The flow behavior is essentially non-linear in the laminar regime ($0 < Re < 1$), and beyond a value of $Re = 1$ it may be necessary to use non-linear or non-Darcy laws. Two non-linear laws

are for instance

$$\frac{\partial \phi}{\partial s} = a\, u_s + b\, u_s^2 \quad \text{(Forchheimer's law)}$$

$$\frac{\partial \phi}{\partial s} = c\, u^m \quad \text{(Missbach's law)}$$

where u_s is the velocity in the direction s and where a, b, c and m are material constants.

For Darcy flows, the governing equation is obtained by substituting equation (1.1) into the unsteady continuity equation to obtain

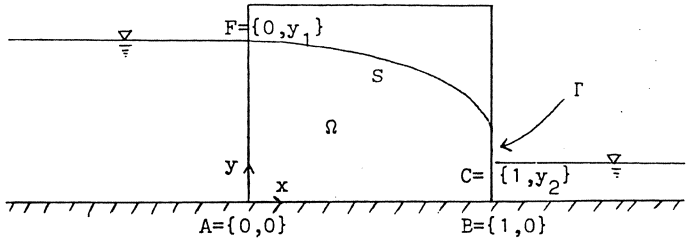
$$\text{(1.2)} \qquad \operatorname{div}(k \operatorname{grad} \phi) + q = \lambda \frac{\partial \phi}{\partial t}$$

where q is the externally added or withdrawn flow and λ the specific storage (confined flow) or effective porosity per unit aquifer thickness (unconfined flow). Notice that k, the permeability tensor, is a function of position.

It is seen from equation (1.2) that porous media flows are, in general, diffusion type problems.

For confined flows, the boundary conditions are generally of the Dirichlet or Neumann type and not especially difficult to handle. In unconfined flows, the location of the free surface is not known a priori (moving or free surface problem) and an iterative procedure is usually adopted. In the stationary case a position of the free surface is initially assumed and it is progressively modified until there is satisfaction of boundary condition of no flow across the surface and total head equalling elevation head.

We will describe here a particular free boundary problem using the trial free boundary method. We look for the bidimensional steady flow of an incompressible liquid through a homogeneuos (k = 1) earth dam which separates two reservoirs of different levels: the basis of the dam is considered to be horizontal and impervious, and we neglect capillarity



Let us denote by $D =]0,1[\times]0,y_1[$ the dam, by Ω the wetted region which is bounded by AF, S, Γ, CB, BA. The boundary S and the end point of Γ are unknown (free boundary problem).

Usually the problem is studied in the unknown piezometric head ϕ given by

$$\phi(x,y) = p(x,y) + y \tag{1.3}$$

where $p(x,y)$ denotes the unknown pressure in D (atmospheric pressure is measured by zero; specific weight of the liquid is supposed to be equal to 1). In terms of ϕ the problem can be formulated as

$$-\Delta\phi = 0 \quad \text{in } \Omega \tag{1.4}$$

$$\phi|_{AF} = y_1,\ \phi|_{BC} = y_2,\ \phi|_{\Gamma} \quad y,\ \frac{\partial\phi}{\partial y}\Big|_{AB} = 0 \tag{1.5}$$

$$\phi = y,\ \frac{\partial\phi}{\partial n} = 0 \quad \text{on } S \tag{1.6}$$

where $\frac{\partial}{\partial n}$ denotes the outward normal derivative on S.
It is a typical free boundary problem; we must solve the elliptic equation (1.4), with usual boundary conditions on the known part of $\partial\Omega \cap \partial D$ and two conditions on the free boundary S.

In the trial free boundary method, the position of the free boundary is estimated, say $S^{(0)}$. Let the nodalpoints of this estimate be given by

$y_i^{(0)}$, $i = 1,\ldots,N$. The first condition of (1.6) is disregarded and the problem for ϕ is solved by means of a finite element method using the boundary condition $\frac{\partial\phi}{\partial n} = 0$ on $S^{(0)}$. Then it must be verified whether along $S^{(0)}$ condition $\phi = y$ is satisfied or not. When this is not the case the position of the free boundary is modified, taking the new values of the y-coordinate equal to

$$y_i^{(m+1)} = y_i^{(m)} + \alpha(\phi_i^{(m)} - y_i^{(m)}) \qquad m = 0$$

with $i = 1,\ldots,N$ and $0 < \alpha < 2$, where the piezometric head $\phi_i^{(0)}$ is obtained from the initially estimated free boundary. This procedure is repeated until finally $|\phi-y|$ is everywhere on the free boundary smaller than a prescribed accuracy ε.

The discrete problem is best solved using a conjugate gradient method for $m = 0$ and an overrelaxation method for $m = 1,2,\ldots$ where the optimal overrelaxation parameter ω is calculated in the course of the iterations (cf. ENGERING [23]).

Obviously this method requires the numerical solution of a sequence of boundary value problems in a sequence of domains which vary at each step $m = 0,1,\ldots$. A finite element method is well suited for these type of problems, especially when an automatic meshgenerator is used to generate the internal nodal points from the boundary points.

The classical method, using this iterative approach, has been used by TAYLOR, BROWN [49], FINN [24], VOLKER [54], DESAI, JAVEL [18] and ENGERING [23].

Schematically the general description of the trial free boundary method is as follows:

step 0: Initial trial free boundary $S^{(0)}$

step 1: Given $S^{(k)}$ let Ω_k be the corresponding domain. Compute ϕ_k (approximation of ϕ) solution of the following problem

$$\left.\begin{array}{l} -\Delta\phi_k = 0 \quad \text{in } \Omega_k \\ \text{Boundary conditions on } \partial\Omega_k \backslash S^{(k)} \\ \text{Boundary condition } \dfrac{\partial\phi_k}{\partial n} = 0 \text{ on } S^{(k)} \end{array}\right\} \Rightarrow \exists!\,\phi_k$$

step 2: Given $S^{(k)}$ and ϕ_k, compute a new trial $S^{(k+1)}$ by requiring that $\phi_k - y$ should be approximately equal to zero on $S^{(k+1)}$; i.e. move the boundary $S^{(k)}$ to $S^{(k+1)}$.

Let us summarize some generalities concerning the trial free boundary method (cf. CRYER [10]).

Advantages of the trial free boundary method are:

(i) The method is, in principle, applicable to all free boundary problems and require no preliminary analysis.

(ii) All the computations are carried out in the physical domain so that intuition can be used and a feeling for the solution can develop as the the computation proceeds.

(iii) Standard implementation of step 1 be means of a finite element, a finite difference, a Galerkin or an integral method is possible.

(iv) There are many free boundary problems, such as free boundary problems in viscous fluid mechanics, for which trial free boundary methods are the only available methods.

As disadvantages of the trial free boundary method we could mention:

(i) It is difficult to define step 2 precisely. Each problem seems to require a different technique and while an individual research worker can accumulate experience, it is hard to transmit this knowledge to others. Many cases of non-convergence have been reported.

(ii) Error estimates and convergence proofs are lacking in general (a counter example is given in section V, where a convergence result does exist).

(iii) It is hard to achieve high accuracy and it is hard to estimate the error in ϕ_k and $S^{(k)}$. In some problems the shape of the free boundary is very sensitive to small deviations in the relaxed condition on the free boundary (for instance when the radius of curvature appears in one of the boundary conditions).

II. A FREE BOUNDARY PROBLEM IN HYDRODYNAMIC PARTIAL LUBRICATION - VARIATIONAL INEQUALITY.

It can be proved that under some conditions, for which we refer to CAMERON [6], the general equations describing the motion of a lubricant (for instance oil) reduce to the so-called Reynolds equation for the pressure p. This equation is a second order equation of elliptic type and for the unique solvability we have to impose boundary conditions of Dirichlet or Neumann type. For the numerical solution of this problem a standard finite element can be applied; we refer to REDDI [44], and ARGYRIS, SCHARPH [1].

This mathematically well posed boundary value problem may predict negative values for the pressure in some parts of the oil film. Physically this is impossible and we observe the formation of cavitation bubbles filled with the vapour phase of the oil. This phenomenon is called partial lubrication and one of the main problems is that the position of the cavities is not known a priori (free boundary problem). In a simplified model we can assume that on the interface of the liquid and the cavitation zone the following transition conditions hold

$$(2.1) \qquad p = 0 \quad \frac{\partial p}{\partial n} = 0$$

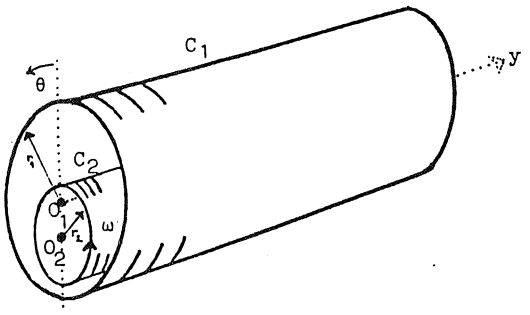
where p denotes the pressure in the lubricant, n the unit normal on the boundary of the cavitation region and where, for simplicity, the vapour pressure is taken equal to zero.

In the case of a journal bearing the unknown function is the pressure distribution p in a thin film of lubricant, with viscosity μ, contained in the narrow clearance between two cylinders C_1 (with radius r_1) and C_2 (with radius r_2) of equal length L, open at both sides at atmospheric pressure (which we put equal to zero). The axes of C_1 (with radius r_1) and C_2 (with radius r_2) of equal length L, open at both sides at atmospheric pressure (which we put equal to zero). The axes of C_1 and C_2 are parallel though distinct and the motion of C_2 is a rotation around its axis with constant angular velocity ω.

The Reynolds equation has the following form:

$$(2.2) \qquad -\frac{\partial}{\partial x_1}\left(h^3 \frac{\partial p}{\partial x_1}\right) - \frac{\partial}{\partial x_2}\left(h^3 \frac{\partial p}{\partial x_2}\right) = -\frac{6\mu UD}{c^2}\frac{\partial h}{\partial x_1}$$

where $\{x_1, x_2\} \in \Omega =]0,1[\times]0, \frac{L}{D}[$, $D = 2\pi r_2$, $h = 1+\varepsilon\cos 2\pi x_1$ and $U = \omega r_2$.



c = radial clearance $= r_1 - r_2$

ε = eccentricity $= \frac{e}{c}$, $0 \leq \varepsilon < 1$

e = $|O_1 - O_2|$ = distance between the central axes.

hc = film thickness.

$\theta = 2\pi x_1$

$y = Dx_2$

Taking now into account the possibility of cavitation, the free boundary problem can be stated as follows:

(2.3)

Find a function p (= pressure) and a subset Ω_c (= cavitation region) of Ω such that

$$-\frac{\partial}{\partial x_1}\left(h^3 \frac{\partial p}{\partial x_1}\right) - \frac{\partial}{\partial x_2}\left(h^3 \frac{\partial p}{\partial x_2}\right) = -\frac{6\mu UD}{c^2}\frac{\partial h}{\partial x_1} \quad \text{in } \Omega \setminus \Omega_c$$

$$p = 0 \quad \text{in } \Omega_c \qquad p > 0 \text{ in } \Omega \setminus \Omega_c$$

$$p = 0 \quad \text{on } \Gamma = \partial\Omega = \text{boundary of } \Omega$$

$$p = 0, \ \frac{\partial p}{\partial n} = 0 \quad \text{on } S = \partial\Omega_c = \text{interface of liquid and vapour}$$

The condition $p = 0$ at $x_1 = 0$ and $x_1 = 1$ indicates the presence of an oil groove at $x_1 = 0$. Boundary conditions of periodic type could also be taken. Setting $f = -\frac{6\mu UD}{c^2}\frac{\partial h}{\partial x_1}$, we make the following assumption

$$\Omega_c \subset \{x \in \Omega \mid f(x) \leq 0\}$$

which amounts to saying that the cavities occur in the divergent part of the journal bearing. With this assumption we can formulate problem (2.3) as a variational inequality (cf. CIMATTI [7], DUVAUT, LIONS [22]):

(2.4)

Find $p \in K = \{q \in H_0^1(\Omega) \mid q \geq 0 \text{ in } \Omega\}$ such that

$$a(p, q-p) \geq (f, q-p) \qquad \forall\, q \in K$$

where

$$H_0^1(\Omega) = \{q \in H^1(\Omega) \mid q|_\Gamma = 0 \text{ in the trace sense}\}$$

and

$$H^1(\Omega) = \{q \in L_2(\Omega) \mid \frac{\partial q}{\partial x_i} \in L_2(\Omega), \ \frac{\partial}{\partial x_i} \text{ in generalized sense}\}$$

is the Sobolev space of order 1. Furthermore

$$a(p,r) = \sum_{i=1}^{2} \int_\Omega h^3 \frac{\partial p}{\partial x_i}\frac{\partial r}{\partial x_i}\, dx$$

$$(f,r) = \int_\Omega f\, r\, dx \, .$$

Notice that in problem (2.3) the free boundary S is one of the unknowns, while in the variational inequality formulation (2.4) this geometrical unknown has been eliminated. Problem (2.4) can equivalently be formulated as a constraint optimization problem:

$$\left|\begin{array}{l} \text{Find } p \in K \text{ such that} \\ \\ J(p) = \inf_{q \in K} J(q) \end{array}\right.$$

where

$$J(q) = \tfrac{1}{2}a(q,q) - (f,q)$$

By standard methods we can prove existence and uniqueness of a solution $p \in K$. (cf. STAMPACCHIA [48], LIONS [35], CUVELIER [12]). Once the solution p of (2.4) or (2.5) is calculated, the free boundary is defined as the interface of the region where $p > 0$ and the region where $p = 0$.

Problem (2.4) is called a variational inequality and is connected with problems of constraint optimization, which explains the fact that a variational inequality can be considered as a generalization of the Euler equation in optimization theory.

Many generalizations of the notion of variational inequality are possible. For applications of variational inequalities (which include results on existence, uniqueness and regularity) to problems of physics and mechanics, we refer to DUVAUT, LIONS [22].

Concerning the numerical solution we notice that most of the stationary (elliptic) variational inequalities lead to one of the following cases:

- Minimization of a differentiable functional on a closed convex set of constraints.
- Minimization of a non-differentiable functional in the whole function space.

In order to solve these problems numerically we approximate the functional by a finite difference or finite element discretization method. For a general theory of numerical methods for variational inequalities we refer to GLOWINSKI, LIONS, TREMOLIERES [30], [31].

Specific difficulties, which essentially distinguish between the numerical analysis of (elliptic) variational inequalities and of elliptic equations are:

- The finite element or finite difference approximation of the set of constraints.
- Solution of the discrete problem, which is a non linear programming problem in a finite dimensional space

Concerning the solution of the discrete problem, it turns out that, if the set of constraints has a local character, the methods which are 'optimal' relative to computation time, computer storage and programming simplicity are of overrelaxation type, in particular for second order elliptical variational inequalities (cf. GLOWINSKI et al [30]. This method will be applied to the problem of hydrodynamic partial lubrication discussed in the beginning of this section.

Duality methods have a wider field of applicability, especially for fourth order elliptic problems and problems with non-differentiable functionals, but are more difficult to program and need more computer storage and computation time than the relaxation methods (at least if relaxation methods can be applied).

Penalization methods are in general difficult to program, are computer time consuming and need more computer storage than the preceeding methods.

An interesting method, which is in some sense related to duality and penalty, is the augmented Lagrangian method (cf. FORTIN [25], FORTIN, GLOWINSKI [26]). This method will be applied to the problems of section IV and V.

The variational inequality (2.4) (or (2.5)) will now be solved using an overrelaxation method.

We introduce a triangulation of the (fixed) domain Ω and we define a conforming finite element approximation V_h of $H_0^1(\Omega)$

$$V_h = \{v_h : \Omega \to \mathbb{R} \mid v_h \text{ piece-wise linear conforming on an element, } v_h \text{ equal to zero on } \Gamma\} \tag{2.6}$$



In this case the approximation K_h of the convex closed set K is easily defined

$$K_h = \{v_h \in V_h \mid v_h(M) \geq 0, \ \forall \text{ nodal points M of the triangulation}\} \tag{2.7}$$

The discrete problem is defined as follows:

$$(2.8)\qquad \left|\begin{array}{l} \text{Find } p_h \in K_h \text{ such that} \\ J(p_h) = \inf\limits_{q_h \in K_h} J(q_h) \end{array}\right.$$

It is proved in GLOWINSKI et al [30] that there exists a unique solution p_h of problem (2.8) and that $p_h \to p$ in $H_0^1(\Omega)$ as $h \to 0$. Since the space V_h is finite dimensional, the elements q_h of V_h are characterized by the values of q_h in the nodal points of the triangulation: $q_h = \{q_{h_1},\ldots,q_{h_N}\}$, $N = \dim(V_h)$. $J(q_h)$ can be written as $J(q_{h_1},\ldots,q_{h_N})$.

The relaxation method for solving (2.8) is now the following. We start with an arbitrary $p_h^o = \{p_{h_1}^o,\ldots,p_{h_N}^o\}$, $p_h^o \in K_h$. Once $p_h^1,\ldots,p_h^m$ are known, we calculate $p_{h_i}^{m+\frac{1}{2}}$ and $p_{h_i}^{m+1}$ subsequently for $i = 1,\ldots,N$ by

$$(2.9)\qquad \left|\begin{array}{l} p_{h_i}^{m+\frac{1}{2}} \in \mathbb{R} \text{ such that} \\ J(p_{h_1}^{m+1},\ldots,p_{h_{i-1}}^{m+1}, p_{h_i}^{m+\frac{1}{2}}, p_{h_{i+1}}^{m},\ldots,p_{h_N}^{m}) = \\ = \inf\limits_{q\in\mathbb{R}} J(p_{h_1}^{m+1},\ldots,p_{h_{i-1}}^{m+1}, q, p_{h_{i+1}}^{m},\ldots,p_{h_N}^{m}) \end{array}\right.$$

$$(2.10)\qquad p_{h_i}^{m+1} = \max(0,(1-\omega)p_{h_i}^{m} + \omega p_{h_i}^{m+\frac{1}{2}})$$

ω is called the relaxation parameter and it can be proved that the iterative method (2.9), (2.10) converges for $0 < \omega < 2$ (see GLOWINSKI et al [30], GLOWINSKI [29], CUVELIER [12]). An approximation of the optimal relaxation parameter can be achieved during the iteration process using a power technique.

Let us remark that (2.9) is equivalent with the following system of Galerkin equations:

$$(2.11)\qquad \sum_{\Omega_e} \sum_{j=1} \int_{\Omega_e} h^3 \frac{\partial \bar{p}_h}{\partial x_j} \frac{\partial \bar{q}_h}{\partial x_j}\, dx = \int_{\Omega_e} f\, \bar{q}_h\, dx$$

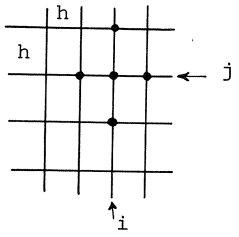
where the summation $\sum_{\Omega_e}$ is taken over all elements and where

$$\bar{p}_h \doteq \{p_{h_1}^{m+1}, \ldots, p_{h_{i-1}}^{m+1}, p_{h_i}^{m+\frac{1}{2}}, p_{h_{i+1}}^{m}, \ldots, p_{h_N}^{m}\}$$

$$\bar{q}_h = \{0, \ldots, 0, \underset{\uparrow}{1}, 0, \ldots, 0\}$$

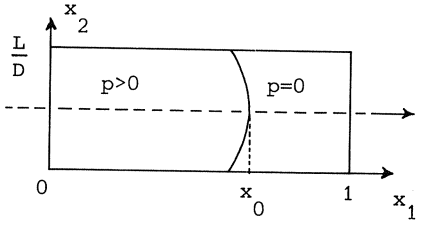
i-th coordinate

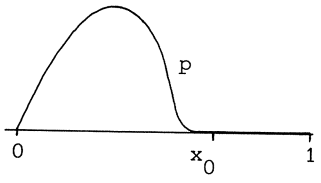
Notice that for $h \equiv 1$ and a regular grid, equation (2.11) takes the well known form



$$4p_{ij}^{n+\frac{1}{2}} - p_{i+1j}^{n} - p_{i-1j}^{n+1} - p_{ij+1}^{n} - p_{ij-1}^{n+1} = h^2 f_{ij}$$

The following figures give an idea of the behavior of the solution of (2.9), (2.10) (see CUVELIER [13]):



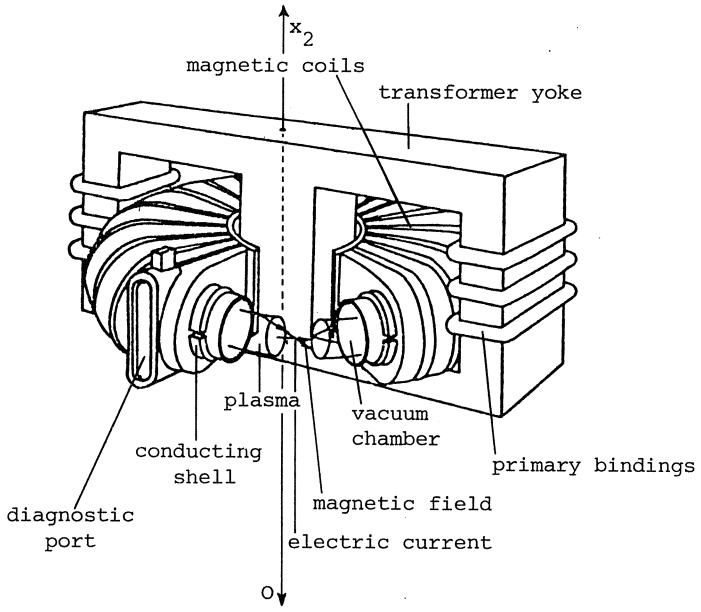


The region where $p > 0$ is the region of the lubricant, the cavitation region is that part of Ω where $p = 0$. The free surface S is the interface of these two regions.

III. A FREE BOUNDARY PROBLEM CONNECTED WITH A PLASMA FLOW IN A TOKOMAK MACHINE - VARIATIONAL FORMULATION

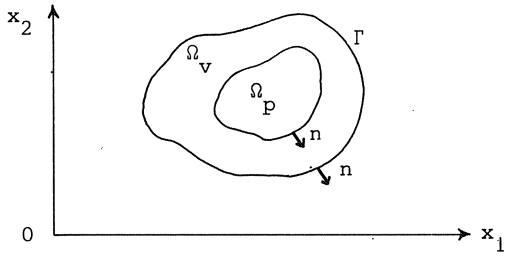
A tokomak machine is considered nowadays as one of the most promising devices for the realization of controlled fusion. Schematically the tokomak machine is constituted by a toroidal shell containing inside the plasma (of toroidal shape too) that one tries to heat intensively without introducing instabilities (see for instance MERCIER, ADAM, SOUBBARAMAYER, SOULE [38], TEMAM [50], LE MONDE [32]).

When a plasma is confined in such a configuration it plays the rôle of the secondary of a transformer, whose primary is explicitly shown in the next figure



By acting on the primary of the transformer, one can induce an intense electrical current inside the plasma, circulating along the meridian lines. This current heats the plasma and also produces a magnetic field; the magnetic lines twist around the surface of the plasma.

In the stationary axisymmetric case the shell and the plasma possess a toroidal symmetry around the axis Ox_2. The geometry in a hall plane limited by the axis Ox_2 is shown in the following figure.



Ω_v = vacuum
Ω_p = plasma
$\Omega = \Omega_v \cup \Omega_p \cup \Gamma_p$

In this stationary axisymmetric case, the equations, which are essentially the Maxwell equations for magnetohydrodynamic fluid flow, are:

$$(3.1)\qquad \left.\begin{array}{l} \operatorname{div} B = 0 \\ \operatorname{curl} B = 0 \qquad \text{(because } j = 0\text{)} \end{array}\right\} \qquad \text{in } \Omega_v$$

with B = magnetic induction, J = current density, and

$$(3.2)\qquad \left.\begin{array}{l} \operatorname{div} B = 0 \\ \operatorname{curl} B = \mu_0 J \\ \operatorname{grad} p = J \times B \end{array}\right\} \qquad \text{in } \Omega_p$$

where μ_0 = magnetic permeability, p = pressure. Notice that equation (3.2c) is the magneto hydrodynamic equation, which is in fact a simplification of the Navier-Stokes equations with an external magnetic force $J \times B$.

The boundary conditions are

$$(3.3)\qquad \begin{array}{l} B.n = 0 \qquad \text{on } \Gamma_p \text{ and } \Gamma \quad (n \text{ is unit outward normal}) \\ B.\tau \quad \text{is continuous across } \Gamma_p \quad (\tau \text{ is unit tangent}) \end{array}$$

Due to the fact that div B = 0, we can introduce a flux function u for which

$$B_1 = \frac{\partial u}{\partial x_1} \qquad B_2 = -\frac{\partial u}{\partial x_1} .$$

Next we suppose that $p = a_0 u^2$, $a_0 > 0$ (cf. TEMAM [51]), and the system of equations can be written as follows:

$$(3.4)\qquad \begin{array}{l} -\Delta u = 0 \qquad \text{in } \Omega_v \\ -\Delta u - \lambda u = 0 \quad \text{in } \Omega_p \qquad \lambda = 2\mu_0 a_0 \end{array}$$

For determining the boundary condition for u, we notice that u is defined up to an additive constant and that u is constant on Γ_p and on Γ. We choose the additive constant such that u vanishes on Γ_p. The boundary conditions are now (cf. MERCIER [37]):

$$
(3.5)\qquad
\left|\begin{array}{l}
u = 0 \\
\dfrac{\partial u}{\partial n} \text{ continuous across } \Gamma_p \\
u = \text{(unknown) constant on } \Gamma \\
\displaystyle\int_{\Gamma_p} \frac{\partial u}{\partial n}\, d\Gamma \quad \left(= \int_{\Gamma} \frac{\partial u}{\partial n}\, d\Gamma \right) = I \quad \text{(known positive constant)} \\
u \neq 0 \quad \text{in } \Omega_p .
\end{array}\right.
$$

Let us remark that the boundary Γ_p is unknown (free boundary).

It is possible to prove (see TEMAM [51]) that $u > 0$ in Ω_v and $u < 0$ in Ω_p, so that the problem can equivalently be defined as

$$
(3.6)\qquad
\left|\begin{array}{l}
\text{Find } u : \Omega \to \mathbb{R} \text{ such that} \\
-\Delta u + \lambda u^- = 0 \text{ in } \Omega,\ u^- = \max(-u,0) \\
u = \text{constant on } \Gamma \\
\displaystyle\int_{\Gamma} \frac{\partial u}{\partial n}\, d\Gamma = I
\end{array}\right.
$$

We see that the position of the free boundary is eliminated from the formulation of the problem. A consequence, however, is that a non-linearity has been introduced into the partial differential equation. Nevertheless such a non-linearity is easier to handle than a free boundary.

The variational formulation of problem (3.6) reads

$$
(3.7)\qquad
\left|\begin{array}{l}
\text{Find } u \in K = \{ v \in H^1(\Omega) \mid v = \text{const. on } \Gamma \} \equiv H_0^1(\Omega) \oplus \mathbb{R} \text{ such that} \\
a(u,v) = L(v) \qquad \forall\, v \in K
\end{array}\right.
$$

with $H_0^1(\Omega)$ and $H^1(\Omega)$ as defined in section II.

$$
a(u,v) = \sum_{i=1}^{2} \int_{\Omega} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}\, dx + \lambda \int_{\Omega} u^- v\, dx
$$

$$
L(v) = I.v\big|_{\Gamma}.
$$

For existence and uniqueness of a solution of problem (3.7), we refer to TEMAM [51], [52]. In terms of optimization theory the problem is formulated as follows:

$$(3.8)\qquad \left|\begin{array}{l} \text{Find } u \in K \text{ such that} \\ \\ J(u) = \inf_{v\in K} J(v) \end{array}\right.$$

with

$$J(v) = \tfrac{1}{2} \sum_{i=1}^{2} \int_{\Omega} \left(\frac{\partial v}{\partial x_i}\right)^2 dx - \frac{\lambda}{2} \int_{\Omega} (v^-)^2 \, dx - I\, v\big|_{\Gamma}$$

Notice that the free boundary Γ_p does not appear explicitly in the formulation (3.7) (or (3.8)) of the 'free boundary problem'. Once we have obtained the solution, the free boundary Γ_p is given by

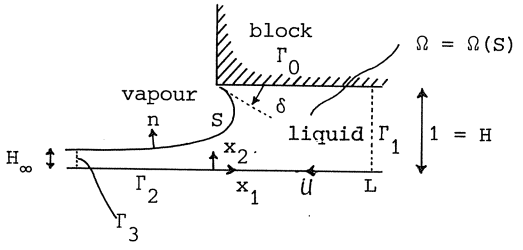
$$\Gamma_p = \{x \in \Omega \mid u(x) = 0\}$$

For numerical calculations we refer to SERMANGE [46]. For a study and a finite element analysis of more complicated (non standard) problems in plasma physics, we refer to MOSSINO [39], MOSSINO, TEMAM [40].

IV. A CAPILLARY FREE BOUNDARY PROBLEM GOVERNED BY THE STOKES EQUATIONS - AN OPTIMAL CONTROL APPROACH

The free boundary problem which will be discussed here is motivated by a phenomenon in lubrication theory, for which the Reynolds equation is not valid (see COYNE, ELROD [8], BIRKHOFF [3], DOWSON, GODET, TAYLOR [19], CUVELIER [14].

The physical situation is shown in the figure:



We see a plane sliding past a parallel semi-infinite block at constant velocity U. The liquid vapour interface of a ruptured thin liquid film separates at the corner of the block and is swept down out into the moving plane.

The object is to find the shape of the liquid vapour interface (free boundary). We assume that the liquid is incompressible (density ρ) and that the Stokes approximation is valid with constant viscosity μ and surface tension T. An external volume force is given by g (for instance gravity force).

In dimensionless form the problem can be formulated as follows.

Find a velocity vector $u = \{u_1, u_2\}$, a pressure field p and an interface S with 'wetting angle' δ such that

$$\left|\begin{array}{ll} -\Delta\, u + \operatorname{grad} p = G \ \text{ in } \Omega & \text{(Stokes equations)} \quad (4.1) \\ \operatorname{div} u = 0 \ \text{ in } \Omega & \text{(incompressibility)} \quad (4.2) \end{array}\right.$$

with $G = \dfrac{\rho H^2 g}{\mu U}$, together with the following boundary conditions

$$\left|\begin{array}{ll} u = 0 \ \text{ on } \Gamma_0 & (4.3) \\ u = \{ -[1-(4-6H_\infty)x_2-(6H_\infty-3)x_2^2],0\} \equiv \{h(x_2),0\} \ \text{ on } \Gamma_1 & (4.4) \\ \text{(asymptotic condition obtained from lubrication theory).} & \\ u = \{-1,0\} \ \text{ on } \Gamma_2 \cup \Gamma_3 & (4.5) \\ \sigma_\tau = 0 \quad \sigma_n = \dfrac{1}{NR} - p_a \ \text{ on S} & (4.6) \\ u_n = 0 \ \text{ on S} & (4.7) \end{array}\right.$$

where (4.7) is a condition on the Cauchy traction tensor, i.e. σ_τ and σ_n are the tangential, resp. normal stress, defined by

$$\sigma_\tau = \sigma_{ij} n_i \tau_j = \frac{\partial u_\tau}{\partial n} - \frac{\partial u_n}{\partial \tau}, \qquad \sigma_n = \sigma_{ij} n_i n_j = -p + 2\frac{\partial u_n}{\partial n}$$

$$\sigma_{ij} = -p\delta_{ij} + \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \quad \text{stress tensor,}$$

δ_{ij} = Kronecker symbol, with $n = \{n_1, n_2\}$ the unit outer normal on S and $\tau = \{\tau_1, \tau_2\}$ the unit tangent on S. R is the radius of curvature, p_a the vapour pressure, $N = \dfrac{\mu U}{T}$ = Saffman-Taylor number, $u_\tau = u \cdot \tau$ and $u_n = u.n$.

For fixed S, problem (4.1),...,(4.6) admits a unique weak solution $\{u,p\} = \{u(S),p(S)\}$. Next we try to select a boundary S_{opt} (belonging to some class $\mathcal{S}$) for which (4.7) is satisfied in the $L_2(S)$ sense. The problem of optimal control (or optimum design, since the control function is a part of the boundary) can be formulated as:

(4.8)
$$\left|\begin{array}{l} \text{Find } S_{opt} \in \mathcal{S} \text{ such that} \\ E(S_{opt}) = \inf_{S\in\mathcal{S}} E(S), \qquad E(S) = \int_S |u_n|^2 \, d\Gamma . \end{array}\right.$$

For solving this problem a descent method in a variable domain can be used. Suppose $S = \{x_2 = \phi(s) \,|\, s \in \mathcal{Q}\} \in \mathcal{S}$, s = curve length, and let $S_\alpha = \{x_2 = \phi(s) + \alpha(s)\, n(s) \,|\, s \in \mathcal{Q}\} \in \mathcal{S}$ be close to S where n(s) is the outward unit normal on S at s.

Using techniques of optimal control theory (see LIONS [33], PIRONNEAU [41], [42], [43]) we can calculate the first order variation $\delta E = E(S_\alpha) - E(S)$ of E. Introducing the adjoint state ψ by

(4.9)
$$\left|\begin{array}{l} -\Delta\psi + \operatorname{grad} q = 0, \quad \operatorname{div}\psi = 0 \quad \text{in } \Omega \\ \psi = 0 \quad \text{on } \Gamma_0 \cup \Gamma_2 \cup \Gamma_3 \\ \tilde{\sigma}_n = 0 \quad \psi_2 = 0 \quad \text{on } \Gamma_1 \\ \tilde{\sigma}_n = u_n \quad \tilde{\sigma}_\tau = 0 \quad \text{on } S \end{array}\right. \qquad (\tilde{\sigma} = \tilde{\sigma}(q,\psi) \text{ stress tensor})$$

it can be proved (see CUVELIER [14]) that

$$\delta E = 2\int_S \Big[\alpha\Big(\frac{\partial u_n}{\partial n} - \frac{u_n}{2R}\Big)u_n - \frac{1}{2}\sum_{i,j=1}^{2}\Big(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\Big)\Big(\frac{\partial \psi_i}{\partial x_j} + \frac{\partial \psi_j}{\partial x_i}\Big) + G\psi +$$

(4.10)
$$-p_a\Big(\frac{\partial \psi_n}{\partial \alpha} - \frac{\psi_n}{R}\Big) + \frac{1}{NR}\,\frac{\partial \psi_n}{\partial n} + \frac{1}{N}\,\frac{\partial^2 \phi_n}{\partial s^2}\Big]d\Gamma + o(\alpha,\dot{\alpha},\ddot{\alpha})$$

$$\simeq \frac{\partial E}{\partial \alpha}\cdot\alpha + o(\alpha,\dot{\alpha},\ddot{\alpha}).$$

The set of admissible boundaries is given by the property that the wetting angle is equal to δ. (In fact a spline function approximation was used).

The descent algorithm is now defined as follows:
Choose $S^o \in \mathcal{S}$. Suppose $S^1,\ldots,S^n$ are known. Solve $u(S^n)$ and $\psi(S^n)$, and calculate $\frac{\partial E}{\partial \alpha}$ by (4.10). Next we set

$$\alpha = -\frac{\partial E}{\partial \alpha}$$

and S^{n+1} can be defined by

$$S^{n+1} = \{x_2 = \phi(s) + \rho\ \alpha(s)\ n(s) \mid s \in \mathcal{Q}\}, \text{ for optimal } \rho > 0.$$

At each iteration step we have to solve two Stokes problems with boundary conditions of Neumann and Dirichlet type. Notice that the Neumann boundary condition is a relation between the velocity and the pressure.

In order to apply a finite element method for solving the Stokes problems, we give a weak formulation of problem (4.1),...,(4.6) (see CUVELIER [14]):

$$\text{(4.11)} \qquad \left| \begin{array}{l} \text{Find } u \in V \text{ such that} \\ \\ a(u,v) = (f,v)_{\mathbb{L}_2} + \int_S (\frac{1}{NR} - p_a) v_n \, d\Gamma \qquad \forall\, v \in V_o \end{array} \right.$$

with

$$a(u,v) = \frac{1}{2} \sum_{i,j=1}^{2} \int_\Omega (\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i})\, dx\ ,$$

$$(f,v)_{\mathbb{L}_2} = \sum_{i=1}^{2} \int_\Omega f_i v_i dx$$

$$V = \{v \in (H^1(\Omega))^2 \mid \text{div } v = 0,\ v = \text{given on } \Gamma_o \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$$

(see (4.3), (4.4), (4.5)}

$$V_0 = \{v \in (H^1(\Omega))^2 \mid \text{div } v = 0,\ v = 0 \text{ on } \Gamma_o \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\}.$$

Notice that this system of equations is the same as the equations of linear elasticity. Due to the continuity and the V ellipticity of the bilinear form a(.,.), which is a consequence of Korn's inequality (cf.

DUVAUT, LIONS [22]), we can prove existence and uniqueness of a solution u of problem (4.11) (cf. TEMAM [51]).

We now introduce the pressure p as a Lagrange multiplier and moreover we add a penalty term associated with the constraint 'div u = 0'. The penalized weak formulation is then:

$$
(4.12)\qquad \left|\begin{array}{l} \text{Find } u_\sigma \in \tilde{V} \text{ and } p_\sigma \text{ such that} \\ a(u_\sigma, v) + \sigma \int_\Omega \operatorname{div} u_\sigma \operatorname{div} v \, dx - \int_\Omega p_\sigma \operatorname{div} v \, dx = \\ = (f,v)_{\mathbb{L}_2} + \int_S \left(\frac{1}{NR} - p_a\right) v_n \, d\Gamma, \qquad \forall\, v \in \tilde{V}_o \end{array}\right.
$$

with σ large positive number ($\approx 10^6$)

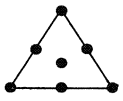
$$
\tilde{V} = \{v \in (H^1(\Omega))^2 \mid v = \text{given on } \Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3,
$$

(see (4.3), (4.4), (4.5)}

$$
\tilde{V}_o = \{v \in (H^1(\Omega))^2 \mid v = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\}.
$$

For solving problem (4.12) we define a triangulation of the domain Ω and we introduce a finite element approximation of the space $\tilde{V}$. Among other choices one can take (see GIRAULT, RAVIART [28], CROUZEIX, RAVIART [9]):

$\tilde{V}_h$ = approximation of $\tilde{V} = \{v_h : \Omega \to \mathbb{R}^2 \mid v_h$ extended quadratic conforming, v_h satisfying the essential boundard conditions}

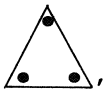


$\tilde{V}_{ho} = \tilde{V}_h$ + homogeneous essential boundary conditions.

The augmented Lagrangian method of Hestenes and Powell (see FORTIN [25], FORTIN, GLOWINSKI [26], SEGAL [45]) is defined as follows:

Choose p_h arbitrary in

$$
H_h = \{q_h : \Omega \to \mathbb{R} \mid q_h \text{ linear nonconforming, } \triangle, \int_\Omega q_h \, dx = 0\}.
$$



Once $u_h^1, p_h^1, \ldots, p_h^m$ have been calculated, u_h^{m+1} is given by

$$(4.13)\qquad \left|\begin{array}{l} u_h^{m+1} \in \tilde{V}_h \\ a(u_h^{m+1}, v_h) + \sigma_\Omega \int_\Omega \Pi\,(\mathrm{div}\, u_h^{m+1})\, \Pi(\mathrm{div}\, v_h)\, dx = (f, v_h)_{L_2} + \\ \qquad + \int_S \left(\frac{1}{NR} - p_a\right) v_{hn}\, d\Gamma + \int_\Omega p_h^m\, \mathrm{div}\, v_h\, dx \qquad \forall\, v_h \in \tilde{V}_{ho} \end{array}\right.$$

and $p_h^{m+1} \in H_h$ by

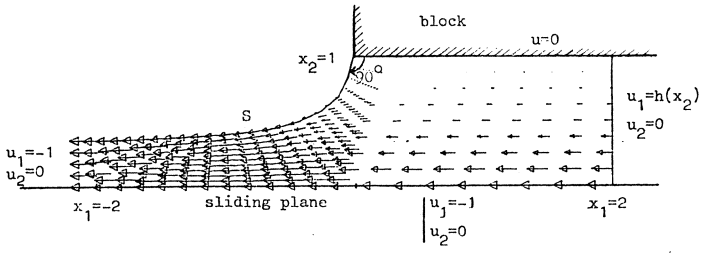
$$(4.14)\qquad \left| (p_h^{m+1} - p_h^m, q_h)_{L_2} - \sigma(\mathrm{div}\, u_h^{m+1}, q_h)_{L_2} = 0 \qquad \forall\, q_h \in H_h \right.$$

where Π denotes the orthogonal projection in $L_2(\Omega)$ onto H_h.

It can be proved that $\{u^m, p^m\} \to \{u(S), p(S)\}$, the unique weak solution of (4.1),...,(4.6). Solving the adjoint system in the same way, we can calculate the first order variation of the discrete analogue of the functional E. This quantity is used to determine the next approximation of the free boundary S, i.e. the discrete analogue of (4.10).

The finite element method is ideally suited for these type of problems. By assuming the shape of the boundary, solving the problem and then correcting the shape and repeating the calculation, the location of the free boundary can be determined. The advantage of the finite element method in handling diverse geometries is usefull here, especially when an automatic triangulation subroutine is used to generate the internal nodal points from the boundary points.

When we take G = 0, the characteristic number of the problem is the Saffman-Taylor number $N = \frac{\mu U}{T}$. For the special case N = 1.0, $\delta = 90^o$, L = 2 the value of the $E(S^n)$ diminishes in 15 iterations from $E(S^o) = 0.2486 * 10^{-2}$ to $E(S^{15}) = 0.1177 * 10^{-2}$. The velocity field of the 15-th approximation is given in the next figure: (for other results we refer to CUVELIER [14], [15]).



V. A CAPILLARY FREE BOUNDARY PROBLEM GOVERNED BY THE NAVIER-STOKES EQUATIONS - EXISTENCE AND UNIQUENESS

Let us consider the following problem

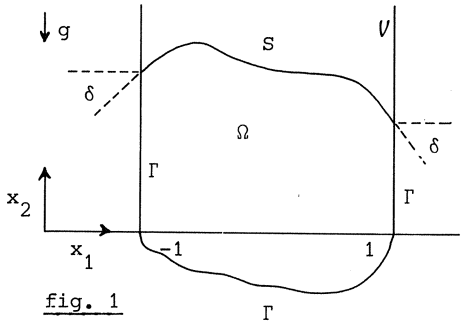


fig. 1

A fluid is contained in an open vessel V (fig. 1) placed in the field of gravity g. On the smooth boundary Γ of the vessel, Dirichlet type boundary conditions are prescribed. In the stationary situation the surface S of the liquid is not known a priori and is called a free boundary. The object is to find the shape of this meniscus

We assume that the liquid is incompressible (density ρ) with constant viscosity μ and surface tension T. The problem can be formulated in the following non dimensional form:

Find a velocity vector $u = \{u_1, u_2\}$, a pressure field p and an interface S (given by $\phi : x_1 \to x_2 = \phi(x_1)$, $-1 \le x_1 \le 1$, with given contact angle δ at $x_1 = -1, +1$), such that

$$(5.1) \qquad -\Delta u + \sum_{i=1}^{2} u_i \frac{\partial u}{\partial x_i} + \operatorname{grad} p = 0 \text{ in } \Omega \quad \text{(Navier-Stokes eq.}^{s}\text{)}$$

$$(5.2) \qquad \operatorname{div} u = 0 \text{ in } \Omega \text{ (incompressibility condition)}$$

with the following boundary conditions

$$(5.3) \qquad u(x) = \operatorname{Re} h(x) \text{ on } \Gamma$$

$$(5.4) \qquad \sigma_\tau = 0 \text{ on } S$$

$$(5.5) \qquad u_n = u.n = 0 \text{ on } S$$

$$(5.6) \qquad -\left(\frac{\phi'(x_1)}{\sqrt{1+(\phi'(x_1))^2}}\right)' + \beta\phi(x_1) = -\gamma\,\sigma_n \text{ on } S \quad \left(\phi' = \frac{d\phi}{dx_1}\right)$$

$$\text{with } \phi'(-1) = -\phi'(1) = \operatorname{tg}\delta$$

where σ_τ and σ_n are the tangential and normal component of the stress tensor σ_{ij}, defined in section IV.

The positive parameters Re, β and γ are defined by

$$\mathrm{Re} = \frac{\rho H}{\mu} = \text{'Reynolds' number} \qquad (H = \text{length scale})$$

$$\beta = \frac{\rho\, g\, H^2}{T} = \text{Bond number} \qquad \gamma = \frac{\mu^2}{\rho\, T\, H} = \text{Ohnesorge number}$$

Notice that

$$\beta = G\,N \qquad \gamma = \frac{N}{\mathrm{Re}\; h}$$

where G and N are defined in section IV.

The numerical method discussed in this section is defined by an existence and uniqueness result of the free boundary problem in weighted Hölder spaces $C^{\ell}_{s}(\Omega)$ due to SOLONNIKOV [47]. The principle of the proof is as follows. A shape of the free boundary is assigned and the flow-field within that shape is calculated after disregarding condition (5.6) on S. Next a new meniscus is computed which satisfies as closely as possible the boundary condition (5.6). This procedure is iterated until convergence is attained.

The weighted Hölder spaces $C^{\ell}_{s}(\Omega)$ consist of elements with bounded $|\cdot|^{(\ell)}_{s,\Omega}$ norm defined by

$$|u|^{(\ell)}_{s,\Omega} = |u|^{(s)}_{\Omega} + \sum_{s<|\alpha|<\ell} \sup_{x\in\Omega} \rho^{|\alpha|-s}(x)\,|D^{\alpha}u(x)| +$$

$$+ \sum_{|\alpha|=|\ell|} \sup_{x,y\in\Omega} R(x,y)\,\frac{|D^{\alpha}u(x)-D^{\alpha}u(y)|}{|x-y|^{\ell-[\ell]}}$$

with

$$|u|^{(s)}_{\Omega} = \sum_{|\alpha|<s} \sup_{x\in\Omega} |D^{\alpha}u(x)| + \sum_{|\alpha|=[s]} \sup_{x,y\in\Omega} \frac{D^{\alpha}u(x)-D^{\alpha}u(y)}{|x-y|^{s-[s]}},$$

$$|u|^{(s)}_{\Omega} = 0 \text{ for } s < 0$$

$$\rho(x) = \min\{|x-x^1|,|x-x^2|\},\ x^1=\{-1,\phi(-1)\},\ x^2=\{1,\phi(1)\},$$
$$R(x,y) = \min\{\rho^{\ell-s}(x),\rho^{\ell-s}(y)\}.$$

Similar definition for the space $C_s^{(\ell)}(-1,1)$, with ρ replaced by $\rho_o(t) = \min\{t+1, 1-t\}$.

THEOREM. $\exists\ R_o > 0$, $V_o > 0$, $\nu_o \in (0,1]$ *such that for* $V = \int_{-1}^{1} \phi(x_1)\, dx_1 \geq V_o$, $Re \leq R_o$, $\nu \leq \nu_o$ *problem* (5.1),...,(5.6) *has a unique solution* $\{\phi,u,p\} \in C_{1+\nu}^{3+\nu}(-1,1) \times C_{\nu}^{2+\nu}(\Omega) \times C_{\nu-1}^{1+\nu}(\Omega)$ *for all* $h \in C^{2+\nu}(\Gamma)$ *with* $\int_{\Gamma} h.n\, d\Gamma = 0$ *and* $h(x^i).n(x^i) = 0$, $i = 1,2$.

The proof of this theorem, for which we refer to SOLONNIKOV [47], is based on the following.

LEMMA. *For* ϕ fixed in $C_{1+\nu}^{3+\nu}(-1,1)$ *and with the conditions of the preceeding theorem, problem* (5.1),...,(5.5) *has a unique solution* $\{u,p_o\} \in C_{\nu}^{2+\nu}(\Omega) \times C_{\nu-1}^{1+\nu}(\Omega)$ *with* $\int_{\Omega} p_o\, dx = 0$.

The problem is to move ϕ in such a way that the normal stress balance (5.6) is satisfied.

For Re = 0, the unique solution $\{\phi_o, \bar{u}, \bar{p}\}$ of problem (5.1),...,(5.6) is given by

$$\text{(5.7)} \qquad \left| \begin{array}{l} \bar{u} = o \qquad \bar{p} = \dfrac{1}{\gamma}\left(\dfrac{\beta V}{2} - \sin\delta\right) \end{array} \right.$$

$$\text{(5.8)} \qquad \left| \begin{array}{l} -\left(\dfrac{\phi_o'}{\sqrt{1+(\phi_o')^2}}\right)' + \beta\,\phi_o = \dfrac{\beta V}{2} - \sin\delta \qquad \text{on } (-1,1) \\ \\ \phi_o'(-1) = -\phi_o'(1) = tg\,\delta \end{array} \right.$$

We introduce $\omega = \phi - \phi_o$ and it can be proved that ω satisfies

$$\text{(5.9)} \qquad \left| \begin{array}{l} -\left(\dfrac{\omega'}{\sqrt{1+(\phi_o')^2}}\right)' + \beta\,\omega = A \qquad \text{on } (-1,1) \\ \\ \omega'(-1) = \omega'(1) = 0 \end{array} \right.$$

with

$$A = \gamma(p_o - 2\frac{\partial u_n}{\partial n} - \tfrac{1}{2}\int_{-1}^{1}(p_o - 2\frac{\partial u_n}{\partial n})dx_1) + \frac{1}{R(\phi)} - \frac{1}{R(\phi_o)} - (\frac{\omega'}{\sqrt{1+(\phi_o')^2}})' =$$

$$= A(\phi, u, p_o)$$

We define an operator L: $C^{3+\nu}_{1+\nu}(-1,1) \to C^{3+\nu}_{1+\nu}(-1,1)$ by

$$L: \hat{\omega} \to \hat{\phi} \to A(\hat{\phi}, u(\hat{\phi}), p_o(\hat{\phi})) \xrightarrow{(5.9)} \omega$$

and it is proved that L is a contraction operator, which proves the existence and uniqueness of an element ϕ in the theorem.

The normal stress algorithm is the following

(i) $\phi^o = \phi_o$ (rest state)

(ii) Assume $\phi^1, \ldots, \phi^n$

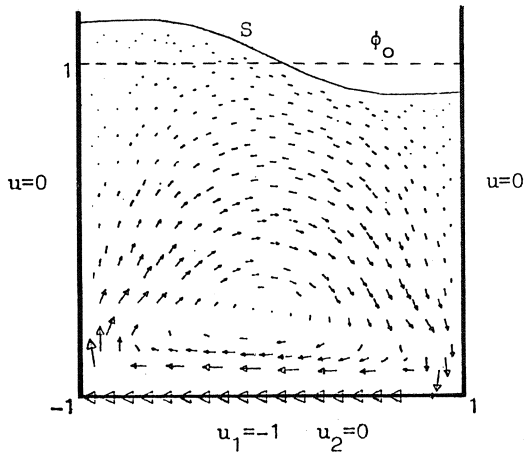
(iii) Solve problem (5.1),...,(5.5) with ϕ^n; solution $\{u^n, p_o^n\}$

(iv) Solve (5.9) with $A = A(\phi^n, u^n, p_o^n)$; solution ω^{n+1}

(v) $\phi^{n+1} = \omega^{n+1} + \phi_o$

The preceding algorithm is discretized by a finite element method in the case of the Stokes equations. The main point in the algorithm is the numerical solution of the Stokes equation in a fixed domain. This has been done using the augmented Lagrangian method (see section IV). The velocity is approximated by an extended quadratic conforming element (i.e. quadratic polynomials plus one third order term) and the pressure by a non-conforming linear element. It is proved in GIRAULT, RAVIART [28], CROUZEIR, RAVIART [9], that this method is of the second order.

The next figure shows the result of a numerical calculation using the proposed algorithm. For other results we refer to CUVELIER [15].



γ = 5.0
β = 0.0
δ = 0.0
iterations = 9

For an application of the optimal control method (discussed in section IV) to this problem, we refer to CUVELIER [16].

Finally let us remark that in the optimal control method two Stokes problems must be solved per iteration, but no pressure calculation is necessary (at least for 1 iteration with the augmented Lagrangian method). The normal stress method needs the solution of one Stokes problem per iteration, moreover a convergence proof can be given.

REFERENCES

[1] ARGYRIS, J.H. and SCHARPF, D.W., *The incompressible lubrication problem* J. Roy. Aero. Soc., 73, 1969, p. 1044-1046.

[2] BAIOCCHI, C., COMINCIOLI, V., MAGENES, E., and POZZI, G.A., *Free boundary problems in the theory of fluid flow through porous media. Existence and uniqueness theorems,* Annali di Mat. Pura ed Appl., 92, 1973, p. 1-82.

[3] BIRKHOFF, G., *Free boundary problems for viscous flows in channels.* In R. Davies (ed.), Cavitation in real liquids, Proceedings of the symposium on cavitation in real liquids, Michigan, 1964, p. 102-121.

[4] BREZIS, H., *Equations et inéquations non linéaires dan les espaces vectoriels en dualité.* Ann. Inst. Fourier, 18, 1968, p. 115-175.

[5] BREZIS, H., STAMPACCHIA, G., *Sur la regularité de la solution d'inéquations elliptiques*. Bull. Soc. Math. France, 96, 1968, p. 153-180.

[6] CAMERON, A., *Principles of lubrication,* Longmans, London, 1966.

[7] CIMATTI, G., *On a problem of the theory of lubrication governed by a variational inequality,* Appl. Math. and comp., 3, 1977, p. 227-242.

[8] COYNE, J.C., and ELROD, H.G., *Conditions for the rupture of a lubricating film,* Part I: Theoretical model. J. Lub. Tech., 92, 1970, p. 156-167.

[9] CROUZEIX, M. AND RAVIART, P.A., *Conforming and non-conforming finite element methods for solving the stationary Stokes equations*. Rev. Française Automat. Informat. Recherche Opérationelle, 7, 1973, p. 33-76.

[10] CRYER, C.W., *A survey of trial free boundary methods for the numerical solution of free boundary problems,* M.R.C. Technical Summary Report 1693, University of Wisconsin, 1976.

[11] CRYER, C.W., *A bibliography of free boundary problems,* M.R.C. Technical Summary Report, 1793, University of Wisconsin, 1977.

[12] CUVELIER, C. *Introduction to the numerical analysis of variational inequalities*. Report NA-22, Delft University of Technology, 1978.

[13] CUVELIER, C. *A free boundary problem in hydrodynamic lubrication including surface tension*. Proc. VIth Int. Conf. on Num. Meth. in Fluid Dynamics. Tbilisi (USSR), 1978, p. 39-44.

[14] CUVELIER, C., *A free boundary problem in hydrodynamic lubrication governed by the Stokes equations,* Proc. 9th IFIP Conference, 1979, Warsaw. Lecture notes in Control and Information Sciences, no. 22, p. 375-384.

[15] CUVELIER, C., *Capillary free boundary problems governed by the Navier-Stokes equations* (to appear).

[16] CUVELIER, C., *On the numerical solution of a capillary free boundary problem governed by the Navier-Stokes equations*. Report NA-34, 1980, Delft University of Technology. (To be published in Proc. 7th Int. Conf. Num. Meth. on Fluid Mech., June, 1980, Stanford, California, USA).

[17] CUVELIER, C., PRAAGMAN, N., SEGAL, A., *A survey of finite element methods in fluid mechanics*, Report NA-26, 1979, Delft University of Technology.

[18] DESAI, C.S., and JAVEL, J.F., *Introduction to the finite element method*, A numerical method for engineering analysis. Von Nastrand Reinhold, New York, 1972.

[19] DOWSON, D., GODET, M. and TAYLOR, C.H., (eds), *Cavitation and related phenomena in lubrication*, Proceedings on the 1st Leeds - Lyon Symposium on Tribology, Leeds, 1974.

[20] DUVAUT, G., *Résolution d'un problème de Stefan*, C.R.A.S., 276, 1973, p. 1461-1463.

[21] DUVAUT, G., *Report Universidade Federal*, Rio-de-Janeiro, 1975.

[22] DUVAUT, G., and LIONS, J.L., *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[23] ENGERING, F.P.H., *A free boundary problem in the theory of fluid flow through a porous dam*, Report, 1976, Delft University of Technology.

[24] FINN, W.D., *Finite element analysis of seepage through dams*, J. Soil Mech. Found. Div. A.S.C.E., 93, 1967, p. 41-48.

[25] FORTIN, M., *Minimization of some non-differentiable functionals by the augmented Lagrangian method of Hestenes and Powell*, Appl. Math. and Opt., 2, 1975, P. 236-250.

[26] FORTIN, M., and GLOWINSKI, R. (eds), *Résolution numérique de problèmes aux limites par des méthodes de Lagrangiens augmentés*, To appear.

[27] FRÉMOND, M., *Variational formulation of the Stefan problem, coupled Stefan problems, frost propagation in porous media*. Int. Conf. on Comp. Methods in non-linear mechanics, Austin, Texas, 1974.

[28] GIRAULT, V., and RAVIART, P.A., *Finite element approximation of the Navier-Stokes equations*, Lecture notes in mathematics 749, 1979.

[29] GLOWINSKI, R., *La méthode de relaxation*, Rendiconti di Matematica 14, Universita di Roma, 1971.

[30] GLOWINSKI, R., LIONS, J.L., and TREMOLIERES, R., *Analyse numerique des inequations variationelles*, I. Théorie générale, premières applications, Dunod, Paris, 1976.

[31] GLOWINSKI, R., LIONS, J.L. and TREMOLIERES, R., *Analyse numérique des inéquations variationelles,* II. Applications aux phénomènes stationaires et d'évolution, Dunod, Paris, 1976.

[32] LE MONDE, *La recherche sur la fusion thermonucleaire. Le JET: pour obtenir des températures inconnues sur la Terre,* de M. ARVONNY, article in 'LE MONDE', novembre 9th 1977.

[33] LIONS, J.L., *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles,* Dunod-Gautheirs-Villars, Paris (1968).

[34] LIONS, J.L., *Some aspects of the optimal control of distributed parameter systems,* SIAM (C.B.M.S. Series), Philadelphia (1972).

[35] LIONS, J.L., *Sur les inéquations aux dérivées partielles,* Uspekhi Mat. Nauk, 1971, t XXVI (2) p. 205-263.

[36] LIONS, J.L. and STAMPACCHIA, G., *Variational inequalities,* Comm. in Pure and Appl. Math. 20, 1967, p. 493-519.

[37] MERCIER, C., *The magnetohydrodynamic approach to the problem of plasma confinment in closed magnetic conficuration,* Publ. EURATOM-CEA, Luxembourg, 1974.

[38] MERCIER, C., ADAM, J.C., SOUBBARAMAYER, SOULE, J.L., *Problèmes et méthodes numériques en physique des plasmas à tres haute température,* Proc. Int. Symp. Versailles, Dec. 1973. Lecture notes in Computer Science, no. 11, p. 65-106.

[39] MOSSINO, J., *Etude de quelques problèmes non linéaires d'un type nouveau apparaissant en physique des plasma,* Thèse Université de Paris-Sud, Orsay, 1977 (also: CRAS 282,A,p.187).

[40] MOSSINO, J., TEMAM, R., *Certains problèmes non linéaires de la physique des plasma,* Proc. Conf. Math. Aspects of FEM, Rome, 1977, Lecture notes in Mathematics no. 606.

[41] PIRONNEAU, O., *On optimum design in Stokes flow,* J. Fluid Mech. 59, 1973, p. 117-128.

[42] PIRONNEAU, O., *On optimum design in fluid mechanics,* J. Fluid Mech. 64, 1974, p. 97-111.

[43] PIRONNEAU, O., *Sur les problèmes d'optimisation de structures en mécanique des fluides,* Thèse de doctorat, Université Paris VI, 1976.

[44] REDDI, M.H., *Finite element solution of the incompressible lubrication problem,* Trans. Am. Soc. Mech. Eng. 91, ser. F, 1969, p. 524.

[45] SEGAL, A., *On the numerical solution of Stokes equations using the finite element method,* Comp. Meth. in Appl. Mech. and Eng., 19, 1979, p. 165-185.

[46] SERMANGE, M., *Une methode numerique de bifurcation,* Rapport IRIA no. 310, 1978.

[47] SOLONNIKOV, V.A., *Solvability of the boundary value problem describing the motion of a viscous incompressible capillary fluid in an open vessel in the two dimensional case.* Ivzestia-AN.SSR, 1979, 43' p. 203-236 (in Russian).

[48] STAMPACCHIA, G., *Variational inequalities,* In Theory and Applications of Monotone Operators, Oderisi, Venise, 1968, p. 101-192.

[49] TAYLOR, R.L. and BROWN, C.B., *Darcy flow with free surface,* J. Hyd. Div. A.S.C.E. 93, 1967, p. 25-33.

[50] TEMAM, R., *A nonlinear eigenvalue problem: the shape of equilibrium of a confined plasma,* A.R.M.A. 60, 1976, p. 51-73.

[51] TEMAM, R., *Nonlinear boundary value problems arising in physics,* In 'Differential Equations and Applications' (ed. W. Echhaus), p. 27-38.

[52] TEMAM, R., *Remarks on a free boundary value problem arising in plasma physics,* Comm. Part Diff. Eq. vol 2. no. 6, 1977, p. 563-586.

[53] TEMAM, R., *Navier-Stokes equations: Theory and numerical analysis,* North-Holland Publishing Comp. Amsterdam, 1977.

[54] VOLKER, R.E., *Non-linear flow in porous media by finite elements.* J. Hyd. Div. A.S.C.E., 95, 1969, p. 2093-2114.

ON MODIFIED INCOMPLETE FACTORIZATION

I. GUSTAFSSON

1. INTRODUCTION

In this contribution we present the main theoretical results for the modified incomplete factorization (MIC) methods. For a more detailed study including proofs see [1], where also a more complete list of references can be found. A number of numerical results concerning different classes of problems is included.

We solve the system of linear equations

$$Au = f$$

by a preconditioned iterative process, which basic form is

$$C(u^{\ell+1}-u^{\ell}) = -\beta_{\ell}(Au^{\ell}-f), \qquad \ell = 0,1,\ldots,$$

where C is the preconditioning matrix.

In the case when A and C are symmetric, positive definite we can use the Chebyshev method or a conjugate gradient type of method as an acceleration procedure. As we shall see for some kinds of non-symmetric problems this is still possible.

In order to get an efficient method, the preconditioning matrix C has to have the following properties:

1. It should be easily calculated.
2. It should not need too much storage.
3. Systems with the matrix C should be easily solved. Typically C = LU, with sparse lower and upper triangular factors L and U.
4. The spectral condition number $\varkappa(C^{-\frac{1}{2}}AC^{-\frac{1}{2}})$ of $C^{-\frac{1}{2}}AC^{-\frac{1}{2}}$ should be much smaller than $\varkappa(A)$, the spectral condition number of A.

Our choice of C is based on modified incomplete factorization (defined in the following section) of A and fulfills the desired properties stated above.

In this paper, we present the methods for well-structured matrices such as those arising in finite element (FEM) or finite difference (FDM) methods. The idea of modification can obviously even be used in a more general context, where the accepted fill-in during the approximate factorization is dynamically controlled, see e.g. [2].

2. MODIFIED INCOMPLETE FACTORIZATION

It is well known that a complete Gaussian or Cholesky factorization of a sparse matrix produces fill in within the band in the upper and lower triangular factors. This leads to a relatively high computational complexity for the factorization and to considerably large storage requirement. By using an incomplete factorization we keep the sparsity and hence need much less computational labour and storage.

The incomplete factorization is used as a preconditioning for the iterative process or, equivalently, one makes an incomplete factorization followed by a number of iterative refinement steps.

For well structured FDM and FEM matrices the positions, where fill-in is allowed during the elimination, can be chosen in advance. Let $A = (a_{ij})$ be the $N \times N$ matrix to be factored and let

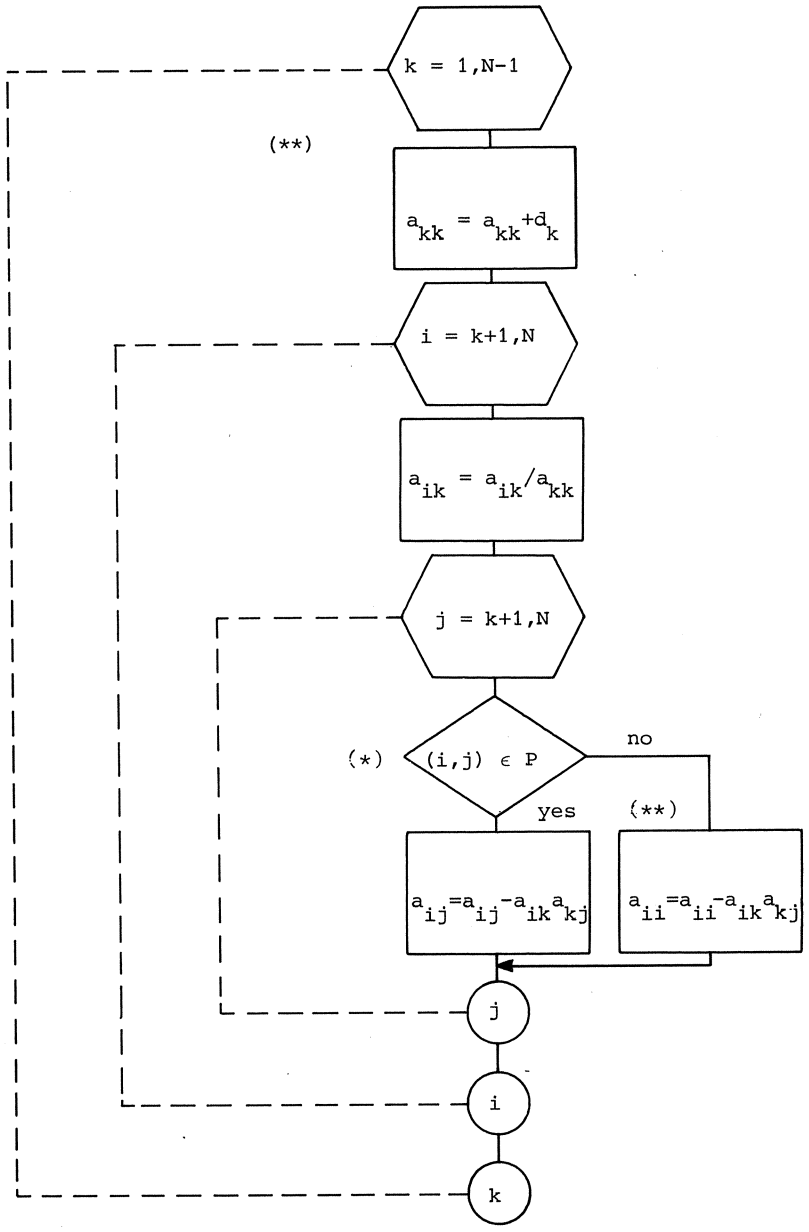
$$P^* = \{(i,j);\ a_{ij} \neq 0\}.$$

Further, let P be positions, where we allow fill-in in the factors $L = (\ell_{ij})$ and $U = (u_{ij})$ of C, i.e.

$$P = \{(i,j);\ \ell_{ij} \neq 0 \text{ or } u_{ij} \neq 0\}.$$

In this paper, we assume that $P^* \subseteq P$, that is, we have fill-in in at least positions where A has nonzero entries.

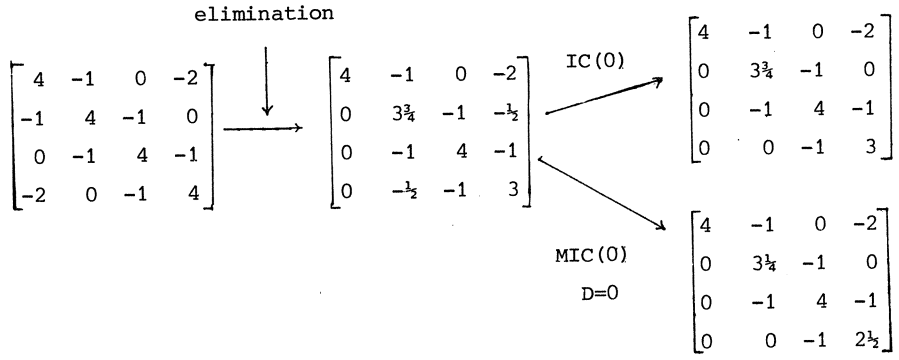
In the following flow-chart, the MIC factorization algorithm, with normalization diag(L) = I, is described in a general context. For the definition of the diagonal matrix $D = \text{diag}(d_k)$ see the following analysis.



From the flow-chart representing a complete factorization (Gaussian-elimination) we obtain an *incomplete* factorization (of our kind) by introducing the test (*) and a *modified incomplete* factorization by further introducing the statements (**). In a MIC factorization we do not drop elements but keep the information by modifying the diagonal.

If $P = P^*$, (that is, if we allow no fill in), we use the notations IC(0)

and MIC(0), respectively. We describe these factorizations for the following simple example.



After three elimination steps we have

$$C = LU = A + R, \text{ where}$$

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \end{bmatrix} \quad \text{for IC(0) and } R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix}$$

for MIC(0). Notice that for the MIC(0) method, R has rowsum = 0 and that R is negative semidefinite. This is true in a more general context as will be stated below.

Let A be of order N. After N-1 modified incomplete factorization steps we have

$$C = LU = A + D + R,$$

where R is the defect matrix. Obviously, rowsum (R) = 0. Furthermore, if A is an M-matrix ($A^{-1} \geq 0$, $a_{ij} \leq 0$, $i \neq j$), then R is negative semidefinite. This is so since all off-diagonal entries of R are non-negative.

3. STABILITY

DEFINITION 3.1. An incomplete LU factorization is said to be *stable* iff diag(U) > 0.

Observe that in the case when A is symmetric, positive definite, stability means that C(= LL^T) is also symmetric, positive definite.

THEOREM 3.1. *If* A *is weakly diagonally dominant* ($a_{ii} \geq \sum_{j \neq i} |a_{ij}|$, $i = 1,\ldots,N$) *then any* MIC *factorization of* A *is stable.*

The MIC algorithms are in general not stable for M-matrices, which is the case for the IC algorithms, see [3].

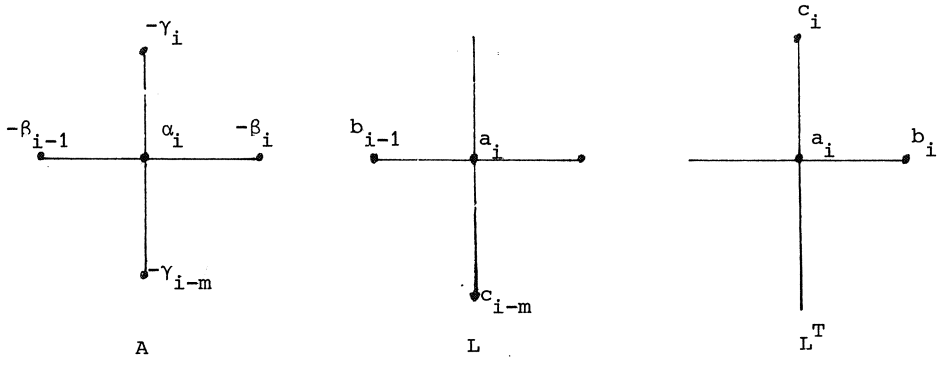
4. MIC(d) ALGORITHMS FOR FEM MATRICES

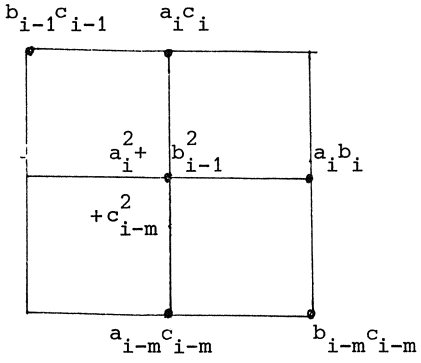
Recall that in MIC(0), $P = P^*$, that is, we allow no fill-in. We can obtain more accurate factorizations by admitting fill-in to some extent. We may then use the following strategy:

1. At first let L and U contain non-zero entries in the same positions as A. (This represents the MIC(0) method)
2. Form C = LU and R = C - A.
3. Re-define L and U in such a way that these matrices are allowed to contain non-zero entries in positions where R has non-zero entries as well.
4. If you are not satisfied, repeat from stage 2.

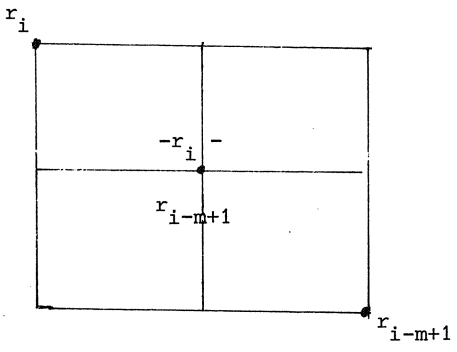
For well structured matrices scuh as FDM and FEM matrices we then get MIC(d) algorithms, where d > 0 indicates that L contains d more non-zero sub-diagonals than the lower part of A. Practical experiments indicate that we get the most efficient method after one or two cycles of the above strategy.

EXAMPLE 4.1. Consider the 5-point FDM matrix arising from a second order self-adjoint elliptic boundary value problem on the unit square. The matrices A, $C = LL^T$ and R for MIC(0) are defined by the following graphs, where as usual the graph nodes coincides with the FDM nodes. m is the half band width of A.





LL^T

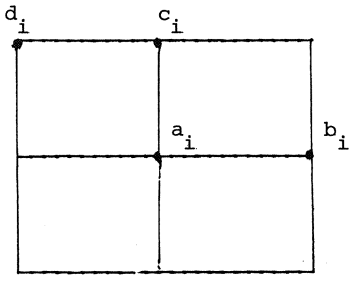


R

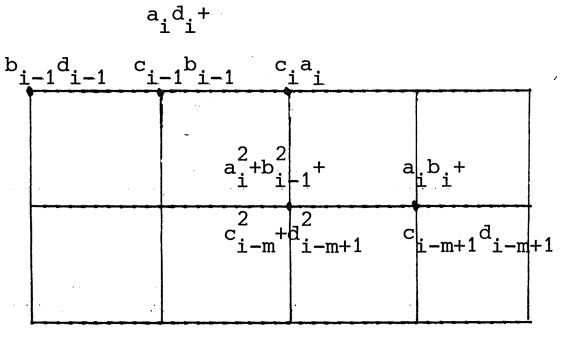
The relation $C = LL^T = A + D + R$ gives the following recursion formulas for the entries of L, defining the MIC(0) algorithm:

$$
\begin{aligned}
a_i^2 &= \alpha_i + \delta_i - b_{i-1}^2 - c_{i-m}^2 - r_i - r_{i-m+1} \\
b_i &= -\beta_i/a_i \\
c_i &= -\gamma_i/a_i \\
r_i &= b_{i-1}\, c_{i-1}.
\end{aligned}
\tag{4.1}
$$

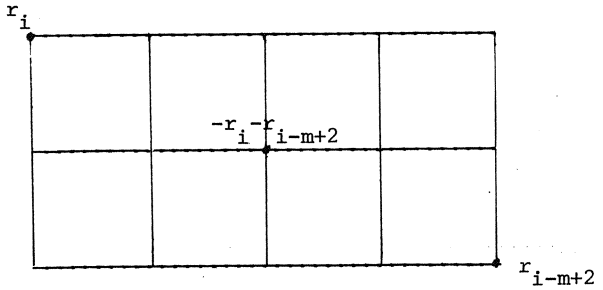
The strategy to obtain a more accurate factorization now leads to the MIC(1) algorithm defined by the graphs;



L^T



LL^T



R

and the recursion formulas

$$a_i^2 = \alpha_i + \delta_i - b_{i-1}^2 - c_{i-m}^2 - d_{i-m+1}^2 - r_i - r_{i-m+2}$$

$$b_i = -(\beta_i + c_{i-m+1} d_{i-m+1})/a_i$$

$$c_i = -\gamma_i / a_i$$

$$d_i = -b_{i-1} c_{i-1} / a_i$$

$$r_i = b_{i-1} d_{i-1}.$$

Continuing in this way we get a sequence of more and more accurate factorizations; MIC(0), MIC(1), MIC(2), MIC(4), MIC(7), MIC(12), etc.

REMARK 4.1. In practice we may avoid the square-roots by making a MIC factorization of the type $C = (L+D_1)D_1^{-1}(L^T+D_1)$, D_1 a diagonal matrix and L strictly lower triangular.

REMARK 4.2. In the recursion formulas (4.1) for the MIC(0) algorithm, a_i can be calculated from

$$a_i^2 = \alpha_i + \delta_i - b_{i-1}(b_{i-1}+c_{i-1}) - c_{i-m}(c_{i-m}+b_{i-m})$$

in order to decrease the number of operations. Similarly for MIC(d), d > 0.

REMARK 4.3. In a Dirichlet problem, we choose $\delta_i = \alpha_i \xi h^2$, $i = 1,\ldots,N$, where $\xi >$ is a parameter and h is a mesh parameter. For the choice of $D = \mathrm{diag}(\delta_i)$ in general see the following analysis.

5. THE MIC(0)* ALGORITHM

We also have the following variant of the MIC(0) algorithm, the MIC(0)* algorithm, which can be thought of as a generalized SSOR method.

The MIC(0)* algorithm is obtained by disregarding in the MIC(0) algorithm all corrections $-a_{ik}a_{kj}$ to the entries a_{ij}, $i \neq j$, see (1.1). These numbers are instead added to the diagonal of U to get rowsum (R) = 0. Apparently, the MIC(0)* algorithm is of the type $C = (\tilde{D}+L)\tilde{D}^{-1}(\tilde{D}+L^T)$, where $\tilde{D} > 0$ is a diagonal matrix and L is the strictly lower triangular part at A, compare with the SSOR method [4].

The advantage with this method is that it needs less storage and less factorization work. On the other hand, the convergence is in general somewhat slower than for the MIC(0) method.

The MIC(0)* algorithm as well as MIC(d) algorithms, $d \geq 0$, are covered by the following analysis of the rate of convergence of the corresponding MICCG method, that is, the MIC factorization combined with the conjugate gradient (CG) method.

6. RATE OF CONVERGENCE OF MICCG METHODS (The symmetric case)

Let h be a mesh parameter and let m_k, $k = 1,2,\ldots$ be independent on h. Further, let n be the space dimension.

Assume that

(i) A is a symmetric M-matrix of order $N = 0(h^{-n})$, $n \to 0$,

(ii) Rowsum (A) ≥ 0,

(iii) $-2\sum_{j>i} a_{ij} \leq a_{ii} + m_1 h$, $\quad i \in N_1$, where $N_1 \subseteq N = \{i;\ 1 \leq i \leq N\}$ and the number of indices in $N_2 = N \backslash N_1$ is $0(h^{-n+1})$, $h \to 0$.

From (i) and (ii) it follows that A is weakly diagonally dominant and hence any MIC factorization is stable.

If we now define $D = \mathrm{diag}(\delta_i)$ in a proper way, namely

(iv) $$\delta_i = \begin{cases} \xi_1 h^2 a_{ii}, & i \in N_1 \\ \xi_2 h\, a_{ii}, & i \in N_2, \end{cases}$$

where $\xi_i > 0$, i = 1,2 are parameters, then we have the following result for the spectral condition number H_1 of $C^{-\frac{1}{2}}AC^{-\frac{1}{2}}$.

THEOREM 6.1. *Assume that* A *satisfies* (i), (ii) *and* (iii). *Then, if* D *is chosen according to* (iv), $H_1 = 0(h^{-1})$, $h \to 0$.

Since, as is well known, the number of iterations in the MICCG method is of order $0(\sqrt{H_1})$, we state the following corollary.

COROLLARY 6.1. *Assume that the conditions* (i) - (iv) *are satisfied. Then the number of iterations in the* MICCG *method is* $0(h^{-\frac{1}{2}})$, $h \to 0$ *and the number of arithmetic operations is* $0(N^{1+1/2n})$, $N \to \infty$.

If A is a weakly diagonally dominant M-matrix, that is, (i) and (ii) are fulfilled, then (iii) is satisfied for the following types of elliptic PDE problems, if the mesh points are numbered in a natural (rowwise or similar) way:

a) Dirichlet problems with constant coefficients (Laplace equation). In this case $N_1 = N$ and $m_1 = 0$.
b) Dirichlet problems with Lipschitz continuous material coefficients. Here, $N_1 = N$.
c) Neumann problems. Then, N_2 represents points on and/or near the Neumann boundary.
d) Problems with discontinuous material coefficients. Here, N_2 represents points on and/or near an interface over which the coefficients are discontinuous.

REMARK. In numerical tests, the number of iterations has turned out to be almost independent of the parameters ξ_1, ξ_2 in a fairly wide range. In fact, the choice $\xi_1 = \xi_2 = 0$, i.e. D = 0, is almost as good as the optimal choice. Hence, in practice, the definition of N_2 offers no problem. For instance, one can use D = 0 or $D = \xi h^2\, \mathrm{diag}(A)$, $\xi > 0$ (say $\xi = 1$) for all types of problems.

7. MIC FOR MORE GENERAL FEM MATRICES

For matrices that are not diagonally dominant, the MIC as well as the IC factorizations may be instable. This can be overcome by using skifted incomplete factorizations (SIC, SMIC), see e.g. [5], [6]:

Positive numbers are added to the diagonal of U (or A) if non-positive diagonal elements are produced.

In general this approach leads to slower convergence, in particular if this skifting has to be done quite often. Although this author has tested several types of FEM problems with different kinds of approximations, the MIC algorithms have never turned out to be instable. Furthermore, the same rate of convergence as was shown in section 6 has been measured for more general matrices as well, see the results presented in section 10. (For the IC algorithms, however, instability has been observed, see also [5]).

For many kinds of problems we can obtain (even theoretically) the same fast rate of convergence as was shown in the previous section, by using the following idea.

Spectral equivalence

Let A_h and B_h of order $N = N(h)$ be two discretizations of a second order elliptic differential operator corresponding to a mesh size parameter h.

DEFINITION 7.1. A_h and B_h are spectrally equivalent if

$$0 < c \leq \frac{(A_h x,x)}{(B_h x,x)} \leq C, \forall x \in R^N,\ x \neq 0,$$

where c,C are independent of h.

Assume now that we have an original, coarse FEM mesh consisting of quadrilateral or triangular elements with all angles $\leq \pi/2$ and let the mesh be refined in a uniform way, see e.g. [7]. Further, let $A_h^{(1)}$ and $A_h^{(p)}$ be the matrices corresponding to piecewise polynomial basis functions of degree 1 (linear) and p, respectively. We then have that $A_h^{(1)}$ and $A_h^{(p)}$ are spectrally equivalent, for detailes see [7].

EXAMPLE 7.1. Consider the Laplace equation, $-\Delta u = f$ in K_e, $u = 0$ on ∂K_e, K_e the unit square, discretized by FEM based on a uniform right angled

triangular mesh. Let $A_e^{(1)}$ and $A_e^{(2)}$ be element stiffness matrices corresponding to linear and quadratic basis functions, respectively. $A_e^{(1)}$ is assembled from the four elements corresponding to the finer mesh, see Fig. 7.1.

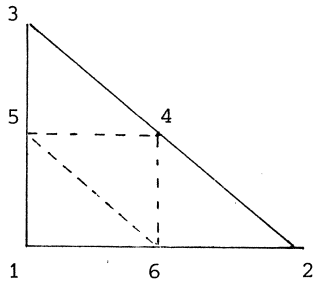


Fig. 7.1. An element

We have

$$A_e^{(1)} = \begin{bmatrix} 2 & 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 4 & -2 & -2 \\ -1 & 0 & -1 & -2 & 4 & 0 \\ -1 & -1 & 0 & -2 & 0 & 4 \end{bmatrix}$$

$$A_e^{(2)} = \begin{bmatrix} 6 & 1 & 1 & 0 & -4 & -4 \\ 1 & 3 & 0 & 0 & 0 & -4 \\ 1 & 0 & 3 & 0 & -4 & 0 \\ 0 & 0 & 0 & 16 & -8 & -8 \\ -4 & 0 & -4 & -8 & 16 & 0 \\ -4 & -4 & 0 & -8 & 0 & 16 \end{bmatrix}$$

and it is easily seen that

$$2(A_e^{(1)}x,x) \le (A_e^{(2)}x,x) \le 4(A_e^{(1)}x,x), \ \forall x \in R^6.$$

Since the global matrices $A_h^{(1)}$ and $A_h^{(2)}$ are "sums" of element matrices, we get

$$2(A_h^{(1)}x,x) \le (A_h^{(2)}x,x) \le 4(A_h^{(1)}x,x), \ \forall x \in R^N.$$

Hence, $A_h^{(1)}$ and $A_h^{(2)}$ are spectrally equivalent and

$H((A_h^{(1)})^{-\frac{1}{2}} A_h^{(2)} (A_h^{(1)})^{-\frac{1}{2}}) \leq 2.$

Now, $A_h^{(1)}$ is a diagonally dominant matrix so any MIC factorization $C = A_h^{(1)} + D + R$ is stable and $H(C^{-\frac{1}{2}} A_h^{(1)} C^{-\frac{1}{2}}) = 0(h^{-1})$, $h \to 0$. Due to the spectrally equivalence we then have

$$H(C^{-\frac{1}{2}} A_h^{(2)} C^{-\frac{1}{2}}) = 0(h^{-1}), \quad h \to 0$$

as well. Note that when we use this method we have to assemble both $A_h^{(1)}$ and $A_h^{(p)}$. On the other hand a MIC factorization of $A_h^{(1)}$ is often much simpler to perform than one of $A_h^{(p)}$.

8. A BIHARMONIC PROBLEM

Let A correspond to the 13-point difference approximation of the biharmonic problem

$$\begin{cases} \Delta^2 u = f & \text{in } K_e \\ u = u_n = 0 & \text{on } \partial K_e \end{cases}$$

and let A_1 correspond to the 5-point difference approximation of the Laplace equation ($\Delta u = g$ in K_e, $u = 0$ on ∂K_e). Define $\tilde{A} = A_1^2$. Then A and $\tilde{A}$ are spectrally equivalent (Definition 7.1) and $H(\tilde{A}^{-\frac{1}{2}} A \tilde{A}^{-\frac{1}{2}}) = 3$, see [8].

Let $C_1 = A_1 + D + R$ be a MIC factorization of A_1. This is stable since A_1 is diagonally dominant and $H(C_1^{-\frac{1}{2}} A_1 C_1^{-\frac{1}{2}}) = 0(h^{-1})$, $h \to 0$.

As preconditioning matrix for A we choose $C = \{(C_1^{-1})^2\}^{-1}$. Then $H(C^{-\frac{1}{2}} A C^{-\frac{1}{2}}) = H(C_1^{-1} A_1 A_1 C_1^{-1}) = 0(h^{-2})$, $h \to 0$ and the number of iterations in MICCG is $0(h^{-1})$, $h \to 0$. Note that the condition number of A, $H(A) = 0(h^{-4})$, $h \to 0$.

In each iteration we solve four triangular systems with the triangular factors of C_1 and the total number of operations is $0(N^{1+1/2})$, $N \to 0$.

REMARK. If we use an iterative method on the form

$$(A_1)^2 (u^{\ell+1} - u^{\ell}) = -\beta_\ell (Au^{\ell} - f), \qquad \ell = 0,1,..$$

we have only 0(1) number of iterations. In each iteration we solve two systems with matrix A_1 by (for instance) a MICCG method (inner iterations). Thus, the number of operations is $0(N^{1+1/4})$, that is of the same order as

for a second order problem. From a practical point of view, however, this seems to be preferable only for fairly small values of h.

9. MIC FOR A CLASS OF NON-SYMMETRIC PROBLEMS

Consider differential equations on the form

$$Lu = -\Delta u + \vec{v}\cdot\nabla u + cu = f, \quad u = u(x), \quad x \in \Omega$$

with suitable boundary conditions on $\partial\Omega$.

Assume that we use a positive difference scheme (upwind, modified upwind, Il'ins etc.). Then the associated matrix A is diagonally dominant and the MIC factorization C = A + D + R is stable.

Let A_0 be the matrix corresponding to $\vec{v} \equiv 0$ and let $A_s = \frac{1}{2}(A+A^T)$ be the symmetric part of A. We assume that

$$(A_s x,x) \geq (A_0 x,x), \quad \forall x \in R^N.$$

A sufficient condition for this to be true is that $\operatorname{div}(\vec{v}) \leq 0$, since $\frac{1}{2}(L+L^*)u = -\Delta u - \frac{1}{2}\operatorname{div}(\vec{v})u + cu$.

Further, assume that $\|\vec{v}\|$ is not too large (compare to h^{-1}), say $\|\vec{v}\| \leq ch^{-1}$, c independent of h.

Then the eigenvalues of $C^{-1}A$ are situated in an ellipse in the complex plane, centered at (1.0), with (after normalization) $\alpha = 1-m_1 h$, $\beta = m_2 h$, see Figure 9.1.

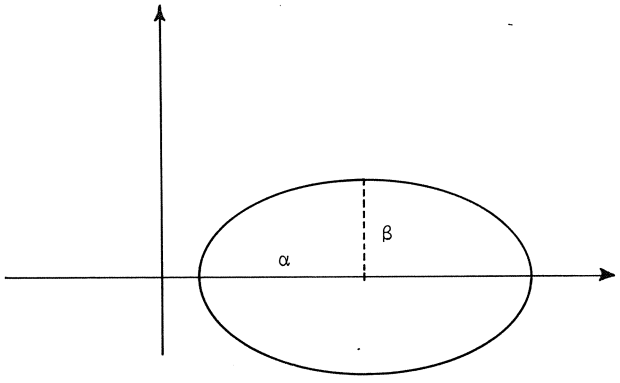


Figure 9.1. The ellipse containing the eigenvalues of $C^{-1}A$

Then the number of iterations, p, in the Chebyshev acceleration method is of order

$$p = O([\ln(1+\sqrt{1-\alpha^2+\beta^2}) - \ln(\alpha+\beta)]^{-1}) = O(h^{-\frac{1}{2}}), \quad h \to 0,$$

for details see [1].

For the unpreconditioned method the results are; $\alpha = 1-m_3h^2$, $\beta = m_4h$ and $p = O(h^{-1})$, $h \to 0$.

We note that a generalized conjugate direction method, see [9], is appropriate even for non-symmetric problems.

10. NUMERICAL RESULTS

We present some typical results from numerical experiments using MICCG methods. For further tests we refer to [1] and the references therein.

As initial approximation u_0 to the conjugate gradient method we have used $u_0 = C^{-1}f$ and the iterations were stopped when the residual error (in ℓ_2-norm) was reduced by a factor ε, that is, when $\|r^{\ell}\| \leq \varepsilon\|r^0\|$

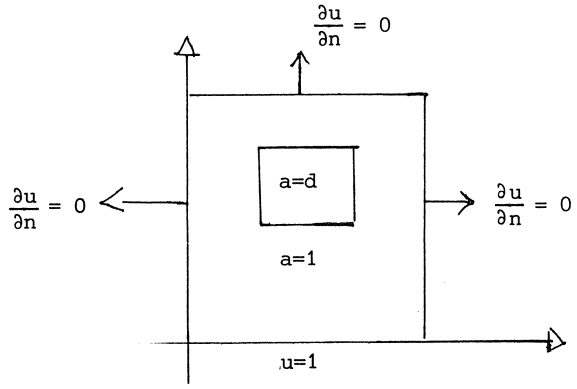
EXAMPLE 10.1. The model Laplace equation on the unit square,

$$\begin{cases} \Delta u = f & \text{in } K_e \\ u = 0 & \text{on } \partial K_e. \end{cases}$$

discretized by bilinear finite elements. The number of iterations for various methods and different values of N, $\varepsilon = 10^{-6}$:

N	MIC(0)	MIC(0)*	SSOR	IC(0)	IC(0)*	MIC(2)
100	8	8	9	8	9	5
400	11	12	12	13	15	7
1600	16	17	18	23	27	10

EXAMPLE 10.2. A discontinuous problem, $-\frac{\partial}{\partial x}(a\frac{\partial u}{\partial x}) - \frac{\partial}{\partial y}(a\frac{du}{\partial y}) = 1$ on Ω, see the figure, where d is a parameter used to vary the degree of discontinuity.



The number of iterations for different methods when the five point FDM approximation was used, $\varepsilon = 10^{-6}$:

d	1			100			10 000		
h^{-1}	SSOR	MIC(0)	MIC(2)	SSOR	MIC(0)	MIC(2)	SSOR	MIC(0)	MIC(2)
6	11	10	6	13	12	7	15	13	7
12	16	14	8	21	17	10	26	18	11
24	26	19	12	34	23	14	42	25	15
27	29	21	13	37	25	15	46	27	16

Notice the remarkable insensitivity to the parameter d observed for the MIC methods.

EXAMPLE 10.3. The problem defined in Example 7.1. The number of iterations for different MIC factorization of $A_h^{(1)}$ and $A_h^{(2)}$, $\varepsilon = 10^{-4}$:

N	factorizations of $A_h^{(2)}$		factorizations of $A_h^{(1)}$		
	MIC(0)	MIC(4)	MIC(0)	MIC(1)	MIC(2)
49	6	4	7	6	6
225	9	5	9	7	6
961	13	7	13	10	8

If one compares the total work for the methods one finds that for N = 961, the MIC(4) factorizations of $A_h^{(2)}$ is only slightly more efficient than the MIC(2) factorization of $A_h^{(1)}$ and that the MIC(0) factorization of $A_h^{(1)}$ is more efficient than the MIC(0) factorization of $A_h^{(2)}$.

EXAMPLE 10.4. The biharmonic problem described in Section 8. The number of iterations for $\varepsilon = 10^{-3}$ and for the MIC factorizations of A and A_1, respectively:

h^{-1}	Factorization of A_1	Factorizations of A	
	MIC(2)	MIC(0)	MIC(4)
5	4	4	2
10	6	7	3
20	10	16	7
40	17	37	14

The work per unknown for $h^{-1} = 40$ is for the MIC(2) factorization of A_1, the MIC(0) and MIC(4) factorizations of A about 630, 1210 and 620 operations, respectively. One realizes that one needs a quite accurate and hence more complicated factorization of A to reach the same efficiency as for a simpler factorization of A_1.

EXAMPLE 10.5. The non-symmetric problem

$$-\Delta u + \gamma(u'_x - u'_y) = 1 \text{ in } K_e \quad u = 0 \quad \text{on } \partial K_e,$$

where $\gamma \geq 0$ is used to vary the degree of non-symmetry. As discretization method we have used standard 5-point approximation for Δu and upwind differences for the derivatives of first order. The number of iterations needed in the minimum residual CG method [9] for various preconditionings and different values of h and γ, $\varepsilon = 10^{-3}$:

h	C \ γ	0	5	10	100	10^5
1/8	I	8	18	19	18	18
	MIC(0)	4	5	5	7	7
	MIC(2)	3	3	3	3	3
1/16	I	18	34	34	34	35
	MIC(0)	6	6	7	10	11
	MIC(2)	4	4	4	4	4
1/32	I	36	72	72	65	62
	MIC(0)	9	9	10	12	13
	MIC(2)	6	6	6	5	5

Observe that, when one uses a more accurate MIC factorization, the number of iterations is almost independent on γ even for fairly large values of γ, that is, for singularly pertubed problems.

11. CONCLUSIONS

The properties of the MICCG methods presented in this contribution can be summarized in the following statements:

- The result $H(C^{-\frac{1}{2}}AC^{-\frac{1}{2}}) = O(h^{-1})$, $h \to 0$ has been proved for several classes of FEM matrices corresponding to 2'nd order elliptic differential equations.
- The number of iterations for MICCG is $O(h^{-\frac{1}{2}}\ln 1/\varepsilon)$ to reach a relative residual error ε.
- The number of operations for MICCG is $O(N^{1+1/2n}\ln N)$, $N \to \infty$, if $\varepsilon = O(N^{-\mu})$, $\mu > 0$ and n is the space dimension.
- For a model biharmonic problem in two dimensions, the number of operations has been shown to be $O(N^{1+1/2}\ln N)$.
- The storage requirement is of optimal order $O(N)$.
- They are easy to program.

REFERENCES

[1] GUSTAFSSON, I., *Stability and rate of convergence of modified incomplete Cholesky factorization methods*, Report 79.02 R, Department of Computer Sciences, Chalmers University of Technology, Göteborg, Sweden (1979).

[2] MUNKSGAARD, N., *Solution of general sparse symmetric sets of linear equations*, Report No. NI-78-02, Inst. for Num. Anal., Technical University of Denmark, Lyngby, Denmark (1978).

[3] MEIJERINK, J.A. & H.A. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp. 31 (1977), 148-162.

[4] YOUNG, D., *Iterative solution of large linear systems*, Academic Press, New York and London (1971).

[5] KERSHAW, D., *The incomplete Cholesky conjugate gradient method for the iterative solution of systems of linear equations*, J. Comput. Phys. 26 (1978), 43-65.

[6] MANTEUFFEL, T.A., *The shifted incomplete Cholesky factorization*, Technical Report, Appl. Math. Division 8325, Sandia Laboratories, Livermore, California, USA (1978).

[7] AXELSSON, O. & I. GUSTAFSSON, *A preconditioned conjugate gradient method for finite element equations, which is stable for rounding errors*, Report 7924, Mathematisch Instituut, Katholieke Universiteit, Nijmegen, The Netherlands (1979).

[8] GUSTAFSSON, I., *On modified incomplete factorizations for a biharmonic problem, in progress.*

[9] AXELSSON, O., *A generalized conjugate direction method and its application on a singular perturbation problem, in* Proceedings, 8th Biennial Numerical Analysis Conference, Dundee, Scotland, June 26-29, 1979 (ed. G.A. Watson), Lecture Notes in Mathematics #773, Springer, Berlin, 1980.

INTRODUCTION TO MULTIGRID METHODS

P.W. HEMKER

A convenient way to give an introduction to multigrid methods is by means of the notion of "Defect Correction Process". Defect correction processes are general iterative processes for the approximation of operator equations. A large number of well known iterative methods can be classified into this category, and among these are the multigrid methods. Therefore, we give an introduction to elementary defect correction processes (DCP) in Section 1. In Section 2 we shall elaborate the idea of DCP to get the framework to fit the multigrid methods in. In Section 3 we give a short introduction to the discretization of analytic problems, with a special emphasis on the discretization on related grids, as they are used in multigrid methods. In Section 4 we treat the principles of multigrid algorithms and we give the basic structure of convergence proofs of multigrid methods. Some examples of applications of multigrid methods are given in Section 5.

1. ELEMENTARY DEFECT CORRECTION PROCESSES

In principle, a defect correction process is an iterative process to solve an equation that we cannot or don't want to solve directly:

(P) $\qquad Fx = y,$

where F is a mapping from A to B. A and B are normed linear spaces. In general the mapping F is non-linear, F is not defined on the whole of A and F is neither injective nor surjective.

We assume that there exist subsets $X \subset A$ and $Y \subset B$ such that F is defined on the whole of X, and the mapping $F:X \to Y$ is surjective. In addition we often require that there exists a unique $x \in X$ such that $Fx = y$ (or: in addition the mapping $F:X \to Y$ is injective and hence it is

bijective). We assume that we *can* solve some approximations $(\tilde{P})$ of the problem (P), i.e. for all $\tilde{y} \in \tilde{Y} \subset Y$ we can solve the equation

$$(\tilde{P}) \qquad \tilde{F}\tilde{x} = \tilde{y}, \qquad \tilde{x} \in \tilde{X},$$

where $\tilde{F}: X \to \tilde{Y}$ is some "approximation" of the operator F.
Formally we describe this as follows:
We assume that for some subset $\tilde{Y} \subset Y$ with $y \in \tilde{Y}$, there exists a mapping

$$\tilde{G}: \tilde{Y} \to X,$$

which we shall call the *approximate inverse* of F.

The meaning of $\tilde{G}$ is that for any $y \in \tilde{Y}$ an approximation to the solution of the equation $Fx = y$ is given by $\tilde{G}y$. The mapping $\tilde{G}$ needs not to be linear and neither injective nor surjective.

REMARK 1. If $\tilde{G}$ is *not* surjective, then possibly $x \notin \tilde{G}\tilde{Y}$, with x the solution of $Fx = y$.

REMARK 2. If $\tilde{G}$ is injective, then an $\tilde{F}: \tilde{G}\tilde{Y} \to \tilde{Y}$ exists such that $\tilde{F}\tilde{G} = I_{\tilde{Y}}$, where $I_{\tilde{Y}}$ is the identity operator on $\tilde{Y}$. Then $\tilde{F}$ is the left-inverse of $\tilde{G}$ and $\tilde{F}$ is "an approximation to F". However, we notice that $\tilde{F}$ is only defined on $\tilde{G}\tilde{Y}$ and not on X!

In a Defect Correction Process the solution of the original problem (P) *is found (or approximated) by the iterative application of one (or more) approximate inverse(s)* $\tilde{G}$.

In its most elementary form we have two versions of the defect correction process for the solution of (P):
the *first defect correction process*

$$\text{DCPA} \qquad \begin{cases} x_0 = \tilde{G}y, \\ x_{i+1} = (I - \tilde{G}F)x_i + \tilde{G}y, \end{cases}$$

and the *second* (or *dual*) *defect correction process*

$$\text{DCPB} \quad \begin{cases} \ell_0 = y, \quad x_i = \tilde{G}\ell_i \ , \\ \ell_{i+1} = (I - F\tilde{G})\ell_i + y. \end{cases}$$

REMARK 3. DCPA is completely described by $F,\tilde{G},y$ and x_0; DCPB is completely described by $F,\tilde{G},y$ and ℓ_0. With DCPA we use the fact that A is a linear space and not the fact that B is. With DCPB we use the fact that B is a linear space and not that A is.

REMARK 4. If $\tilde{G}$ is injective, then we can define its left-inverse $\tilde{F}$ and the DCPB can be shown to be equivalent with the iterative process

$$\text{DCPB}^* \quad \begin{cases} \tilde{F}x_0 = y, \\ \tilde{F}x_{i+1} = (\tilde{F} - F)x_i + y. \end{cases}$$

It is clear that, if $\hat{x}$ is a fixed point of the iteration DCPA then $\tilde{G}Fx = \tilde{G}y$. Hence, if $\tilde{G}$ is injective then $\hat{x}$ is a solution of the original problem (P). Also, if $\hat{\ell}$ is a fixed point of DCPB, then $F\tilde{G}\hat{\ell} = y$ and, hence, $\tilde{G}\hat{\ell}$ is a solution to (P).

If we consider the difference between the iterand x_i (resp. ℓ_i) and the fixed point $\hat{x}$ (resp. $\hat{\ell}$), then we notice that

$$x_{i+1} - \hat{x} = (I - \tilde{G}F)(x_i - \hat{x}) \ ,$$

and

$$\ell_{i+1} - \hat{\ell} = (I - F\tilde{G})(\ell_i - \hat{\ell}) \ .$$

Hence we call $M = I - \tilde{G}F$ the *amplification operator* (of the error) of DCPA and $\hat{M} = I - F\tilde{G}$ the amplification operator of DCPB. It is obvious that a sufficient condition for a DCP to converge to a fixed point is that the norm of its amplification operator is less than one.

In Section 4 we shall need the following relation between $\hat{M}$ and M, which follows immediately from the definition

$$\hat{M} = FMF^{-1}.$$

THEOREM 1. *If* $\tilde{G}$ *is an affine mapping, then the sequences* $\{x_i\}$ *in* DCPA *and* $\{x_i\}$ *in* DCPB *are identical.*

PROOF. Let $\{\ell_i\}_{i=0,1,2,\ldots}$ and $\{x_i\}_{i=0,1,2,\ldots}$ be defined as in DCPB, then:

1) $x_0 = \tilde{G}\ell_0 = \tilde{G}y$, and

2) $$\begin{aligned} x_{i+1} = \tilde{G}\ell_{i+1} &= \tilde{G}(\ell_i - F\tilde{G}\ell_i + y) \\ &= \tilde{G}\ell_i - \tilde{G}F\tilde{G}\ell_i + \tilde{G}y \\ &= x_i - \tilde{G}Fx_i + \tilde{G}y \\ &= (I - \tilde{G}F)x_i + \tilde{G}y; \end{aligned}$$

This means that the values from this sequence $\{x_i\}$ satisfy exactly the generation rules for the sequence $\{x_i\}$ from DCPA. Hence both sequences are identical. □

REMARK 5. It is clear from the proof of the last theorem that for a general mapping $\tilde{G}$ both processes DCPA and DCPB yield different sequences $\{x_i\}$.

A slight generalization of the DCPA, which is often more convenient for non-linear problems is the following defect correction process:

DCPC
$$\begin{cases} x_0 = \tilde{G}y \\ x_{i+1} = x_i + \mu\tilde{G}(\tilde{y} + (y - Fx_i)/\mu - \mu\tilde{G}\tilde{y}. \end{cases}$$

In this iteration step the parameters μ and $\tilde{y}$ are still free to choose.

REMARKS. With respect to this new defect correction process we notice:

1. Near a solution of $Fx = y$ the operator $\tilde{G}$ is applied only in the neighbourhood of $\tilde{y}$.
2. In the general case (i.e. for any μ and $\tilde{y}$) the solution of $Fx = y$ is a fixed point of DCPC.
3. With $\mu = -1$ and $\tilde{y} = y$, DCPC is identical with DCPA.
4. For arbitrary μ and $\tilde{y}$, with $\tilde{G}$ affine DCPC is identical with DCPA (and hence also with DCPB).
5. The amplification factor of DCPC is given by

$$\frac{\|x_{i+1}-\hat{x}\|}{\|x_i-\hat{x}\|} \leq \|I - \tilde{G}'F'\| + \|\tilde{G}'\|\,\|F^*\| + \|\tilde{G}^*\|\,\|F'\| + \|\tilde{G}^*\|\,\|F^*\|,$$

where $\tilde{G}'$ and $\tilde{G}^*$ are defined by

$$\tilde{G}(\tilde{y} + \delta) - \tilde{G}(\tilde{y}) = \tilde{G}'\delta + \tilde{G}^*\delta,$$

with $\tilde{G}'$ linear and $\tilde{G}^*$ such that

$$\|\tilde{G}^*\delta\| = o(\|\delta\|) \text{ as } \delta \to 0 ,$$

and F' and F^* defined analogously.

We conclude this section with some examples of defect correction processes.

EXAMPLE 1. Iterative methods for the solution of linear systems.

Many of the well-known iterative methods for the solution of linear systems can easily be recognized as a defect correction process. For all these methods $\tilde{G}$ is linear and, hence, the three variants are identical. Here we shall identify as a DCP a number of these methods for the solution of the square linear system $Ax = b$.

1.1 The Jacobi method.

The Jacobi-method:

$$\text{diag}(A)\, x_{i+1} = b + (\text{diag}(A) - A)x_i,$$

can be written as

$$x_{i+1} = x_i + \tilde{G}(b - Ax_i) = (I - \tilde{G}A)x_i + \tilde{G}b,$$

with

$$\tilde{G} = [\text{diag}(A)]^{-1}.$$

1.2 The Gauss-Seidel method.

Let A be decomposed as $A = L + U$, where U is strictly upper-triangular

and L is lower triangular; then the Gauss-Seidel process reads

$$L\, x_{i+1} = b - Ux_i,$$

i.e. a DCP with $\tilde{G} = L^{-1}$.

1.3 The relaxation methods JOR, SOR, RF and GRF.

All "stationary fully consistent iterative methods of degree one" for the solution of $Ax = b$ can be written as

$$x_{i+1} = x_i - P(Ax_i - b),$$

where P is a non-singular matrix (cf YOUNG [1971]). With $P = pI$, p a scalar and I the identity matrix it is a stationary Richardson method (RF); with P a non-singular diagonal matrix it is a Generalized stationary Richardson method (GRF); with $P = \omega\tilde{G}$, $\tilde{G}$ as under 1.1 it is a Jacobi relaxation method (JOR) and with $P = \omega\tilde{G}$, $\tilde{G}$ as under 1.2 it is a SOR method.

EXAMPLE 2. Modified Newton iteration.

In this case the problem (P) is the solution of a non-linear equation

$$Fx = y,$$

with a Fréchet-differentiable operator F. The Fréchet-derivative $F'(x)$ is approximated by a non-singular linear operator E. The relation

$$Fx - Fx_i = F'(x_i)(x-x_i) + o(\|x-x_i\|),$$

or equivalently,

$$x - x_i = (F'(x_i))^{-1}(y - Fx_i + \sigma(\|x-x_i\|)),$$

suggests the modified Newton iteration:

$$x_{i+1} = x_i + E^{-1}(y - Fx_i).$$

Clearly, this is a DCPA with $\tilde{G} = E^{-1}$.

We notice that in a proper Newton process (not the modified Newton iteration) the approximate Fréchet-derivative E is updated during the iteration process. This kind of generalization of the elementary DCP will be treated in Section 2.

EXAMPLE 3. An analytic example.

We consider the two-point boundary-value problem (cf. STETTER [1978])

$$(*) \qquad \begin{cases} x'' - e^{x} = 0 & \text{on } (-1,+1) \\ x(-1) = x(+1) = 0. \end{cases}$$

This defines the problem

$$Fx = 0,$$

where

$$F: C_0^2[-1,+1] \to C(-1,+1).$$

We construct an approximate problem, replacing e^{x} by $0.99 + 0.81x$ (i.e. a reasonable approximation if $-0.4 \le x \le 0.0$). Thus we get the approximate problem $\tilde{F}x = y$, viz.

$$\begin{cases} x'' - 0.81x - 0.99 = y & \text{on } (-1,-1) \\ x(-1) = x(+1) = 0. \end{cases}$$

Hence, we can write the solution of $\tilde{F}x = y$ as

$$x(t) = \int_{-1}^{+1} K(t,z)(y(z) + 0.99)\, dz,$$

for some suitable kernel-function $K(t,z)$. This integral operator defines an approximate inverse $\tilde{G}$ for the problem (*). With this $\tilde{G}$ we can construct a DCP to find the solution of (*).

2. EXTENSION OF THE DCP PRINCIPLE

In this section we shall extend the idea of the defect correction process in several ways: we allow different approximate inverses to serve in one iteration process and we consider a sequence of problems that converges to a final problem of which the solution is wanted. We also consider the process obtained when a fixed combination of approximate inverses is used all over in a defect correction process.

2.1 Non-stationary defect correction processes.

In order to find a solution to the problem (P) it is not necessary to use one fixed approximate inverse in an iteration process as described in the preceeding section. As we anticipated in the example with Newton's method, it is possible to use another approximate inverse in each iteration step. Then the iteration steps in DCPA and DCPB read respectively

$$x_{i+1} = x_i - \tilde{G}_i F x_i + \tilde{G}_i y$$

and

$$\ell_{i+1} = \ell_i - F\tilde{G}_i \ell_i + y.$$

A similar modification for DCPC can be given.

Various methods are known to find a proper sequence of $\{\tilde{G}_i\}$. Here we mention a few.

EXAMPLE 1. $\tilde{G}_i = \tilde{G}(x_{i-1})$.

The approximate inverse depends on the last iterand computed. This is the case e.g. in Newton's method for the solution of non-linear equations, where $\tilde{G}(x) = (F'(x))^{-1}$. $F'(x)$ is the Fréchet derivative of the operator in the problem (P).

EXAMPLE 2. $\tilde{G}_i = \tilde{G}(\omega_i)$.

The approximate inverse depends on a single real parameter. This is the case e.g. in non-stationary relaxation processes for the solution of linear systems.

EXAMPLE 3. $\tilde{G}_i \in \{\tilde{G}_1, \tilde{G}_2\}$.

In each iteration step the approximate inverse is chosen out of a set of two (or more) fixed approximate inverses. This is the case e.g. in Brakhage's and Atkinson's methods for the solution of Fredholm integral equations of the 2nd kind. (See ATKINSON [1976] and BRAKHAGE [1960].)

2.2 A fixed combination of approximate inverses.

We consider two iteration steps in the non-stationary DCPA in which, in turn, one or the other of two approximate inverses is used. The iteration steps

$$x_{i+\frac{1}{2}} = (I - \tilde{G}F)x_i + \tilde{G}y$$
$$x_{i+1} = (I - \tilde{\tilde{G}}F)x_{i+\frac{1}{2}} + \tilde{\tilde{G}}y$$

combine into a single iteration step of the form

$$x_{i+1} = (I - \tilde{\tilde{G}}F)(I - \tilde{G}F)x_i + (\tilde{\tilde{G}} - \tilde{\tilde{G}}F\tilde{G} + \tilde{G})y.$$

This is easily recognized as a new iteration step of the type DCPA, now with the approximate inverse

$$\hat{G} = \tilde{\tilde{G}} - \tilde{\tilde{G}}F\tilde{G} + \tilde{G}.$$

We conclude that a fixed combination of DCPA-steps can be considered as a new DCPA-step with a more complex approximate inverse. The amplification operator of the new DCPA process is the product of the amplification operators of the elementary processes.

REMARK. Generally the above observation with respect to DCPA does not hold for DCPB processes.

2.3 σ applications of the same approximate inverse.

In order not to make the notation unnecessarily intricate, from now on we shall only consider linear problems, unless explicitly stated otherwise.

We can describe the DCPA in matrix notation by

$$\begin{pmatrix} x_{i+1} \\ y \end{pmatrix} = \begin{pmatrix} I-\tilde{G}F & \tilde{G} \\ \emptyset & I \end{pmatrix} \begin{pmatrix} x_i \\ y \end{pmatrix} \quad .$$

σ times an application of the same iteration step yields

$$\begin{pmatrix} x_{i+\sigma} \\ y \end{pmatrix} = \begin{pmatrix} I-\tilde{G}F & \tilde{G} \\ \emptyset & I \end{pmatrix}^{\sigma} \begin{pmatrix} x_i \\ y \end{pmatrix} = \begin{pmatrix} (I-\tilde{G}F)^{\sigma} & \sum_{m=0}^{\sigma-1} (I-\tilde{G}F)^{m}\tilde{G} \\ \emptyset & I \end{pmatrix} \begin{pmatrix} x_i \\ y \end{pmatrix} .$$

Thus, one iteration step which consists of σ applications of DCPA-steps results in a DCPA with the amplification operator

$$M = (I-\tilde{G}F)^{\sigma}$$

and the approximate inverse

$$\hat{G} = \sum_{m=0}^{\sigma-1} (I-\tilde{G}F)^{m}\,\tilde{G} = [I - (I-\tilde{G}F)^{\sigma}]F^{-1} .$$

2.4 Iterative application of DCP.

It is possible not only to change the approximate inverse $\tilde{G}$ during the iteration process, often it makes sense also to substitute different operators F_i for F during iteration. In general, the operators $\{F_i\}$ will be simple to evaluate in the beginning of the iteration and they will converge to F, the operator in the original problem, as the iteration proceeds.

One example of such a process is the IUDeC (Iteratively Updated Defect Correction) process described by STETTER [1978]. Here $\{F_i\}$ are discrete approximations of higher and higher order to an analytic operator F. The approximate inverse $G = F_0^{-1}$ is kept constant during the process. An analysis of this kind of process is given in Section 3.3, when we have introduced discretizations.

Another example is the Full Multigrid method [BRANDT, 1979] in which $\{F_i\}$ are discretizations on finer and finer nets of an analytic operator F.

2.5 Recursive application of DCP.

Generally, the evaluation of the approximate inverse operator $\tilde{G}_i$ implies the solution of an equation which is (essentially) of a simpler type than the original equation. However, also this simpler equation may be of a kind that we want to solve by means of a DCP. For this we need an even simpler equation to solve, etc.. Thus, the execution of a single iteration step may imply the activation of a new (simpler to solve) DCP. In this way we can construct a recursive construction of DCPs in which only on the lowest level of recursion a very simple equation is to be solved.

Independently, this is probably not a real meaningful construction, but in combination with non-stationary processes, where also other (non-recursive) approximate inverses are available, it describes the essentials of the multigrid algorithm.

Such a combination of a non-stationary process with some recursive approximate inverses can be described by the following sequence of DCPs.

DCP_1: $x := x - \tilde{G}_1 (F_1 - f_1)$ $\quad \tilde{G}_j$ fixed, $\quad j=1,2,\ldots,n$,

DCP_2: $x := x - \tilde{G}_{2,i} (F_2 - f_2)$

$\quad\vdots \qquad\qquad\vdots \qquad\qquad \tilde{G}_{j,i} \in \{\tilde{G}_j, F_{j-1}^{-1}\},$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j=2,3,\ldots,n.$

DCP_n: $x := x - \tilde{G}_{n,i} (F_n - f_n)$

A full use of the sequence of DCPs is made by combining also the iterative application: first DCP_1 is solved and its solution is used as a starting value for DCP_2 etc.. In a multigrid context

$$DCP_1, DCP_2, \ldots, DCP_n ,$$

are processes to solve operator equations, discretized on finer and finer grids. The complete iterative process is called: Full Multigrid Algorithm [BRANDT,1979].

3. DISCRETIZATION ON RELATED GRIDS, RELATED DISCRETIZATIONS.

In this section we give definitions for related discretizations of spaces and problems and we define relative order of approximation,

consistency and convergence between related discretizations. In Section 3.3 we give an approximation theorem for successive approximations in the iterative application of a DCP.

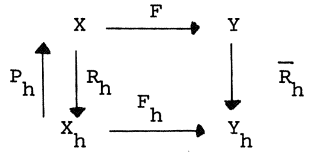
3.1 Discretization of spaces and operators.

Let's be given a problem $Fx = y$, where $F:X \to Y$ and $y \in Y$ are given and where X and Y are (infinite dimensional) vector spaces. The problem is discretized by associating it with a problem $F_h x_h = y_h$, where $F_h: X_h \to Y_h$ and $y_h \in Y_h$ are given and X_h and Y_h are finite dimensional vector spaces. By selecting $h \in H$ (H an index-set) different discretizations of the same problem are possible.

A relation between the problem and its discretization is obtained by introducing surjections $R_h: X \to X_h$ and $\bar{R}_h: Y \to Y_h$. (Notice that $\dim(X) \geq \dim(X_h)$, $\dim(Y) \geq \dim(Y_h)$ and, in most cases, X_h and Y_h are selected such that $\dim(X_h) = \dim(Y_h)$.)

In order to interpret the solution of the discretized problem as an approximation to the solution of the original problem, we have to define an injection $P_h: X_h \to X$.

The mappings P_h are called *prolongations*, the mappings R_h and $\bar{R}_h$ are *restrictions*. The relation between the different spaces and mappings is summarized in the following diagram



$h \in H$

DEFINITION. Given the discretization of the spaces X and Y by X_h, Y_h, P_h, R_h and $\bar{R}_h$, $h \in H$, we can associate with the problem $Fx = y$ its *canonical discretization* $F_h x_h = y_h$ by defining $F_h = \bar{R}_h F P_h$ and $y_h = \bar{R}_h y$.

DEFINITION. Given two discretizations of the spaces X and Y by $(X_h, Y_h, P_h, R_h, \bar{R}_h)$ and $(X_H, Y_H, P_H, R_H, \bar{R}_H)$, $h, H \in H$, these are called *related discretizations* if surjective mappings R_{Hh} and $\bar{R}_{Hh}$ and an injection P_{hH} exist such that

$$
\begin{aligned}
R_{Hh} &: X_h \to X_H, & R_{Hh}\, R_h &= R_H,\\
\bar{R}_{Hh} &: Y_h \to Y_H, & \bar{R}_{Hh}\, \bar{R}_h &= \bar{R}_H,\\
P_{hH} &: X_H \to X_h, & P_h\, P_{hH} &= P_H .
\end{aligned}
$$

It should be clear that $\dim(X_H) \le \dim(X_h)$ and $\dim(Y_H) \le \dim(Y_h)$. We see also that, if two discretizations (with $h,H \in \mathcal{H}$) of the spaces X and Y are related, then the *coarse discretization* (with $H \in \mathcal{H}$) can be considered as a discretization of the *fine discretization* (with $h \in \mathcal{H}$) of the finite dimensional spaces X_h and Y_h.

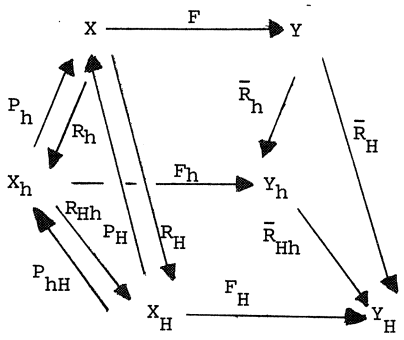
From the definitions it follows immediately that, if the coarse discretization $F_H\, x_H = y_H$ and the fine discretization $F_h\, x_h = y_h$ are both canonical discretizations of the same problem $Fx = y$, we have

$$F_H = \bar{R}_{Hh}\, F_h\, P_{hH} \quad \text{and} \quad y_H = \bar{R}_{Hh}\, y_h .$$

Because P_h is an injection, it has a left-inverse $\hat{R}_h$ such that $\hat{R}_h\, P_h$ is the identity operator on X_h; because R_h and $\bar{R}_h$ are surjective, right-inverses $\hat{P}_h$ and $\bar{P}_{\hat{h}}$ exist such that $R_h\, \hat{P}_h : X_h \to X_h$ and $\bar{R}_h\, \bar{P}_{\hat{h}} : Y_h \to Y_h$ are identity operators. From these definitions of $\hat{R}_h$, $\hat{P}_h$ and $\bar{P}_{\hat{h}}$ follows:

$$
\begin{aligned}
R_{Hh} &= R_H\, \hat{P}_h ,\\
\bar{R}_{Hh} &= \bar{R}_H\, \bar{P}_{\hat{h}} ,\\
P_{hH} &= \hat{R}_h\, P_H .
\end{aligned}
$$

The relation between the different spaces is summarized in the following diagram.



It is important to notice that, in general, differeht norms can be used to (trans-) form the above mentioned vector spaces into normed vector spaces or Banach spaces. Indeed, each of the above vector spaces, say Z, can be formed into a scale of normed vector spaces $\{Z^\alpha\}$, $\alpha \in \mathbb{R}$, with $Z^\alpha = Z$ and norms $\|\cdot\|_\alpha$ such that with $u \in Z$ we have

$$\|u\|_\alpha \leq \|u\|_\beta \quad \text{if } \alpha \leq \beta.$$

DEFINITIONS. An operator $F : X \to Y$ is called *bounded* if

$$\|F\|_{X^\alpha \to Y^\alpha} \leq C \qquad \text{uniformly in } \alpha,$$

and σ-*stable* if

$$\|F^{-1}\|_{Y^\alpha \to X^{\alpha-\sigma}} \leq C \qquad \text{uniformly in } \alpha.$$

In the following we shall assume that all restrictions and prolongations and their right- resp. left-inverses are bounded, uniformly in $h \in H$. The conditions on the inverses imply for the prolongations P_h that

$$\inf_{v \neq 0} \frac{\|P_h v\|}{\|v\|} > C > 0 \quad , \qquad C \text{ independent of } h \in H,$$

and for the restrictions R_h that

$$\inf_{\substack{w \in R_h \\ w \neq 0}} \sup_{\{v \mid R_h v = w\}} \frac{\|R_h v\|}{\|v\|} > C > 0, \quad C \text{ independent of } h \in H.$$

To each discretization, characterized by $h \in H$, a mesh-size $m(h) > 0$ is associated. Discretizations X_h and X_H of X with $\dim(X_h) \geq \dim(X_H)$ generally have mesh-sizes related by $m(h) < m(H)$. If no confusion is possible we denote $m(h)$ simply by h. Often we consider infinite sequences $\{X_h\}$ with $h > 0$ and $\lim_{h \to 0} \dim(X_h) = \infty$.

3.2 Relative consistency and convergence

DEFINITIONS. A sequence of discretizations of X characterized by $(X_h, P_h, R_h)_{h>0}$ is called *convergent* if

$$\lim_{h\to 0} \| I - P_h R_h \| = 0;$$

the *order of approximation* is p if

$$\| I - P_h R_h \| = O(h^p) \text{ for } h \to 0.$$

DEFINITION. A sequence of discretizations of a problem Fx = y is *consistent* if

$$\lim_{h\to 0} \| F_h R_h - \bar{R}_h F \| = 0;$$

its *order of consistency* is p if

$$\| F_h R_h - \bar{R}_h F \| = O(h^p) \text{ for } h \to 0.$$

DEFINITION. A sequence of discretizations of a problem Fx = y is *σ-stable* if $F_h^{-1}: Y_h^{\alpha} \to X_h^{\alpha-\sigma}$ is bounded uniformly in h and α. It is called *stable* if it is 0-stable.

DEFINITION. A sequence of discretizations of a problem Fx = y is *convergent* if

$$\lim_{h\to 0} \| F^{-1} - P_h F_h^{-1} \bar{R}_h \| = 0$$

its *order of convergence* is p if

$$\| F^{-1} - P_h F_h^{-1} \bar{R}_h \| = O(h^p) \text{ for } h \to 0.$$

Analogously, for related discretizations characterized by H > h > 0, we can define the corresponding relative properties (without reference to the original problem), i.e.

the *relative order of approximation* p:

$$\| I_h - P_{hH} R_{Hh} \| = O(H^p),$$

the *relative order of consistency* p

$$\| F_H R_{Hh} - \bar{R}_{Hh} F_h \| = O(H^p),$$

the *relative order of convergence* p:

$$\| F_h^{-1} - P_{hH} F_H^{-1} \bar{R}_{Hh} \| = O(H^p).$$

THEOREM. *If two related discretizations of the same problem are consistent of order* p_1 *and* p_2 *respectively, then they are relatively consistent of the order* $\min(p_1,p_2)$.

PROOF. The simple proof is left to the reader.

NOTE 1. The following identity is useful if we consider DCPs with related discretizations

$$I_h - P_{hH} F_H^{-1} \bar{R}_{Hh} F_h = (I_h - P_{hH} R_{Hh}) + P_{hH} F_H^{-1} (F_H R_{Hh} - \bar{R}_{Hh} F_h).$$

NOTE 2. Let $F_h x_h = y_h$ and $F_H x_H = Y_H$ be two related canonical discretizations of the same problem, then, for any restriction $\tilde{R}_{Hh} : X_h \to X_H$ we have

$$I_h - P_{hH} F_H^{-1} \bar{R}_{Hh} F_h = (I_h - P_{hH} F_H^{-1} \bar{R}_{Hh} F_h)(I_h - P_{hH} \tilde{R}_{Hh}).$$

3.3 The accuracy of successive approximations in a DCP iteration with different discretizations of the same problem.

Let us consider (different) discretizations of the problem Fx = y, viz.

$$F_h^i x_h = y_h \text{ , with } F_h^i : X_h \to Y_h \text{ for all } i = 0,1,2,\ldots;$$

and let X, X_h, Y and Y_h be related by

$$R_h : X \to X_h \text{ and } \bar{R}_h : Y \to Y_h.$$

Let the order of consistency of the discretizations be p_i, and let the first discretization be stable. We will study the iterative application of DCPA, with the equations $F_h^i x_h = y_h = \bar{R}_h y$ to solve in the i-th iteration step and with the same approximate inverse $\tilde{G}_h = (F_h^o)^{-1}$ in all iteration steps. Then the DCPA reads

$$\begin{cases} u_1 = \tilde{G}_h y_h = \tilde{G}_h \bar{R}_h\, y \\ u_{i+1} = (I_h - \tilde{G}_h F_h^i)\, u_i + \tilde{G}_h\, y_h\ . \end{cases}$$

We are going to estimate the relative error of approximation for a finite number of iteration steps:

$$k_i = \| u_i - R_h\, x \| \,/\, \|x\| .$$

THEOREM. *For the relative error of approximation in the i-th iteration step of the iterative DCPA process:*

$$k_i = \| u_i - R_h\, x \| \,/\, \|x\|,$$

we have

$$k_0 = \|\tilde{G}_h\|\, \|\bar{R}_h F - F_h^0 R_h\| = O(h^{p_0})$$

$$k_i = \|\tilde{G}_h\|\, \|\bar{R}_h F - F_h^{i-1} R_h\| + \|\tilde{G}_h\|\, \|F_h^1 - F_h^{i-1}\|\, k_{i-1}$$

$$= O(h^{\min_{0\le j\le i}(p_j + (i-j)p_0)}), \quad i = 1,2,\ldots\ .$$

PROOF.

$$u_0 - R_h x = \tilde{G}_h \bar{R}_h\, y - R_h\, x = \tilde{G}_h(\bar{R}_h F - F_h^o R_h)\, x.$$

The given estimate now follows from the stability of F_h^o (i.e. $\tilde{G}_h$ is uniformly bounded) and the consistency of F_h^o.

$$u_{i+1} - R_h\, x = u_i - R_h x - \tilde{G}_h F_h^i\, u_i + \tilde{G}_h\, y_h$$

$$= u_i - R_h x + \tilde{G}_h(\bar{R}_h F x - F_h^i R_h x + F_h^i R_h x - F_h^i u_i)$$

$$= (I_h - \tilde{G}_h F_h^i)(u_i - R_h x) + \tilde{G}_h(\bar{R}_h F - F_h^i R_h)x\ .$$

Hence, for i = 0,1,2,...,

$$k_{i+1} \leq \| I_h - G_h F_h^i \| k_i + \|\tilde{G}_h\| \quad \|\bar{R}_h F - F_h^i R_h\|.$$

Here again, the estimate follows from the stability of F_h^0 and the consistency of F_h^i. □

COROLLARY. *If*

$$\begin{cases} p_i \geq i\, p_0 & (i<n) \\ p_i = p_n & (i\geq n) \end{cases}$$

then

$$k_i = O(h^{\min(p_n, i.p_0)}) .$$

4. MULTIGRID ALGORITHMS

In this section we shall describe multigrid algorithms and the structure of their convergence theorems. First we consider a simple form of the multigrid algorithm, "the two-level algorithm" (or TLA), and show how its convergence is proved. Then we show the multi-level algorithm (MLA), which is the recursive application of the two-level algorithm. At the end we show how multigrid algorithms are applied to non-linear problems.

The problems that are solved by multigrid methods are all discretizations of a continuous problem $Lx = f$. The methods find solutions to the finest discretization $L_h x_h = f_h$ by means of discretizations on coarser grids, which we denote by $L_H x_H = f_H$.

4.1 The two-level algorithm.

The two-level algorithm is a non-stationary defect correction process in which only two different approximate inverses are used:
(1) some *relaxation method* (e.g. Jacobi, Gauss-Seidel or the incomplete LU-decomposition, see example 1 Section 1) on the fine grid and
(2) a *coarse grid correction*.

The approximate inverse in the coarse grid correction that is used to solve the discrete problem $L_h x_h = f_h$ is given by $\tilde{G}_i = P_{hH} L_H^{-1} \bar{R}_{Hh}$. Thus, one coarse grid correction step in the two-level algorithm reads

$$x_{i+1} = x_i + P_{hH} L_H^{-1} \bar{R}_{Hh} (f_h - L_h x_i).$$

One step in the two-level algorithm, now consists of p relaxation sweeps of the relaxation method chosen, a coarse grid correction step and again q relaxation sweeps of the relaxation method. Such a step of the two-level algorithm is described in the following ALGOL-like program:

```
proc two level algorithm = (ref gridf u, gridf f) void:
begin
      for i to p
      do relax (u,f) od;

      d := restrict (Lh u - f);
      solve (v,d);                        # solves L_H v = d #
      u := u - prolongate v ;

      for i to q
      do relax (u,f) od
end;
```

Clearly, the amplification operator of one step of the two-level algorithm is given by

$$M_h^{TLA} = (I - B_h L_h)^q \, (I - P_{hH} L_H^{-1} \bar{R}_{Hh} L_h)(I - B_h L_h)^p,$$

where B_h is the approximate inverse of the relaxation process. In this expression we recognize the relative convergence operator and the amplification operators of the relaxation process:

$$M_h^{REL} = (I - B_h L_h),$$
$$\hat{M}_h^{REL} = (I - L_h B_h),$$

and we can write

$$M_h^{TLA} = (M_h^{REL})^q \, (L_h^{-1} - P_{hH} L_H^{-1} R_{Hh}) (\hat{M}_h^{REL})^p L_h,$$

or

$$\hat{M}_h^{TLA} = L_h (M_h^{REL})^q (L_h^{-1} - P_{hH} L_H^{-1} R_{Hh}) (\hat{M}_h^{REL})^p .$$

The structure of the convergence proof for the two-level algorithm is as follows:

Assuming that

(1) the two discrete operators are relatively convergent of order α,

(2) the relaxation satisfies a *proper smoothing property* of order at least α, i.e. $\exists\ C_0(p) > 0$, independent of h, such that $\| (\hat{M}_h^{REL})^p L_h \| < C_0(p) h^{-\alpha}$ and $\lim_{p\to\infty} C_0(p) = 0$

(3) the amplification operator $(M_h^{REL})^q$ is bounded,

(4) the mesh-ratio m(H)/m(h) is bounded, uniformly in h,

then the two-level algorithm converges for p large enough.

PROOF.

$$\| M_h^{TLA} \| \leq \| (M_h^{REL})^q \| \, \| L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh} \| \, \| (\hat{M}_h^{REL})^p L_h \|$$

$$\leq C \cdot C\ (m(H))^{\alpha} \cdot C_0(p)\ (m(h))^{-\alpha}$$

$$= C \cdot C_0(p)\ (m(H)/m(h))^{\alpha} \leq C \cdot C_0(p)$$

Since $C_0(p) \to 0$ for $p \to \infty$ we see that $\| M_h^{TLA} \| < 1$ for p large enough. □

REMARK. In an actual convergence proof the norms in the relevant spaces should be specified and the assumptions should be verified for the particular algorithm under consideration. We have to realize that, apart from the above mentioned structure, the two-level algorithm is determined by the particular discretizations L_h and L_H, by the restrictions and prolongations $\bar{R}_{Hh}$ and P_{hH} and by the particular relaxation method used (characterized by B_h).

If the discretizations L_h and L_H are related canonical discretizations, then we can make use of the relations in the notes 1 and 2 of Section 3.2.

4.2 The multi-level algorithm.

Whereas for the two-level algorithm we have to evaluate L_H^{-1}, i.e. we have to solve a discretized problem on a coarse grid, in the multi-level algorithm we approximate this solution by application of a number of

iteration steps of the same algorithm on the coarse level. As was explained in Section 2.5 we now only have to solve directly a discretized problem on the very coarsest grid. If σ iteration steps of the multi-level algorithm are used to approximate L_H^{-1}, this multi-level algorithm is described in the ALGOL-like program:

```
proc multi level algorithm = (ref gridf u, gridf f) void:
if level of u = 0
then solve (u,f) # on the coarsest grid #
else
      for i to p while ...
      do relax (u,f) od;

      d := restrict (f - Lh u); v := 0;
      for : to σ while ...
      do multi level algorithm (v,d) od;
      u := u + prolongate v

      for i to q while ...
      do relax (u,f) od

fi;
```

By while ... we denote in the program that some iterations may be terminated sooner, depending on the speed of convergence or other conditions that can be checked during the computation. Multigrid algorithms that make use of this possibility are said to have an *adaptive strategy*, algorithms where the iterations are controled only by the fixed numbers p, σ and q are said to have a *fixed strategy*. Although the adaptive strategy may be very efficient (cf. BRANDT, 1979), the fixed strategy is better accessible for a theoretical analysis.

For some fixed strategies, we show in figure 1 how it is switched between the different levels of discretization. We see that - essentially - most relaxation sweeps are performed on the lower levels.

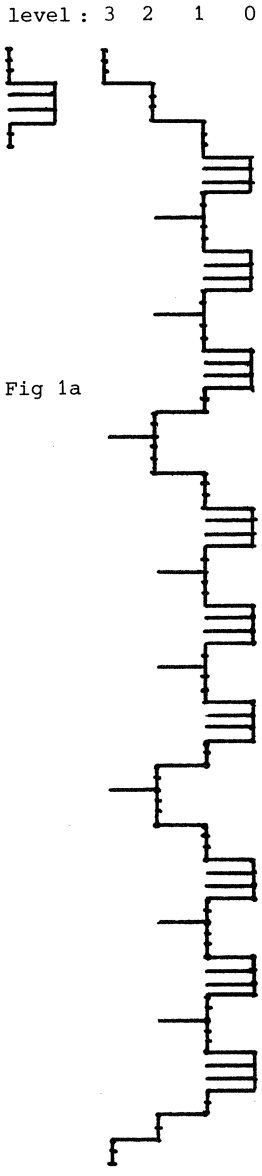


Fig 1a

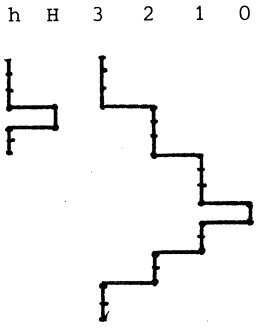


Fig 1b

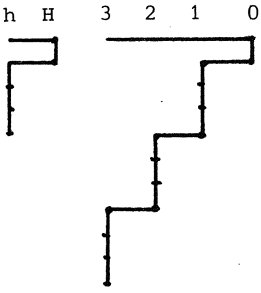


Fig 1c

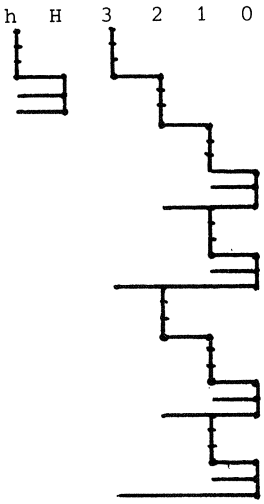


Fig 1d

Figure 1. The recursive structure of multigrid algorithms with a fixed strategy.

In all diagrams the number of levels is 3, the very coarsest level is denoted by 0. In each diagram 1a, 1b, 1c or 1d, the basic structure on the levels h and H is given as well as the recursive structure of one iteration step on level 3. Tick marks on a level > 0 denote the execution of a relaxation step on this level, a tick-mark on level 0 denotes the direct solution on the very coarsest level. The different structures shown are:

1a. A general structure with $p = 3$, $\sigma = 3$, and $q = 2$.

1b. A structure with $\sigma = 1$ (NICOLAIDES, 1979) $p = 3$, $q = 2$.

1c. A structure with $\sigma = 1$, $p = 0$ (FREDERICKSON, 1975) $q = 3$.

1d. A structure with $q = 0$ (HACKBUSCH, 1979) $p = 3$, $\sigma = 2$.

The amplification operator of a multi-level iteration step on the h-level of discretisation we denote by M_h^{MLA}, this amplification operator on the next coarser level we denote by M_H^{MLA}. The approximate inverse of the coarse grid correction in the multigrid algorithm is not given by L_H^{-1}, but it is obtained by σ steps in the DCP for the approximation of L_H^{-1}. The amplification operator of such a single DCP-step is given by M_H^{MLA}. Hence, the approximate inverse of the σ iteration steps together is given by (see Section 2.3):

$$(I - (M_H^{MLA})^{\sigma}) L_H^{-1}.$$

Consequently, the amplification operator of the coarse grid correction is

$$(I - P_{hH}(I - (M_H^{MLA})^{\sigma})\, L_H^{-1}\, \bar{R}_{Hh}\, L_h)$$

and we have

$$M_h^{MLA} = (M_h^{REL})^q (I - P_{hH}(I - (M_H^{MLA})^{\sigma}) L_H^{-1}\, \bar{R}_{Hh}\, L_h)(M_h^{REL})^p$$

$$= M_h^{TLA} + (M_h^{REL})^q\, P_{hH} (M_H^{MLA})^{\sigma}\, L_H^{-1}\, \bar{R}_{Hh} (\hat{M}_h^{REL})^p L_h$$

or

$$\hat{M}_h^{MLA} = \hat{M}_h^{TLA} + L_h (M_h^{REL})^q\, P_{hH} (M_H^{MLA})^{\sigma} L_H^{-1}\, \bar{R}_{Hh} (\hat{M}_h^{REL})^p.$$

Therefore, if the (coarse) discretized operator L_H is stable and the assumptions (2) and (3) of Section 4.1 hold, then

$$\|M_h^{MLA}\| \leq \|M_h^{TLA}\| + C\|M_H^{MLA}\|^{\sigma}.$$

Here we get a recursive expression, where the rate of convergence of the MLA on the level h is expressed in the rate of convergence of the TLA and the rate of convergence of the MLA on the next coarser level H. Further we notice that on the coarsest level we have $M_0^{MLA} = M_0^{TLA}$.

On each level we have $\|M_h^{TLA}\| \leq \rho < 1$ if p is large enough, hence we can find a σ such that $\|M_H^{MLA}\| < 1$. Often a small value of σ (e.g. σ=2) can be shown to be sufficient to have $\|M_h^{MLA}\| \leq \rho < 1$ on all levels, ρ independent of h.

4.3 The non-linear multi-level algorithm.

The multi-level algorithm in Section 4.2 essentially used the fact that the operator L and its discretizations are linear. By a slight change of the algorithm we can adapt it for nonlinear problems. For this purpose we make use of the DCPC as treated in Section 1. We describe the nonlinear algorithm - again - in an ALGOL-like program

```
proc non linear mla = (ref grid u, gridf f ) void:
if level of u = 0
then solve (u,f)
      # e.g. by a Newton type method #
else
      for i to p
      do relax (u,f) od;

      y := w := restrict u;
      d := LH y + restrictbar (f - Lh u)/mu ;
      for i to σ
      do nonlinear mla (w,d) od;
      u := u + mu × prolongate (w-y);

      for i to q
      do relax (u,f) od
fi;
```

Here, of course, the relaxation should be of a non-linear type. The coarse grid correction of the TLA corresponding with this MLA (i.e. the

MLA with $\sigma = \infty$) is here

$$x_{i+1} = x_i + \mu\, P_{hH}(L_H^{-1}(L_H R_{Hh}\, x_i + \bar{R}_{Hh}(f_h - L_h\, x_i)\,/\,\mu) - R_{Hh}\, x_i).$$

This can be recognized as the DCPC in Section 1, with $\tilde{y}$ such that $L_H\, R_{Hh} x_i = R_{Hh}\tilde{y}$.

If we fit the nonlinear mla-step into a Full Multigrid Method (see Section 2.4), then we may replace $R_{Hh}\, x_i$ (i.e. the best approximation of the solution that is available at the level H) by the last solution obtained on the next coarser grid. In that case, there is no need for recomputing y and LH y in each call of the nonlinear mla.

5. EXAMPLES OF MULTIGRID METHODS

In this section we give two examples of multigrid methods. In the first example we show Fredericson's method for the solution of a differential equation and in the second we treat a multigrid method for the solution of a Fredholm integral equation of the 2nd kind. The essential difference between both problems is that a regular differential operator, $L : A \to B$, maps a space with a stronger into a space with a weaker topology, whereas a compact integral operator, $K : A \to B$, maps a space with a weaker into one with a stronger topology. The effect is, that for the differential equation we can get an amplification factor $\|M_h^{MLA}\|$ which is bounded by a constant (less than one) uniformly in h. We call this a *multigrid method of the first kind*. For the integral equation we can get an amplification factor $\|M_h^{MLA}\|$ which is bounded by a constant of order $O(h^m)$ for some $m > 0$. This we call a *multigrid method of the second kind*.

REMARK. With Jacobi-type iteration similar differences are found for the two different problems: for the differential equation we have the bound $\|M_h^{REL}\| \le 1 - Ch^{2m}$ and for the integral equation the bound is $\|M_h^{REL}\| \le C < 1$ as $h \to 0$. These bounds also clearly show the supremacy of the MLA-iteration over the classical iteration methods.

5.1 The multigrid method of Fredericson for the solution of a differential equation.

For Fredericson's multigrid method we have $p = 0$ and $\sigma = 1$. Because

of $\sigma = 1$ the amplification operator is much simpler than in the general case. For a 3-level method (see figure 1.c) this operator is given by

$$\begin{aligned} M_3^{MLA} = & (I-B_3L_3)^q && (L_3^{-1}-PL_2^{-1}R)\ L_3 \\ & + (I-B_3L_3)^q\ P(I-B_2L_2)^q && (L_2^{-1}-PL_1^{-1}R)R\ L_3 \\ & + (I-B_3L_3)^q\ P(I-B_2L_2)^q\ P(I-B_1L_1)^q && (L_1^{-1}-PL_0^{-1}R)RRL_3, \end{aligned}$$

where L_i is the discretized operator at level i, $(I-B_iL_i)$ is the amplification operator of the relaxation at level i, and P and R are the prolongation and restriction operators between the various levels.
First we look at the first term of this operator:

$$(I-B_3L_3)^q(I-PL_2^{-1}RL_3)$$

Here $(I - PL_2^{-1}RL_3)$ reduces the low frequences in the error and $(I-B_3L_3)^q$ reduces the high frequencies in the error of the approximation to the solution. This can be seen e.g. if L_2 and L_3 are related canonical discretizations: $L_2 = RL_3P$. Then the first term can be rewritten as

$$(I-B_3L_3)^q(I-PL_2^{-1}RL_3)(I-P\tilde{R})$$

If $\tilde{R}$ denotes restriction to gridpoints and P denotes piecewise polynomial interpolation of degree k-1 then it is clear that for $I - P\tilde{R} : H^k \to H^0$ we have $\|I-P\tilde{R}\|_{H^k \to H^0} \leq Ch^k$.
$(I-PL_2^{-1}\ RL_3): H^0 \to H^0$ being bounded we need for smoothing property

$$\|(I-B_3L_3)^q\|_{H^0 \to H^k} \leq C(q)h^{-k}$$

with C(q) sufficiently small for large enough q, i.e. components in the error with large derivatives should be damped sufficiently. Such estimates can be proved. E.g. HACKBUSH [1979] proves for regular elliptic differential problems of order 2m and (damped) Jacobi relaxation:

$$\|(I-B_3L_3)^q\|_{H^\alpha \to H^{\alpha+2m}} \leq q^{-1}\ h^{-2m}.$$

Analogously, in the third term of M_3^{MLA}, the factor $L_1^{-1} - PL_0^{-1}R$ reduces the lowest frequencies, whereas the factors $(I-B_iL_i)$, = i,1,2,3, reduce each a particular range of higher frequencies. The final effect

is that a bound for $\|M_h^{MLA}\|$ can be found that is less than one *uniformly in h*. This is in contrast with a plain relaxation method for the solution of a discretized differential equation for which $\|M_h^{REL}\| \to 1$ as $h \to 0$.

5.2 A multigrid method for the solution of a Fredholm integral equation of the 2nd kind.

In this example we consider the integral equation

$$x(s) - \int_a^b k(s,t)\, x(t)\, dt = y(s),$$

or, in operator notation,

$$Lx \equiv x - Kx = y,$$

and we consider a sequence of related discretizations

$$L_p\, x \equiv x - K_p\, x = y_p\,, \qquad p = 0,1,2,\ldots\,,$$

with $h_p \to 0$ as $p \to \infty$.
A simple method to solve the discrete equation is by means of successive substitution

$$x_{i+1} = K_p\, x_i + y_p.$$

This is a Jacobi-type iteration: it is a DCPA with approximate inverse $\tilde{G} = I$. It converges if $\|K_p\| < 1$ and, for a compact operator K, it has a smoothing property.

For $p > 0$, also a coarse grid correction is possible by using - in the DCPA - a coarse grid solution operator $L_{p-1}^{-1} = (I-K_{p-1})^{-1}$ for the approximate inverse.

Combination of one relaxation step and one coarse grid correction step yield the TLA with

$$M_p^{TLA} = (I-L_{p-1}^{-1} L_p)K_p$$
$$= (I-K_{p-1})^{-1} (K_p-K_{p-1})K_p.$$

Under suitable conditions (see HEMKER & SCHIPPERS, 1979) it can be shown that - if the repeated trapezoidal rule is used for the discretization of the integral equation - we have

$$\|M_p^{TLA}\| \leq \|(I-K_{p-1})^{-1}\| \, \|(K_p-K_{p-1})K_p\| \leq C.h_p^2, \text{ for } p \to \infty.$$

The TLA still needs the exact solution of the discretized equation on the lower level p-1. Approximating this solution by recursive application of σ MLA iterations on lower levels we have the MLA with

$$\begin{aligned} M_p^{MLA} &= (I-(I-(M_{p-1}^{MLA})^{\sigma})L_{p-1}^{-1}L_p)K_p \\ &= M_p^{TLA} + (M_{p-1}^{MLA})^{\sigma}L_{p-1}^{-1}L_p K_p \\ &= M_p^{TLA} + (M_{p-1}^{MLA})^{\sigma}(K_p-M_p^{TLA}), \qquad p = 1,2,3,... \end{aligned}$$

Hence,

$$\rho_p \equiv \|M_p^{MLA}\| \leq \|M_p^{TLA}\| + \rho_{p-1}^{\sigma}(\|K_p\| + \|M_p^{TLA}\|).$$

From this it can be derived that, for $\sigma = 2$ and with $\rho_0 = \|M_0^{MLA}\| = \|M_0^{TLA}\|$ small enough, we have

$$\rho_p \leq C \, \|M_p^{TLA}\| = O(h_p^2) \text{ as } p \to \infty.$$

This is the typical behaviour of the multigrid iteration of the second kind: the finer the discretization of the analytical problem is, the faster the iterative process to solve the discrete system of equations converges.

REFERENCES

ATKINSON, K.E., *A survey of numerical methods for the solution of Fredholm integral equations of the second kind*, SIAM, 1976.

BRAKHAGE, H., *Ueber die Numerische Behandlung von Integral gleichungen nach der Quadraturformelmethode*, Num. Math., 2(1960) 183-196.

BRANDT, A., *Multi-level adaptive techniques for singular perturbation problems*, In: Numerical Analysis of Singular Perturbation Problems, P.W. Hemker & J.J.H. Miller eds., Academic Press, London, 1979.

FREDERICKSON, P.O., *Fast approximate inversion of large sparse linear systems,* Report 7-75, Lakehead University, 1975.

HACKBUSCH, W., *On the Convergence of Multigrid Iterations,* Report 79-4, Mathematisches Institut, Universität Köln, 1979.

HEMKER, P.W. and SCHIPPERS, H., *Multiple grid methods for the solution of Fredholm integral equations of the second kind,* Report NW75, Mathematisch Centrum, Amsterdam, 1979.

NICOLAIDES, R.A., *On some theoretical and practical aspects of multigrid methods,* Math. Comp. 147 (1979) 933-952.

STETTER, H.J., *The defect correction principle and discretization methods* Num. Math. 29 (1978) 425-443.

YOUNG, D.M., *Iterative solution of large linear systems,* Academic Press, 1971.

BIBLIOGRAPHY

R.E. ALCOUFFE, A. BRANDT, J. DENDY Jr & J.W. PAINTER.
The multi-grid methods for the diffusion equation with strongly discontinuous coefficients.
Los Alamos Sc. Lab. Report; subm. to: SIAM J on Scientific & Statistical Computation.

G.P. ASTRACHANCEV.
An iterative method of solving elliptic net problems.
Žh. Vyčisl. Mat. Fiz. 11(1971) 439-448.

The iterative improvement of eigenvalues.
Žh. Vyčisl. Mat. Fiz. 16(1976) 131-139.

N.S. BAKHVALOV.
On the convergence of a relaxation method with natural constraints on the elliptic operator.
Žh. Vyčisl. Mat. Fiz. 6(1966) 861-885.

R.E. BANK.
A comparison of two multi-level iterative methods for non-symmetric and indefinite elliptic finite element equations.
CNA Report 154, Univ. of Texas at Austin, (Jan. 1980).

------ & T. DUPONT.
An optimal order process for solving elliptic finite element equations.
Dept. of Math, Univ. of Chicago, (1978).

------ & A.H. SHERMAN.
Algorithmic aspects of the multi-level solution of finite element equations.
CNA Report 144, Univ. of Texas at Austin, (Oct. 1978).

------ & ------
A multi-level iterative method for solving finite element equations.
CNA Report 145, Univ. of Texas at Austin, (Oct. 1978).

------ & ------
PLTMG Users' guide
CNA Report 152, Univ. of Texas at Austin, (Sept. 1979).

K. BÖHMER.
Defect corrections via neighbouring problems; I. General theory.
MRC Technical Summary Rept. 1750, Univ. of Wisconsin, Madison, (1977).

D. BRAESS.
The contraction number of a multi-grid method for solving the Poisson equation.
Preprint, Math. Inst. Ruhr-Univ. Bochum, Oct. 1980.

A. BRANDT.
Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems.
In: Procs of the 3rd Int. Conf. on Num. Meth. in Fluid Mechanics (Paris 1972), Lecture Notes in Physics 18, pp.82-89, Springer, Berlin & New York, 1973.

Multi-level adaptive techniques
IBM Res. Report RC6026, Yorktown Heights, New York, 1976.

Multi-level adaptive solutions to boundary value problems.
Math. Comp. 31(1977) 333-390.

Multi-level adaptive solutions to partial differential equations; ideas and software.
In: Procs of Symp. on Math. Software (J. Rice ed.) pp.277-318, Academic Press, New York, 1977.

Lecture notes of the ICASE workshop on multi-grid methods.
(with contributions by: J.C. South, J. Oliger, F. Gustavson, C.E. Grosch, D.J. Jones and T.C. Poling).
ICASE,NASA Langley Research Center, Hampton, Virginia, 1978.

Multi-grid solutions to flow problems; numerical methods for partial differential equations.
MRC Procs of Advanced Seminar, Univ. of Wisconsin, Madison; to appear.

Multi-level adaptive techniques (MLAT) for singular perturbation problems.
In: Numerical Analysis of Singular Perturbation Problems. (P.W. Hemker & J.J.H. Miller eds), Academic Press, London, 1979.

Multi-level adaptive computations in fluid dynamics.
Procs of the AIAA 4th Comp. Fluid Dynamics Conf., Williamsburg, Virginia, July 1979.

Multi-level adaptive finite element methods: I. *Variational Problems.*
ICASE Res. Report 79-8 (1979).

Numerical stability and fast solutions to boundary value problems.
In: Boundary and Interior Layers-Computational and Asymptotic Methods. (J.J.H. Miller ed.). Boole Press, Dublin, 1980.

---, J.E. DENDY Jr & H. RUPPEL.
The multi-grid for the pressure iteration in Eulerian and Lagrangian hydrodynamics.
Los Alamos Sc. Lab. Report LA-UR 77-2995.

(BRANDT, DENDY & RUPPEL)

------ , ------ & ------
The multi-grid method for semi-implicit hydrodynamic codes.
Los Alamos Sc. Lab. Report LA-UR 78-3066 and J. Comp. Phys. 34(1980) 348-370.

------ & N. DINAR.
Multigrid solutions to elliptic flow problems.
In: Numerical Methods for Partial Differential Equations. (S.V. Parter ed.), Academic Press, 1979.

N.I. BULEEV.
Numerical method of solving two-dimensional and three-dimensional diffusion equations.
Matematicheskii Sbornik, Vol. 5, No. 2 (1960).

J.W. DANIEL & A.J. MARTIN.
Numerov's method with deferred corrections for two-point boundary value problems.
SIAM J. Numer. Anal. 14(1977) 1033-1050.

N. DINAR.
On several aspects and applications of the multi-grid method for solving partial differential equations.
NASA Contractor Report 158947, NASA Langley Res. Center, Hampton, Virginia, (Sept. 1978).

Fast methods for the numerical solution of boundary value problems
Ph.D. Thesis, Weizmann Institute of Science, Rehovot, Israel (1979).

R.P. FEDORENKO.
A relaxation method for solving elliptic difference equations.
Žh. Vyčisl. Mat. Fiz. 1(1961) 922-927.

The speed of convergence of an iteration process.
Žh. Vyčisl. Mat. Fiz. 4 (1964) 559-564.

H. FÖRSTER, K. STÜBEN & U. TROTTENBERG.
Non-standard multi-grid techniques using checkered relaxation and intermediate grids.
Preprint IMA-GMD, St. Augustin, 1980; to appear in: Elliptic Problem Solvers (M. Schultz ed.), Academic Press, New York, 1980.

R. FRANK.
The method of iterated defect-correction and its application to two-point boundary value problems.
Num. Math. 25(1976) 409-419.

------ & C.W. UEBERHUBER.
Iterated defect correction for the efficient solution of stiff systems of ordinary differential equations.
BIT 17(1977) 146-159.

P.O. FREDERICKSON.
Fast approximate inversion of large sparse linear systems.
Report 7-75, Lakehead Univ., Ontario, Canada 1975.

L. FUCHS.
A Newton-multi-grid method for the solution of non-linear partial differential equations.
In: Boundary and Interior Layers - Computational and Asymptotic Methods. (J.J.H. Miller ed.), Boole Press, Dublin, 1980.

F.G. GUSTAVSON.
Implementation of the multi-grid method for solving partial differential equations.
IBM Res. Report RA82 (Nr. 26690) pp.51-57, Yorktown Heights, New York, 1976.

W. HACKBUSCH.
Ein iteratives Verfahren zur schnellen Auflösung elliptischer Randwertproblemen.
Report 78-12, Math. Inst. Univ. Köln, 1976.

On the convergence of a multi-grid iteration applied to finite element equations.
Report 77-8, Math. Inst. Univ. Köln, July 1977.

A fast numerical method for elliptic boundary value problems with variable coefficients.
In: Second GAMM-Conference on Num. Meth. in Fluid Mechanics. (E.H. Hirschel & W. Geller eds), DFVLR, Köln, 1977.

A multi-grid method applied to a boundary value problem with variable coefficients in a rectangle.
Report 77-17, Math. Inst. Univ. Köln, 1977.

On the multi-grid method applied to difference equations.
Computing 20 (1978) 291-306.

On the fast solution of nonlinear elliptic equations.
Num. Math. 32 (1979) 291-306.

On the fast solving of parabolic boundary control problems.
SIAM J. Contr. Opt. 17 (1979) 231-244.

A fast iterative method for solving Poisson's equation in a general region.
In: Numerical treatment of differential equations, Oberwolfach, 1976. (R. Bulirsch, R.D. Grigorieff & J. Schröder eds), LNM 631, Springer, Berlin, 1978.

(HACKBUSCH)

On the fast solving of elliptic control problems.
To appear in: J. Optimization Theory and Applications.

Die schnelle Auflösung der Fredholmschen Integralgleichung zweiter Art.
Report 78-4, Math. Inst. Univ. Köln, 1978; to appear in: Beiträge zur Numerischen Mathematik 9.

An error analysis of the nonlinear multi-grid method of the second kind.
Report 78-15, Math, Inst. Univ. Köln, 1978; to appear in: Apl. Mat.

A fast iterative method for solving Helmholtz's equation in a general region.
In: Fast Elliptic Solvers (U. Schumann ed.) Advance Publications, London, 1978.

On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method.
SIAM J. Numer. Anal. 16 (1979) 201-215.

On the convergence of multi-grid iterations.
Report 79-4, Math. Inst. Univ. Köln, 1979; to appear in: Beiträge zur Numerischen Mathematik 9.

Convergence of multi-grid iterations applied to difference equations.
Report 79-5, Math. Inst. Univ. Köln 1979;
Math. Comp. 34(1980) 425-440.

Optimal $H^{p,p/2}$ *error estimates for a parabolic Galerkin method.*
Preprint Math. Inst. Univ. Köln, 1978.

On the regularity of difference schemes.
To appear in: Ark. Mat.

The fast numerical solution of very large elliptic difference schemes.
Preprint Math. Inst. Univ. Köln, 1978.

Bemerkungen zur iterierten Defektkorrektur und zu ihrer Kombination mit Mehrgitterverfahren.
Report 79-13, Math. Inst. Univ. Köln 1979.

The fast numerical solution of time-periodic parabolic problems.
Report 79-14, Math. Inst. Univ. Köln 1979.

(HACKBUSCH)

Survey of convergence proofs for multi-grid iterations.
In: Special Topics of Applied Mathematics. (J. Frehse, D. Pallaschke & U. Trottenberg eds), North Holland Publ. Comp., 1980.

Multi-grid solutions to linear and nonlinear eigenvalue problems for integral and differential equations.
Report 80-3, Math. Inst. Univ. Köln 1980.

A note on the penalty correction method.
Report 80-6, Math. Inst., Univ. Köln, (July 1980).

Regularity of difference schemes: II. *Regularity estimates for linear and nonlinear problems.*
Report 80-13, Math. Inst. Univ. Köln, Aug. 1980.

------ & G. HOFMANN.
Results of the eigenvalue problem for the plate equation.
Report 80-4, Math. Inst., Univ. Köln 1980.

P.W. HEMKER.
On the structure of an adaptive multi-level algorithm.
MC Report NW 65/79, Math. Centre, Amsterdam 1979;
To appear in BIT 1980.

The incomplete LU-decomposition as a relaxation method in multi-grid algorithms.
In: Boundary and Interior Layers - Computational and Asymptotic Methods. (J.J.H. Miller ed.). Boole Press, Dublin, 1980.

Introduction to multi-grid methods.
to appear in: Nw. Arch. Wisk.

------ & H. SCHIPPERS
Multiple grid methods for the solution of Fredholm integral equations of the 2nd kind.
MC Report NW 75/79, Math. Centre, Amsterdam 1979;
to appear in: Math. Comp. 1981.

J.M. HYMAN
Mesh refinement and local inversion of elliptic partial differential equations.
J. Comp. Phys. 23(1977) 124-134.

A. JAMESON
Accelleration of transonic potential flow calculations on arbitrary meshes by the multiple grid method.
AIAA Paper 79-1458, AIAA 4th Comp. Fluid Dynamics Conf., Williamsburg, July 1979.

L. KRONSJÖ
A note on the nested iterations method.
BIT 15(1975) 107-110.

------ & G. DAHLQUIST.
On the design of nested iterations for elliptic difference equations.
BIT 11(1971) 63-71.

H.N. LEE & R.E. MEYERS.
On time dependent multi-grid numerical technique.
To appear in: Computers and Mathematics with Application, 1979.

B. LINDBERG.
Error estimates and iterative improvement for the numerical solution of operator equations.
UIUCDCS-Report-79-820, Univ. of Illinois, Urbana (July 1976).

J. LINDEN, U. TROTTENBERG & K. WITSCH.
Ein Mehrgitterprogramm zur Bestimmung von 2D-Lösungen für die Poisson-Gleichung in Spalt konzentrischer Kugeln, mit Dirichlet'schen oder Neumann'schen Randbedingungen.
Preprint GMD, St. Augustin; Inst. Angew. Math., Univ. Bonn, 1980.

S.F. McCORMICK.
Mesh refinement for integral equations.
Preprint 1979.

A mesh refinement for $Ax = \lambda Bx$.
Preprint 1979.

Multi-grid methods: an alternate viewpoint.
Los Alamos Sc. Lab. Preprint UCRL, Oct. 1979.

An algebriac interpretation of multi-grid methods.
Preprint 1980.

W.L. MIRANKER.
Hierarchical relaxation.
IBM Res. Report RC6884, Yorktown Heights, New York (1977);
Computing 23 (1979) 267-285.

W. MOL.
A multi-grid method applied to some simple problems.
Memorandum Nr. 287, Appl. Math. Dept. Twente Univ. of Technology, 1979.

MUGTAPE
A tape of multi-grid software and programs.
Contributions by A. Brandt, N. Dinar, F. Gustavson & D. Ophir.
Distributed at the ICASE Workshop on Multi-Grid Methods (1978).

R.A. NICOLAIDES.
On multiple grid and related techniques for solving discrete elliptic systems.
J. Comp. Phys. 19(1975) 418-431.

On the ℓ^2 convergence of an algorithm for solving finite element equations.
Math. Comp. 31(1977) 892-906.

On finite element multi-grid algorithms and their use.
ICASE-Res. Report 78-8 (March 1978).

On the observed rate of convergence of an iterative method applied to a model elliptic difference equation.
Math. Comp. 32(1978) 127-133.

On multi-grid convergence in the indefinite case.
Math. Comp. 32(1978) 1082-1087.

On some theoretical and practical aspects of multi-grid methods.
Math. Comp. 33(1979) 933-952.

D. OPHIR.
Language for processes of numerical solutions to differential equations.
Ph.D. Thesis, Weizmann Institute of Science, Rehovot, Israel (1979).

J.W. PAINTER.
Multi-grid experience with the neutron diffusion equation.
Los Alamos Sc. Lab. Report LA-UR 79-1634, (1979).

S.J. POLAK, A. WACHLERS, Th. BEELEN & P.W. HEMKER.
A mesh-parameter-continuation method.
MC Report NW89/80, Math. Centre, Amsterdam, 1980;
to appear in: Elliptic Problem Solvers (M. Schultz ed.), Academic Press, New York, 1980.

T.C. POLING.
Numerical experiments with multi-grid methods.
M.A. Thesis, Dept. Mathematics, The college of William and Mary, Williamsburg, Virginia (1978).

M. RIES & U. TROTTENBERG.
MGR- Ein blitzschneller elliptischer Löser.
Preprint Nr. 277, Univ. Bonn, 1978.

H. REUTERSBERG.
Bibliography: Fast elliptic solvers and related topics.
Report IMA-GMD, Gesellschaft für Mathematik und Datenverarbeitung MBH Bonn, 1980.

H. SCHIPPERS.
Multi-grid techniques for the solution of Fredholm integral equations of the 2nd kind.
In: Colloquium on the numerical treatment of integral equations, MCS 41, Math. Centre, Amsterdam, 1979.

Multiple grid methods for oscillating disc flow.
In: Boundary and Interior Layers - Computational and Asymptotic Methods. (J.J.H. Miller ed.). Boole Press, Dublin, 1980.

Y. SHIFTAN.
Multi-grid methods for solving elliptic difference equations (in Hebrew)
M.Sc. Thesis, Weizmann Institute of Science, Rehovot, Israel (1972).

K. SOLCHENBACH, U. TROTTENBERG & K. WITSCH.
Efficient solution of a special heat conduction problem by use of fast elliptic reduction and multi-grid methods.
Preprint, IMA-GMD, St. Augustin, June 1980.

J.C. SOUTH Jr. & A. BRANDT.
Application of a multi-level grid method to transonic flow calculations.
ICASE Res. Report 76-8, NASA Langley Research Center, Hampton, Virginia, 1976.

H.J. STETTER.
The defect correction principle and discretization methods.
Num. Math. 29(1978) 425-443.

K. STÜBEN.
Das programm MG01 zur Lösung von $au_{xx} + bu_{yy} - c(x,y)u = f(x,y)$ *auf allgemeinen 2D-Gebieten.*
Preprint, IMA-GMD, St. Augustin, Febr. 1980.

E.L. WACHSPRESS.
Iterative solution of elliptic systems and applications to the neutron diffusion equation of reactor physics.
Prentice Hall, Englewood Cliffs, N.J., 1966.

P. WESSELING.
Numerical solution of the stationary Navier-Stokes equations by means of a multiple grid method and Newton iteration.
Report NA-18, Delft Univ. of Technology, 1977.

A convergence proof for a multiple grid method.
Report NA-21, Delft Univ. of Technology (1978).

The rate of convergence of a multiple grid method.
Report NA-30, Delft Univ. of Technology (1979),
In: Numerical Analysis (G.A. Watson ed.)
LNM 773, Springer, Berlin, 1980.

(WESSELING)

------ & P. SONNEVELD.
Numerical experiments with a multiple grid and a preconditioned Lanczos type method.
Report NA-32, Delft Univ. of Technology, (1979).
In: Procs of the IUTAM-Symposium, Paderborn 1979, LNM771, Springer, Berlin, 1980.

H. WOLF.
Multi-grid techniek voor het oplossen van Fredholm integraalvergelijkingen van de tweede soort.
MC Report NN 19/79, Math. Centre, Amsterdam 1979.

MULTISTEP SPLITTING METHODS FOR NONLINEAR INITIAL VALUE PROBLEMS

P.J. van der HOUWEN

1. INTRODUCTION

If an initial-boundary value problem for (nonlinear) hyperbolic or parabolic differential equations is semi-discretized with respect to its space variables, we often obtain a system of ordinary differential equations of the form

$$\frac{d^\nu y}{dt^\nu} = f(t,y), \qquad \nu = 1,2 \tag{1.1}$$

with prescribed values for y (and dy/dt) at $t = t_0$. We integrate this initial value problem with a *highly stable* linear multistep method (e.g. *backward differentiation formulas* if $\nu = 1$). This leads us to the problem to solve an equation of the form

$$y - b_0\tau_n^\nu f(t_{n+1},y) = \sum_{\ell=1}^{k} [a_\ell y_{n+1-\ell} + b_\ell \tau_n^\nu f(t_{n+1-\ell}, y_{n+1-\ell})] \tag{1.2}$$

where y_n denotes the numerical solution at $t = t_n$, $\tau_n = t_{n+1} - t_n$ and $\{a_\ell, b_\ell\}$ are real coefficients. The (approximate) solution of this equation is identified with y_{n+1}.

In this contribution we will consider *splitting methods* to solve (1.2) which are such that only a relatively low number (4 say) of f-evaluations are involved. In particular we will analyse the *stability* of the approximate solution y_{n+1} with respect to perturbations of y_n, $y_{n-1}, \ldots, y_{n+1-k}$. The material presented here is mainly based on [8]. However, the derivations of the splitting methods to solve (1.2) given below will be slightly different, because we have tried to fit them into the frame work of Defect Correction Processes discussed in Chapter 4 of these lecture notes.

2. APPROXIMATION OF THE RIGHT-HAND SIDE FUNCTION

In order to shorten the formulae we write instead of (1.2)

$$y - b_0\tau^\nu f(y) = \Sigma \tag{2.1}$$

where Σ is known. When the original partial differential equation has only one space-dimension this equation can usually be solved by Newton type iteration (we assume that (2.1) always has a solution η). In the case of two or more space variables this offers difficulties. Following the ideas outlined in the contribution of Hemker we replace problem (2.1) by a sequence of easier problems. Let $\tilde{f}_j(y)$, $j = 1,2,\ldots,m$, be functions approximating $f(y)$, then we may define the non-stationary Defect Correction Process of type B (DCPB*) by

$$y^{(j)} - b_0\tau^\nu \tilde{f}_j(y^{(j)}) = \Sigma - b_0\tau^\nu[\tilde{f}_j(y^{(j-1)}) - f(y^{(j-1)})], \qquad j = 1,2,\ldots,m, \tag{2.2}$$

where $y^{(0)}$ is an initial guess for the solution of (2.1). The functions $\tilde{f}_j(y)$ should be such that the equations (2.2) are easily solved for $y^{(j)}$.

EXAMPLE 2.1. The two-dimensional diffusion equation gives rise to the system

$$\frac{dy}{dt} = (A_1+A_2)y$$

where A_1 and A_2 are matrices corresponding to the three-point replacements of the operators $\partial^2/\partial x_1^2$ and $\partial^2/\partial x_2^2$. We may define

$$\tilde{f}_j(y) \begin{cases} = A_1 y + A_2 y^{(j-1)} & \text{for } j \text{ odd} \\ = A_1 y^{(j-1)} + A_2 y & \text{for } j \text{ even} \end{cases}$$

Substitution into (2.2) yields a method of the well known Peaceman-Rachford ADI type [7] ∎

2.1 Method of successive corrections for two-dimensional problems

When the right-hand side function f is such that we can find a *splitting function* F(u,v) with

$$F(y,y) = f(y) \tag{2.3}$$

and with "simply structured" Jacobian matrices $\partial F/\partial u$ and $\partial F/\partial v$, the most obvious choice for $\tilde{f}_j(y)$ is

$$\tilde{f}_j(y) = \begin{cases} F(y,y^{(j-1)}) & \text{for } j \text{ odd} \\ F(y^{(j-1)},y) & \text{for } j \text{ even} \end{cases} . \tag{2.4}$$

We remark that two-dimensional problems usually give rise to right-hand side functions which admit such a splitting function F(u,v).

Substitution of (2.4) into (2.2) yields the *method of successive corrections* (cf. [8, formula (3.3)])

$$y^{(j)} = \begin{cases} \Sigma + b_0\tau^\nu F(y^{(j)},y^{(j-1)}), & j \text{ odd} \\ \Sigma + b_0\tau^\nu F(y^{(j-1)},y^{(j)}), & j \text{ even} \end{cases} , \; j = 1,2,\ldots,m. \tag{2.5}$$

It is easily verified that the local truncation error of the scheme {(1,2),(2,5)} is given by (m even)

$$[b_0\tau^\nu]^m \left[(I-b_0\tau^\nu \frac{\partial F}{\partial u})^{-1} \frac{\partial F}{\partial v} (I-b_0\tau^\nu \frac{\partial F}{\partial v})^{-1} \frac{\partial F}{\partial u}\right]^{m/2} (\eta-y^{(0)}) + O(\tau^{p+1}) = \tag{2.6}$$

$$= O(\tau^{\nu m+q+1}+\tau^{p+1}) \text{ as } \tau \to 0$$

where p and q are the orders of consistency of (1.2) and the predictor formula for $y^{(0)}$. Thus the order of accuracy of the scheme {(1,2),(2.5)} equals

$$\min(m\nu+q,p). \tag{2.7}$$

The $y^{(j)}$ may be assumed to be easily solved from (2.5) by Newton

iteration (by virtue of the simply structured Jacobian matrices $\partial F/\partial u$ and $\partial F/\partial v$). However, since each iteration requires the evaluation of the splitting function F, *the method (2.5) is only feasible from a practical point of view if* $m\mu$, μ *being the number of (internal) Newton iterations, is comparable with the number of right-hand side evaluations required by e.g. the ADI methods* (i.e. 4 in moderately non-linear problems). This leads us to the problem of stability because for small m-values the stability properties of $y^{(m)}$ may differ considerably from those of the fully implicit formula (1.2).
iteration for solving the implicit relations. The corresponding computational scheme then is identical to (2.2) with $\tilde{f}_j(y)$ defined by

$$\tilde{f}_j(y) = \begin{cases} f(y^{(0)}) + \dfrac{\partial F}{\partial u}(y^{(0)},y^{(0)})[y-y^{(0)}], & j \text{ odd} \\ f(y^{(0)}) + \dfrac{\partial F}{\partial v}(y^{(0)},y^{(0)})[y-y^{(0)}], & j \text{ even} \end{cases} \qquad j = 1,2,\ldots,m. \tag{2.8}$$

In [9] experiments are reported where the generating linear multistep method (1.2) is the *four step backward differentiation formula for first order systems*, i.e.

$$\nu=1,\ b_0=\frac{12}{25},\ \Sigma=\frac{1}{25}[48y_n - 36y_{n-1} + 16y_{n-2} - 3y_{n-3}], \tag{2.9}$$

and where $y^{(0)} = y_n$. This method is of *order min(m,4)* and for m=2,4 it was proved [8] to be *unconditionally stable* for problems in which $\partial F/\partial u$ and $\partial F/\partial v$ commute and have negative eigenvalues (therefore, it is a suitable method for *parabolic* equations). For the splitting function F(u,v) we chose the *line hopscotch splitting* [4]

$$F(u,v) = f_{\bullet}(u) + f_{+}(v),$$

where $f_{\bullet}$ corresponds to grid points on "even" rows and f_{+} to points on "odd" rows of the spatial grid.

For three or higher dimensional problems admitting splitting functions F(u,v,w,...) such that F(y,y,y,...) = f(y), the method of successive corrections can be defined in a similar way as (2.5). The stability of these methods, however, is rather poor unless τ is extremely small (cf. [8]), hence we choose alternative functions $\tilde{f}_j(y)$ for higher

dimensional problems (see section 2.2).

2.2 Method of stabilizing corrections for higher dimensional problems

Unlike the method of successive corrections, the method of stabilizing corrections cannot be fitted into the frame work of DCP's by defining such simple approximating functions $\tilde{f}_j(y)$ as given by (2.4) or (2.8). However, if we write (2.2) in the form

$$y^{(j)} = \Sigma + b_0\tau^{\nu} f_j^*(y^{(j)}), \qquad j = 1,2,\ldots,m, \tag{2.2'}$$

where $f^*(y)$ is again sort of approximating right-hand side function which is related to $\tilde{f}_j(y)$ by the relation

$$f_j^*(y) = \tilde{f}_j(y) - \tilde{f}_j(y^{(j-1)}) + f(y^{(j-1)}), \tag{2.9}$$

we may define the *method of stabilizing corrections* by the functions (cf. [8, formula (4.7)])

$$\begin{aligned} f_1^*(y) &= \tfrac{1}{2} F(y,y^{(0)},\ldots,y^{(0)}) + \tfrac{1}{2} f(y^{(0)}) \\ f_2^*(y) &= f_1(y^{(1)}) + \tfrac{1}{2}\left[F(y^{(0)},y,y^{(0)},\ldots,y^{(0)}) - f(y^{(0)})\right] \\ f_j^*(y) &= f_{j-1}^*(y^{(j-1)} + \tfrac{1}{2}\left[F(y^{(0)},\ldots,y^{(0)},y,y^{(0)},\ldots,y^{(0)}) - f(y^{(0)})\right] \\ & \qquad\qquad\qquad\qquad\qquad\qquad j = 3,4,\ldots,m. \end{aligned} \tag{2.10}$$

Here, $F(u,v,w,\ldots)$ is a splitting function with m arguments satisfying $F(y,y,\ldots,y) = f(y)$.

Generally the order of accuracy of the scheme {1.2),(2.10)} equals (an exception is the case m=2 [2])

$$\min(\nu+q,p). \tag{2.11}$$

Unlike the situation with the method of successive corrections the iteration process (2.10) reaches its maximal order after the first iteration and the subsequent iterations serve to stabilize the process

(cf. [2]). For a further discussion of (2.10) we refer to [8].

3. APPROXIMATION OF THE JACOBIAN MATRIX

In section 2 we discarded Newton iteration and directly replaced the problem (2.1) by a sequence of simpler problems which were then solved by a (non-stationary) DCP. In this section we follow the usual approach in which we first replace (by Newton iteration) the problem (2.1) by a sequence of linear problems:

$$
\begin{aligned}
&Ly = \phi^{(j-1)}, \quad j = 1,2,\ldots,m \\
&L = I - b_0\tau^{\nu}J, \; J = \frac{\partial f}{\partial y}(y^{(0)}), \; \phi^{(j)} = \Sigma + b_0\tau^{\nu}[f(\bar{y}) - J\bar{y}],
\end{aligned}
\tag{3.1}
$$

where $\bar{y}$ is the solution of the preceding linear problem with $\bar{y} = y^{(0)}$ for $j = 0$. Here, $y^{(0)}$ is again obtained by some predictor formula. Each of these linear problems is now replaced by a sequence of simpler problems and solved by a DCP. Let $\tilde{J}_i$, $i = 1,2,\ldots,$, denote approximations to the Jacobian J then we may define the non-stationary DCP

$$
\begin{aligned}
&x_0 = \bar{y} \\
&x_i = [I - \tilde{L}_i^{-1}L]x_{i-1} + \tilde{L}_i^{-1}\phi^{(j-1)}, \quad i = 1,2,\ldots . \\
&\tilde{L}_i = I - b_0\tau^{\nu}\tilde{J}_i
\end{aligned}
\tag{3.2}
$$

We assume that $\tilde{J}_i$ is such that the application of $\tilde{L}_i^{-1}$ does not require much computational effort. When this process is terminated after (say) μ iterations we put $\bar{y} = x_{\mu}$ and start with the next linear problem. The iteration processes (3.1) and (3.2) can be combined into one formula:

$$
\begin{aligned}
&y^{(j)} = [I - b_0\tau^{\nu}\tilde{J}_j]^{-1}[b_0\tau^{\nu}(J - \tilde{J}_j)y^{(j-1)} + \phi^{(u(j))}] \\
&\phi^{(u(j))} = \Sigma + b_0\tau^{\nu}[f(y^{(u(j))}) - Jy^{(u(j))}], \quad j = 1,\ldots,M,
\end{aligned}
\tag{3.3}
$$

where the $\tilde{J}_j$ should be rearranged and where u(j) is a piecewise constant, non-decreasing function which assumes integer values such that $0 \le u(j) \le j-1$. We shall call u(j) the *update function* of the right-hand side function.

3.1 Splitting the Jacobian matrix

First of all we remark that (3.3) reduces to the method of successive corrections defined by (2.8) if we put u(j) = j-1, M = m and

$$J = J_1 + J_2,\quad \tilde{J}_j = \begin{cases} J_1 = \dfrac{\partial F}{\partial u}\,(y^{(0)}, y^{(0)}), & j \text{ odd} \\ J_2 = \dfrac{\partial F}{\partial v}\,(y^{(0)}, y^{(0)}), & j \text{ even} \end{cases} . \tag{3.4}$$

However, by choosing the update function u(j) less that j-1, we save evaluations of the function ϕ (and hence evaluations of f(y)) while in most cases the iteration error is not substantially increased. To see this we derive from (3.3) the inequality

$$\|\eta - y^{(j)}\| \le b_0\tau^{\nu}\|(I - b_0\tau^{\nu}\tilde{J}_j)^{-1}\|\{\|J - \tilde{J}_j\|\,\|\eta - y^{(j-1)}\| + C_j\|\eta - y^{(u(j))}\|\} \tag{3.5}$$

where

$$C_j = \sup_{y\in Y_j} \left\|\frac{\partial f}{\partial y}(y) - J\right\|,\quad Y_j = \{y \mid y = \theta\eta + (1-\theta)y^{(u(j))},\ 0\le\theta\le 1\}. \tag{3.6}$$

Thus, if $\partial f/\partial y$ is slowly varying the contribution of the Newton iteration error $\eta - y^{(u(j))}$ to the total error $\eta - y^{(j)}$ may be neglected. This means that the iteration error does not depend strongly on the update function u(j) so that usually only a few f-evaluations are needed per integration step in order to get sufficient accuracy. Note that the order of accuracy is given by (2.7) where m now denotes the number of (external) Newton iterations. The number of internal DCP iterations of type (3.2), however, may be large in order to compensate the large error constant in the iteration error caused by the matrices $J - \tilde{J}_j$ in (3.5) (see also remark 3.3). This leads to large numbers of matrix-vector multiplications but does not increase the number of function evaluations. An additional advantage is an improved stability behaviour because the stability properties of the fully implicit formula (1.2) are better approximated.

REMARK 3.1 Depending on the splitting (3.4), the iteration process {(3.2),(3.4)} presents several of the well known splitting methods for solving linear systems arising from the semi-discretization of partial differential equations.

REMARK 3.2. Instead of (3.4) we may define

$$\tilde{J}_j = J_1 + J_2 - b_0\tau^\nu J_1 J_2 = J - b_0\tau^\nu J_1 J_2.$$

The iteration process (3.3) then assumes the form

$$(3.3') \qquad y^{(j)} = [I-b_0\tau^\nu J_2]^{-1}[I-b_0\tau^\nu J_1]^{-1}[b_0^2\tau^{2\nu} J_1 J_2 y^{(j-1)} + \phi^{u(j)}].$$

It is easily verified that the effect on the iteration error $\eta - y^{(j)}$ by one iteration with (3.3') is roughly the same as two iterations with (3.3).

REMARK 3.3. Since J_1 and J_2 usually have large matrix norms the approximation (3.4') to the Jacobian J is very poor for larger integration steps τ and consequently the iteration process (3.3) will slowly converge. An alternative choice for $\tilde{J}_j$ which is not based on a splitting of J is defined by

$$\tilde{J}_j = \frac{1}{b_0\tau^\nu}\,[I-L^*U^*],$$

where L^*U^* denotes an *incomplete LU-decomposition* of the matrix $I-b_0\tau^\nu J$ (see [5]). Evidently, (3.6) represents an approximation to J and therefore the iteration error (3.5) is expected to be small.

REMARK 3.4. By section 3.1 of Chapter 4 yet another approximation of the Jacobian matrix J is suggested. Let J_H denote the Jacobian matrix J on a coarse grid and let R and P be the restriction and prolongation operators which relate this coarse grid to the grid on which the calculations are performed. Then we may define the approximation (see also (3.13))

$$\tilde{J}_j = PJ_HR.$$

3.2 Coarse grid correction

In Chapter 4 it was shown that an iterative process can be accelerated by defining a second iterative process which is based on a coarser discretization of the iterative operator (cf. p. 18). The two processes were called the *relaxation method* and the *coarse grid correction process*,

respectively. In this section we consider the coarse grid correction in order to accelerate the iteration process (3.2). For the sake of simplicity we assume that the matrix $\tilde{L}_i$ in (3.2) does not depend on i and we denote this matrix by $\tilde{L}_h$ in order to indicate that the original partial differential equation is semi-discretized on a grid Γ_h with grid parameter h. We also write L_h, J_h and $\phi_h^{(j)}$ instead of L, J and $\phi^{(j)}$.

According to section 4.1 of chapter 4 one should alternate μ_j (say) relaxation steps of the type (3.2) with one coarse grid correction step of the form

$$x_{i+1} = x_i + P_{hH} L_H^{-1} R_{Hh}[\phi_h^{(j-1)} - L_h x_i], \tag{3.7}$$

where P_{hH} and R_{Hh} are the prolongator and restrictor defined for the two grids Γ_h and Γ_H (cf. Chapter 4, section 3.1). L_H is defined by

$$L_H = I - b_0\tau^{\nu} J_H, J_H = \frac{\partial \tilde{\tilde{H}}}{\partial y}(R_{Hh} y^{(0)}). \tag{3.8}$$

If the grid Γ_H is sufficiently coarse we may evaluate L_H^{-1} directly, otherwise we have to solve the problem

$$\begin{aligned} L_H z &= \psi_H^{(j)} \\ \psi_H^{(j)} &= R_{Hh}[\phi_h^{(j-1)} - L_h x_{\mu_j}] \end{aligned} \tag{3.9}$$

For instance, the DCP (3.10) may be used:

$$\begin{aligned} z_0 &= \tilde{L}_H^{-1}\psi_H^{(j)} \\ z_\ell &= [I - \tilde{L}_H^{-1} L_H] z_{\ell-1} + \tilde{L}_H^{-1}\psi_H^{(j)}, \quad \ell = 1,2,\dots, \\ \tilde{L}_H &= I - b_0\tau^{\nu}\tilde{J}_H \end{aligned} \tag{3.10}$$

where $\tilde{J}_H$ is some approximation to J_H such that $\tilde{L}_H^{-1}$ is easily evaluated. In practice, $\tilde{J}_H$ may be chosen in an analogous way as J_h, for instance according to remark 3.2 and 3.4.

Let us assume that (3.9) is solved with negligible error and let us insert a coarse grid correction in (3.3) for $j = \mu_1, \mu_1 + \mu_2, \dots = M_1, M_2, \dots$. Thus, for these values of j we replace in (3.3) $y^{(j-1)}$ by

an expression of the form (3.7) to obtain

$$y^{(j)} = [I-b_0\tau^\nu\tilde{J}_h]^{-1}\{b_0\tau^\nu(J_h-\tilde{J}_h)y^{(j-1)} + \phi_h^{(u(j))} + b_0\tau^\nu(J_h-\tilde{J}_h)P_{hH}[I-b_0\tau^\nu J_H]^{-1}R_{Hh}[\phi_h^{(u(j))} + - (I-b_0\tau^\nu J_h)y^{(j-1)}]\}. \tag{3.11}$$

EXAMPLE 3.1. Consider the special two-level algorithm which is obtained if (3.11) is used for all j with u(j) = j-1, i.e.

$$y^{(j)} = [I-b_0\tau^\nu\tilde{J}_h]^{-1}\left\{\sum + b_0\tau^\nu[f(y^{(j-1)}) - \tilde{J}_h y(j-1) + + (J_h-\tilde{J}_h)(I-b_0\tau^\nu J_h^*)^{-1}(\sum+b_0\tau^\nu f(y^{(j-1)})-y^{(j-1)})] + b_0\tau^\nu J^* y^{(j-1)}\right\},$$

where we have introduced the matrix (it is assumed that $R_{Hh}P_{hH} = I$)

$$J_h^* = P_{hH}J_H R_{Hh}. \tag{3.13}$$

For the iteration error we find the relation

$$\eta - y^{(j)} = b_0\tau^\nu[I-b_0\tau^\nu\tilde{J}_h]^{-1}\left\{f_h(\eta) - f_h(y^{(j-1)}) - J_h(\eta-y^{(j-1)}) + b_0\tau^\nu(J_h-\tilde{J}_h)(I-b_0\tau^\nu J_h^*)^{-1}[f(\eta)-f(y^{(j-1)}) - J_h^*(\eta-y^{(j-1)})]\right\}. \tag{3.14}$$

From this relation it follows that

$$\|\eta-y^{(j)}\| \le b_0\tau^\nu\|(I-b_0\tau^\nu\tilde{J}_h)^{-1}\|\left\{C_j+b_0C_j^*\tau^\nu\|J_h-\tilde{J}_h\| \|(I-b_0\tau^\nu J_h^*)^{-1}\|\right\}\|\eta-y^{(j-1)}\|, \tag{3.15}$$

where

$$C_j = \sup_{y\in Y_j}\left\|\frac{\partial f_h}{\partial y}(y) - J_h\right\|, \quad Y_j = \left\{y \mid y = \theta\eta + (1-\theta)y_{j-1},\ 0\le\theta\le 1\right\} \tag{3.16}$$

and a similar definition for C_j^* (J_h replaced by J_h^*). Since C_j and C_j^* are expected to be small for slowly varying Jacobian matrices $\partial f/\partial y$, the

accuracy will be higher than in the case of the method of successive corrections with the same number of iterations. ⊠

REFERENCES

[1] G. DAHLQUIST, *On accuracy and unconditional stability of linear multistep methods for second order differential equations*, BIT, 18 (1978), pp. 133-136.

[2] J. DOUGLAS & J.E. GUNN, *A general formulation of alternating direction methods*, Part I. *Parabolic and hyperbolic problems*, Num. Math. 6 (1964), pp. 428-453.

[3] J. DOUGLAS & H.H. RACHFORD, *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc. 83 (1956), pp. 421-438.

[4] A.R. GOURLAY, *Splitting methods for time-dependent partial differential equations*, in "The state of the art in numerical analysis" ed. D.A.H. Jacobs, Academic Press 1977.

[5] A. GREENBAUM, *Comparison of splittings used with the conjugate gradient algorithm*, Num. Math. 33 (1979), pp. 181-194.

[6] J.D. LAMBERT, *Computational methods in ordinary differential equations*, John Wiley & Sons, London, 1971.

[7] D.W. PEACEMAN & H.H. RACHFORD, *The numerical solution of parabolic and elliptic equations*, J. Soc. Ind. Appl. Math., 3 (1955), pp. 28-41.

[8] P.J. VAN DER HOUWEN, *Multistep splitting methods of high order for initial value problems*, SIAM J. Numer. Anal., 3 (1980), pp. 420-427.

[9] -, B.P. SOMMEIJER & J.G. VERWER, *Comparing time integrators for parabolic equations in two space dimensions with a mixed derivative*, J. Comp. Appl. Math., 5 (1979), pp. 73-83.

[10] N.N. YANENKO, *The method of fractional steps*, Springer Verlag, Berlin, 1971.

ON THE TREATMENT OF TIME-DEPENDENT BOUNDARY CONDITIONS IN SPLITTING METHODS FOR PARABOLIC DIFFERENTIAL EQUATIONS

P.J. van der HOUWEN*

REMARK. Only a summary of the colloquium lecture is given. The complete paper will appear in the "International Journal for Numerical Methods in Engineering".

In [4] Fairweather and Mitchell investigated Alternating Direction Implicit (ADI) methods for the heat condition equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x_1^2} + \frac{\partial^2 U}{\partial x_2^2} \tag{1.1}$$

in a domain Ω with Dirichlet boundary conditions along the boundary $\partial\Omega$. Among other things, they discussed the classical Peaceman-Rachford ADI-method on a square Ω with square meshes of size h, i.e. the scheme

$$\begin{aligned} (I - \tfrac{1}{2}\tfrac{\tau}{h^2}\delta_{x_1^2})u_{\bar{n}} &= (I + \tfrac{1}{2}\tfrac{\tau}{h^2}\delta_{x_2^2})u_n \\ (I - \tfrac{1}{2}\tfrac{\tau}{h^2}\delta_{x_2^2})u_{n+1} &= (I + \tfrac{1}{2}\tfrac{\tau}{h^2}\delta_{x_1^2})u_{\bar{n}} \end{aligned}, \tag{1.2}$$

where $u_{\bar{n}}$, u_n and u_{n+1} denote grid functions defined on the grid $\Gamma_h \cup \partial\Gamma_h$ covering $\Omega \cup \partial\Omega$ and where $\delta_{x_i^2}/h^2$ denotes the usual finite difference replacement of $\partial^2/\partial x_i^2$; furthermore, τ is the integration step and u_n, u_{n+1} present numerical approximations to the exact solution values U at times t_n and t_{n+1}, respectively. By defining u_n and u_{n+1} on the set of boundary grid points $\partial\Gamma_h$ by Dirichlet boundary conditions, the scheme (1.2) can be applied in all internal grid points provided $u_{\bar{n}}$ is prescribed along those parts of the boundary for which the x_1-coordinate is constant. Peaceman and Rachford defined in their paper [6]

$$u_{\bar{n}} = U(t_n + \tfrac{1}{2}\tau, x_1, x_2), \qquad (x_1, x_2) \in \partial\Gamma_h. \tag{1.3}$$

---

* Joint work with B.P. Sommeijer and J.G. Verwer.

the region $\Omega \cup \partial\Omega$ is replaced by a grid $\Gamma_h \cup \partial\Gamma_h$ characterized by the parameter h and which is defined for each $h \in (0,\bar{h}]$ such that $\Gamma_h \cup \partial\Gamma_h$ is dense in $\Omega \cup \partial\Omega$ as $h \to 0$; (ii) the right hand side of the partial differential equation and the boundary condition in (1.5) is discretized on $\Gamma_h \cup \partial\Gamma_h$ in such a way that the equation and the boundary condition together convert into a system of ordinary differential equations

$$\frac{dy}{dt} = f(t,y+b), \quad b(t) = g(t,y(t)). \tag{1.6}$$

Here, to each grid point $\in \Gamma_h \cup \partial\Gamma_h$ there corresponds a component of y, f and b, those of y and f being zero at all boundary grid points $\in \partial\Gamma_h$ and those of b being zero at all internal grid points $\in \Gamma_h$. Thus, y, f and b have as many components as there are grid points in $\Gamma_h \cup \partial\Gamma_h$. Furthermore, system (1.6) has as many non-trivial equations as there are internal grid points. The function g(t,y) expresses the boundary values in terms of t and y. We shall assume that f is defined for each $h \in (0,\bar{h}]$ and that the exact solution y(t) of (1.6) and the grid function $U_h(t)$ obtained by restricting the exact solution $U(t,x_1,x_2)$ of the initial-boundary value problem to the grid $\Gamma_h \cup \partial\Gamma_h$, satisfy the condition

$$U_h(t) - [y(t) + b(t)] = \varepsilon(t,h) \to 0 \qquad \text{as } h \to 0 \tag{1.7}$$

(provided of course that $U_h(t_0) = y(t_0) + b(t_0)$). It should be noted that our assumption on the existence of f for $0 < h \le \bar{h}$ does not mean that f remains bounded as $h \to 0$. Only for sufficiently smooth grid functions (e.g. $U_h(t)$) the right hand side functions will converge as $h \to 0$. This observation turns out to be crucial in deriving the Fairweather-Mitchell correction for the problem (1.5).

By virtue of assumption (1.7) our considerations can be restricted to the time integration of the initial-boundary value problem, that is the integration of the initial value problem for equation (1.6). In section 2 we define a class of one-step splitting formulas for (1.6) and we derive the *error of approximation* of these formulas. This error is the residual left when the exact solution y(t) of (1.6) is substituted into the numerical scheme. Thus, by writing the numerical scheme in the form

D'Yakonov [3] (see also [10, Section 2.9]) and Fairweather and Mitchell [4] showed, however, that the method will lose accuracy if the boundary conditions become time-dependent. In order to improve the accuracy, Fairweather and Mitchell proposed to replace $u_{\bar{n}}$ along the vertical parts of the boundary by

$$u_{\bar{n}}^{*} = \frac{U(t_n,x_1,x_2)+U(t_{n+1},x_1,x_2)}{2} + \frac{\tau}{4h^2}\,\delta_{x_2}^2[U(t_n,x_1,x_2)-U(t_{n+1},x_1,x_2)]. \tag{1.4}$$

The effect of the modification (1.4) is that at points adjacent to a vertical boundary (1.2) becomes an $O(h^2+\tau^2)$ approximation to the equation (1.1), whereas (1.3) yields an $O(h^2+\tau^2/h^2)$ approximation.

The purpose of this paper is to derive the Fairweather-Mitchell modification for a family of splitting methods (including the classical ADI- and LOD-schemes) and for a rather general class of initial-boundary value problems given by

$$\frac{\partial U}{\partial t} = G_1(t,x_1,x_2,U,\frac{\partial U}{\partial x_1},\frac{\partial^2 U}{\partial x_1^2}) + G_2(t,x_1,x_2,U,\frac{\partial U}{\partial x_2},\frac{\partial^2 U}{\partial x_2^2}),$$

$$(x_1,x_2) \in \Omega \cup \partial\Omega, \tag{1.5a}$$

$$U(t_0,x_1,x_2) = U_0(x_1,x_2), \qquad (x_1,x_2) \in \Omega \cup \partial\Omega$$

$$a_0(t,x_1,x_2)U + a_1(t,x_1,x_2)\frac{\partial U}{\partial x_1} + a_2(t,x_1,x_2)\frac{\partial U}{\partial x_2} = a_3(t,x_1,x_2), \tag{1.5b}$$

$$(x_1,x_2) \in \partial\Omega.$$

Throughout the paper it is assumed that Ω is a bounded and path-connected region in the (x_1,x_2)-space. Further, it is assumed that the functions G_1, G_2, and a_i, $i = 0,1,2,3$, as well as the solution U, are sufficiently smooth.

Since the Fairweather-Mitchell modification has to do with the *time-discretization* of (1.5), and is not part of the space-discretization, we follow in our analysis the *method of lines* which more or less separates the discretization of $\partial U/\partial t$ from the discretization of the right hand side of the partial differential equation. In the method of lines we assume that (i)

$$(1.8) \qquad \frac{y_{n+1}-y_n}{\tau} = S_n(y_n, y_{n+1}),$$

the error of approximation A_n over the interval $[t_n, t_n+\tau]$ is defined by

$$(1.9) \qquad A_n = \frac{y(t_{n+1})-y(t_n)}{\tau} - S_n(y(t_n), y(t_{n+1})).$$

Here, S_n denotes an operator defined by the splitting formula and the functions f and g. We observe that A_n is closely related to the local error of (1.8), which is usually considered in the numerical analysis of ordinary differential equations. To see this we consider the local error

$$(1.10) \qquad \rho_n = y(t_{n+1}) - y_{n+1} = y(t_{n+1}) - y(t_n) - \tau S_n(y(t_n), y_{n+1})$$

where it is assumed that $y_n = y(t_n)$. Let $S_n(u,v)$ be differentiable with respect to its second argument, then it follows from (1.9), (1.10) and a mean-value argument (cf. [11, p. 68]) that

$$\rho_n = \tau A_n + \tau[S_n(y(t_n), y(t_{n+1})) - S_n(y(t_n), y_{n+1})]$$

$$= \tau A_n + \tau B_n(y(t_{n+1}), y_{n+1})[y(t_{n+1}) - y_{n+1}],$$

where $B_n(y(t_{n+1}), y_{n+1})$ is a matrix with elements $\partial S_n^{(P)}/\partial v^{(Q)}$ evaluated at $(y(t_n), \bar{v})$, $\bar{v}$ being an intermediate point "between $y(t_{n+1})$ and y_{n+1}" and depending on row and column index P and Q. Thus, A_n and ρ_n are related by the equation

$$(1.11) \qquad [I - \tau B_n(y(t_{n+1}), y_{n+1})]\rho_n = \tau A_n.$$

REFERENCES

[1] DOUGLAS, J. Jnr., *On the numerical integration of* $u_{xx} + u_{yy} = u_t$ *by implicit methods,* J. Soc. Ind. Appl. Math. 3, 42-65, 1955.

[2] DOUGLAS, J. Jnr. & H.H. RACHFORD Jnr., *On the numerical solution of heat conduction problems in two and three space variables,* Trans. Amer. Math. Soc. 82, 421-439, 1956.

[3] D'YAKONOV, YE.G., *Some difference schemes for solving boundary problems,* U.S.S.R. Comput. Math. and Math. Phys., No. 1, 55-77, 1963.

[4] FAIRWEATHER, G. & A.R. MITCHELL, *A new computational procedure for ADI-methods,* SIAM J. Numer. Anal. 4, 163-170, 1967.

[5] HUBBARD, B.E., *Alternating direction schemes for the heat equation in a general domain,* J. SIAM Numer. Anal. (series B) 2, 448-463, 1965.

[6] PEACEMAN, D.W. & H.H. RACHFORD Jnr., *The numerical solution of parabolic and elliptic differential equations,* J. Soc. Ind. Appl. Math. 3, 28-41, 1955.

[7] SAMARSKII, A.A., *On an economical difference method for the solution of a multidimensional parabolic equation in an arbitrary region,* U.S.S.R. Comput. Math. and Math. Phys., No. 5, 894-926, 1963.

[8] VAN DER HOUWEN, P.J. & J.G. VERWER, *One-step splitting methods for semi-discrete parabolic equations,* Computing 22, 291-309, 1979.

[9] VARGA, R.S., *Matrix iterative analysis,* Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[10] YANENKO, N.N., *The method of fractional steps,* Springer-Verlag, Berlin, 1971.

[11] ORTEGA, J.M. & W.C. RHEINBOLDT, *Iterative solution of nonlinear equations in several variables,* Academic Press, New York, 1970.

ON THE STABILITY OF A CLASS OF DIFFERENCE METHODS FOR BOUNDARY VALUE PROBLEMS OF THE HEAT EQUATION WITH TIME DEPENDENT COEFFICIENTS

R.M.M. MATTEIJ

1. INTRODUCTION

An investigation is made to the L_2 stability of "θ methods" for solving the heat equation by finite differences. The boundary conditions are supposed to contain derivatives. For Neumann conditions only a weak result is given (if the problem is non separable). If the boundary conditions are somewhat relaxed it can be shown that we have stability if the time discretization is of the order of the space discretization.

We consider the PDE

$$\frac{\partial u}{\partial t} = b(t,x)\,\frac{\partial^2 u}{\partial x^2} + c(t,x)\,\frac{\partial u}{\partial x}, \qquad b > 0 \tag{1.1}$$

where $u(t,x)$ is defined on a strip $S = \{[t,x] \mid t \geq 0,\ 0 \leq x \leq 1\}$, subject to the boundary conditions

$$\begin{aligned} &\text{(a)}\ u(0,x) = f(x),\ 0 < x < 1 \\ &\text{(b)}\ \frac{\partial}{\partial x}\,u(t,0) = h_0(t)\,u(t,0) \\ &\text{(c)}\ \frac{\partial}{\partial x}\,u(t,1) = -h_1(t)\,u(t,1), \end{aligned} \tag{1.2}$$

where $h_0, h_1 \geq 0$. Later on we shall impose more conditions upon the functions a, b, f, h_0 and h_1, where necessary.

Our purpose is to consider some ways to prove the stability of a difference scheme for (1.1) and (1.2). Of course the subject is not new; in fact it has already received much attention. Just to mention some papers, see [1,2,3,5,6]. However, these papers are either restricted to constant coefficients, or coefficients that only depend on the space variable x, or to Dirichlet boundary conditions. By the present analysis we show that also for more general coefficients stability of such schemes can be proved.

Rather than considering general space and time discretizations, we restrict ourselves to the well-known three point spatial discretization for $\frac{\partial^2}{\partial x^2}$, the central difference for $\frac{\partial}{\partial x}$ and the weighted average of Euler forward and Euler backward ("θ method") for $\frac{\partial}{\partial t}$. We remark that this difference scheme is mainly taken for demonstrative reasons, not because of their particular usefulness. If e.g. b(t,x) may be unduly small (as in boundary layer problems) other methods should be employed (cf. e.g. [4]).

In order to have a feeling what might be expected, it makes sense to look at the stability of the trapezoidal rule, applied to the ODE

$$\frac{du}{dt} = q(t)u, \quad q < 0. \tag{1.3}$$

So discretize (1.3) by

$$v_{i+1} = v_i + \frac{\Delta t}{2}\,(q_i v_i + q_{i+1} v_{i+1}) \tag{1.4}$$

(Δt time discretization, v an approximation for $u(i\Delta t)$)

Then

$$v_{i+1} = \frac{1 + \frac{\Delta t}{2}\, q_i}{1 - \frac{\Delta t}{2}\, q_{i+1}}\, v_i. \tag{1.5}$$

Hence we have *A stability* if

$$\left|\frac{2 + \Delta t q_i}{2 - \Delta t q_{i+1}}\right| \leq 1. \tag{1.6}$$

This holds true if $q_{i+1} - q_i < \frac{4}{\Delta t}$, so

PROPERTY 1.6. *The trapezoidal rule is A stable if either* $q' < 0$ *or* $q' > 0$ *but* $q' < \frac{4}{(\Delta t)^2}$.

It may seem that this last result is a bit pessimistic, regarding the following estimation: Let e_j be an error at time t_j then we have for the propagated error e_i at time t_i:

$$e_i = \frac{2 + \Delta t q_{i-1}}{2 - \Delta t q_i}\ \frac{2 + \Delta t q_{i-2}}{2 - \Delta t q_{i-1}} \cdot \cdots \cdot \frac{2 + \Delta t q_j}{2 - \Delta t q_{j+1}}\ e_j. \tag{1.7}$$

Obviously all factors

$$\left|\frac{2 + \Delta t q_\ell}{2 - \Delta t q_\ell}\right|, \quad \ell = i-1,\ldots,j+1$$

are bounded by 1 so we end up with

$$|e_i| \leq \left|\frac{2 + \Delta t q_j}{2 - \Delta t q_i}\right| |e_j|. \tag{1.8}$$

(1.8) again puts restrictions on the growth of the $\{q_i\}$. For N dimensional ODE the problem is how to match the differently growing basis solution, like we can simply do in the one dimensional case. In §§ 4,5 we shall encounter counterparts of both 1.6 and (1.8)

2. THE DIFFERENCE SCHEME

Let [0.1] be divided into (N-1) subintervals of length Δx. Let Δt denote the time step. By v_{ij} we denote the approximating value for $u(i\Delta t,(j-1)\Delta x)$, $1 \leq j \leq N$. Moreover we write h_{i1} for $h_0(t)$ and h_{iN} for $h_1(t)$. A difference scheme is then given by

$$\begin{aligned} v_{i+1,j} = v_{ij} + \frac{\Delta t}{\Delta x^2}\Big\{ & \theta b_{i+1,j}\Big[v_{i+1,j-1} - 2v_{i+1,j} + v_{i+1,j+1} + \\ & + \frac{c_{i+1,j}}{b_{i+1,j}} \frac{\Delta x}{2}\left(v_{i+1,j+1} - v_{i+1,j-1}\right)\Big] + \\ & + (1-\theta) b_{i,j}\Big[v_{i,j-1} - 2v_{ij} + v_{i,j+1} + \frac{c_{ij}}{b_{ij}} \frac{\Delta x}{2}\left(v_{i,j+1} - v_{i,j-1}\right)\Big]\Big\}. \end{aligned} \tag{2.1}$$

If we discretize (1.2) (b) and (1.2) (c) by central differences, thus introducing virtual points at $x = -\Delta x$ and $x = 1+\Delta x$, we can eliminate them in the usual way using (2.1). So for the left boundary we obtain:

$$\begin{aligned} v_{i+1,1} = v_{i1} + \frac{\Delta t}{\Delta x^2}\Big\{ & \theta b_{i+1,1}\Big[2v_{i+1,2} - \\ & + (2+2h_{i+1,1}(1 + \frac{c_{i+1,1}}{b_{i+1,1}} \frac{\Delta x}{2})\Delta x) v_{i+1,1}\Big] \\ & + (1-\theta) b_{i1}\Big[2v_{i2} - (2+2h_{i1}(1 + \frac{c_{i1}}{b_{i1}} \frac{\Delta x}{2})\Delta x) v_{i1}\Big]\Big\}. \end{aligned} \tag{2.2}$$

A similar result is obtained for the boundary condition on the right. Writing for the vector of approximating values v_{ij} at time $i\Delta t$

$$V_i = (v_{i1},\ldots,v_{iN})^T, \tag{2.3}$$

we obtain the following recursion

$$(I + \theta H_{i+1})\, V_{i+1} = (I-(1-\theta)H_i)V_i, \tag{2.4}$$

where

$$H_i = \frac{\Delta t}{\Delta x^2}\,\mathrm{diag}(b_{ij}) \begin{bmatrix} 2+p_{i1} & -2 & & & \emptyset \\ -(1-q_{i2}) & 2 & -(1+q_{i2}) & & \\ & \ddots & \ddots & \ddots & \\ & & & 2 & -(1+q_{i,N-1}) \\ \emptyset & & & -2 & 2+p_{iN} \end{bmatrix} \tag{2.5}$$

with $p_{ij} = 2h_{ij}\,\Delta x(1+\frac{c_{ij}}{b_{ij}}\frac{\Delta x}{2})$, $q_{ij} = \frac{c_{ij}}{b_{ij}}\frac{\Delta x}{2}$ for short

Note that it follows from the assumptions in §1 that $p_{ij} \geq 0$ if Δx is sufficiently small, more precisely if $|\frac{c_{ij}}{b_{ij}}\frac{\Delta x}{2}| \leq 1$; in this case we also may conclude that the off diagonal elements are non positive and we have

PROPERTY 2.6. *Let* $\forall_{i,j}\ |q_{ij}| < 1$. *Then the eigenvalues of* H_i *are contained in the region*

$$[0, \frac{4\Delta t}{\Delta x^2}\max_{2\leq j\leq N-1} b_{ij}] \cup [0, \frac{4\Delta t}{\Delta x^2}\max_{j=1,N} b_{ij}(1+\tfrac{1}{2}p_{ij})].$$

PROOF. As we shall see in §3, H_i is similar to a symmetric matrix; hence it has only real eigenvalues. Gershgorin's theorem gives the estimates. □

COROLLARY 2.7. *If* $\forall_{ij}|q_{ij}| < 1$ *then* $(I + \theta H_i)$ *is non-singular.*

From now on we assume that Δx is chosen so small that $\forall_{i,j}|q_{ij}| \leq \rho < 1$, so that both $1+q_{ij}$ and $1-q_{ij}$ are uniformly bounded away from zero.

3. STABILITY

We provide $\mathbb{R}^n$ with the Euclidean norm $\|\cdot\|$, i.e. $\|v_i\| = \mathrm{sqrt}(\sum_{j=1}^{N} v_{ij}^2)$. We will investigate stability of the scheme (2.4) with respect to this norm, both for finite (case I) and for infinite (case II) time intervals. Introducing the operator

$$B_i = (I + \theta_i H_{i+1})^{-1} (I - (1-\theta_i) H_i) \tag{3.1}$$

(N.B. θ_i may depend on i; (3.1) is a generalization of (2.4))
we have the following well-known criterion (see 3.4)

DEFINITION 3.2. (a) Case I: Let $T > 0$ be given. The scheme (2.4) is called *stable* if $\|\Pi_{\ell=j}^{i} B_\ell\|$ is uniformly bounded for all j,i with $j\,\Delta t$, $i\,\Delta t \le T$ $(i \ge j)$.
(b) Case II: The scheme (2.4) is called *asymptotically stable* if $\|\Pi_{\ell=j}^{i} B_\ell\|$ is uniformly bounded for all i,j with $i \ge j$.

PROPERTY 3.3. *A sufficient condition in order that* (2.4) *is stable, is in case I that* $\|B_\ell\| \le 1 + 0(\Delta t)$ *and in case II that* $\|B_\ell\| \le 1$.

REMARK 3.4. These kinds of stability are sufficient to guarantee that a single rounding error will not "blow up". Moreover, if the solution of the exact problem is sufficiently regular they also imply convergence of the scheme (cf. a remark in [2, p. 111]) as a straightforward analysis shows.

REMARK 3.5. If both h_0 and h_1 are zero (Neumann-conditions) then the matrices H_i are singular and B_i obviously has the eigenvalue 1, corresponding to the eigenvector $(1,1,\ldots,1)^T$. Therefore, if we want to consider the Neumann case, we cannot hope for a general result where $\|B_i\| < 1$ (a situation which would make stability easier to establish).

Our considerations will be based on estimates for symmetric matrices. We can find a symmetric matrix, which is similar to H_i using the diagonal matrix

$$D_i = \mathrm{diag}\,(d_{i1},\ldots,d_{iN}) \tag{3.6}$$

where

$$\text{(a)}\quad d_{i1} = \frac{1}{\sqrt{2\,b_{i1}}}\,; \quad d_{i2} = \frac{1}{\sqrt{b_{i2}(1-q_{i2})}}$$

$$\text{(3.7)}\qquad \text{(b)}\quad d_{i,j+1} = d_{i,j}\sqrt{\frac{b_{ij}}{b_{i,j+1}}\,\frac{1+q_{ij}}{1-q_{i,j+1}}}\,, \quad j = 2,\ldots,N-2$$

$$\text{(c)}\quad d_{i\,N} = d_{i,N-1}\sqrt{\frac{b_{i,N-1}}{b_{iN}}\,\frac{1+q_{i,N-1}}{2}}$$

Indeed, the matrix K_i, defined by

$$\text{(3.8)}\qquad K_i = D_i\,H_i\,D_i^{-1},$$

is symmetric.

In some fairly special cases we can then give immediate stability results, viz if we have separability:

DEFINITION 3.9. The PDE (1.1) is called *separable*, if there exist functions $\beta(t)$, $\gamma(x)$ and $\delta(x)$, such that $b(t,x) = \beta(t)\gamma(x)$ and $c(t,x) = \beta(t)\delta(x)$.

For these cases we have the following generalization of [2]:

THEOREM 3.10. *Let* (1.1) *be separable and let* h_0, h_1 *be constant, then the scheme* (2.4) *is asymptotically stable under the following conditions* (a) *or* (b):

(a) *If* $\forall_i \dfrac{\beta(t_i)}{\beta(t_i)+\beta(t_{i+1})} \le \theta_i \le 1$ *for all* $\Delta t,\Delta x$ (*if* Δx *is small enough, see* §2).

(b) *If* $\forall_i 0 \le \theta_i < \dfrac{\beta(t_i)}{\beta(t_i)+\beta(t_{i+1})}$ *for* $\Delta t,\Delta x$ *satisfying*

$$\frac{\Delta t}{\Delta x^2} \le \tfrac{1}{2}\,\frac{\beta(t_0)}{(1-\theta_i)\beta(t_i)-\theta_i\beta(t_{(i+1)})}\,\max\left\{\min_{2\le j\le N-p}\frac{1}{b_{0j}}\,,\ \min_{j=1,N}\frac{1}{b_{0j}(1+\frac{1}{4}p_{0j})}\right\}.$$

PROOF. By assumption there exists a sequence α_i such that $H_i = \alpha_i H_0$ ($\alpha_i = \beta(t_i)/\beta(t_0)$). Hence all H_i and so B_i are simultaneously diagonalizable. Now λ is an eigenvalue of B_i iff μ is an eigenvalue of H_0 such that

$$\lambda = \frac{1-(1-\theta_i)\alpha_i\mu}{1+\theta_i\alpha_{i+1}\mu}\,.$$

In Figure 3.1 we have drawn a picture of

$$g_i(\mu) = \frac{1-(1-\theta_i)\alpha_i\mu}{1+\theta_i\alpha_{i+1}\mu} .$$

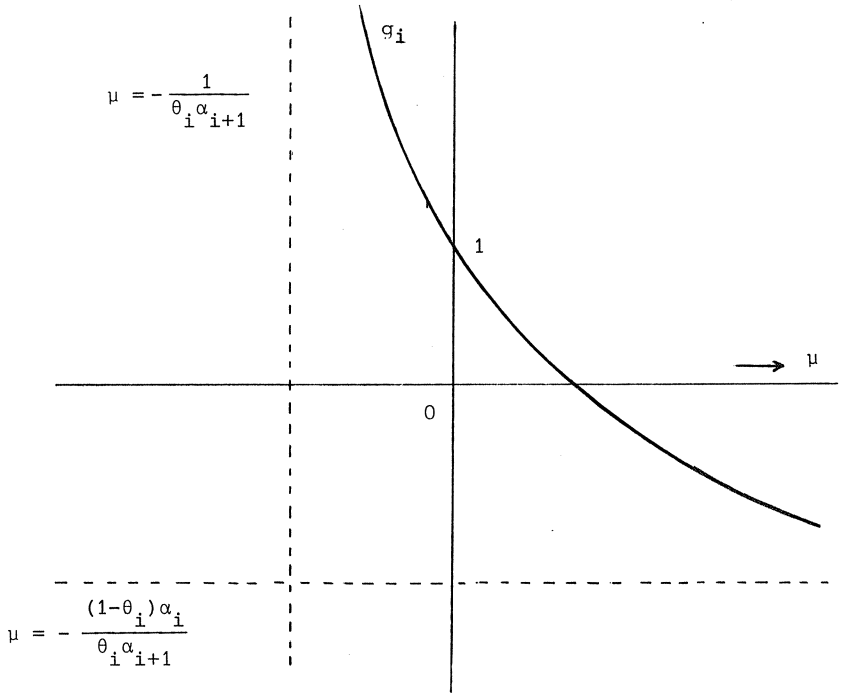


Figure 3.1

Clearly the interesting region is $-1 \le g_i(\mu) \le 1$ (cf. 3.3). In any case $g_i(0) = 1$, and since $\mu \ge 0$ (see 2.6) we only have to worry about $g_i(\mu) \ge -1$. In case (a) this is trivial again. In case (b) we find from intersecting $g_i(\mu)$ with -1 the value $\mu = 2/((1-\theta)\alpha_i - \theta\alpha_{i+1})$. Combination with 2.6 gives the bounds in (b). Since $\Pi_j^i \beta_\ell$ and the symmetrized version (cf.(3.8)) only differ a similarity transformation by a constant matrix D_0 (= D_i for all i), we find $\|\Pi_j^i B_\ell\| \le \|D_0\|\|D_0^{-1}\|$, and hence uniformly bounded. □

COROLLARY 3.11. *Scheme* (2.4) *is in particular asymptotically stable if* (1.1) *is separable and we have Neumann boundary conditions and moreover* 3.10 (a) *or* (b) *is fulfilled.*

4. STABILITY FOR VARIABLE COEFFICIENTS; NEUMANN CONDITIONS

If the matrices B_i, see (3.1), are not separable, theorem 3.10 is not applicable. We can, however, still prove stability results. A first approach is given in this section. First we consider finite time intervals, (case I). We need two lemmata.

LEMMA 4.1. *Let* $\frac{\partial b}{\partial t}$ *and* $\frac{\partial c}{\partial t}$ *be continuous (in both variables) on some finite time interval* $[0,T]$. *Let* $\ell\,\Delta t \le T$. *Then*

$$\| D_\ell \, D_{\ell-1}^{-1} \| \le 1 + O(\Delta t).$$

PROOF. In order to find a bound for $\| D_\ell \, D_{\ell-1}^{-1} \|$ we seek a bound for, cf. (3.7),

$$\max_j \left\{ \frac{b_{\ell-1,j}}{b_{\ell j}} \cdot \prod_{s=2}^{j} \left[\frac{1+q_{\ell,s}}{1-q_{\ell,s+1}} \right] \Big/ \left[\frac{1+q_{\ell-1,s}}{1-q_{\ell-1,s+1}} \right] \right\}^{\frac{1}{2}} \qquad \text{(a)}$$

From the assumptions above, the general assumptions on b (viz b bounded away from zero) and Δx (sufficiently small) it now follows that

$$1 + q_{\ell s} = 1 + q_{\ell-1,s} + O(\Delta t\, \Delta x)$$

and

$$1 - q_{\ell-1,s+1} = 1 - q_{\ell,s+1} + O(\Delta t\, \Delta x) \qquad \text{(both uniformly)}$$

and clearly $b_{\ell-1,j} = b_{\ell,j} + O(\Delta t)$; the expression (a) can thus be estimated uniformly by

$$\left\{ (1 + O(\Delta t))\,(1 + O(\Delta t \cdot \Delta x))^N \right\}^{\frac{1}{2}} = 1 + O(\Delta t) \qquad \text{(b)}$$

In a similar way it is seen that $d_{\ell 1}/d_{\ell-1,1} = 1 + O(\Delta t)$, so that $\max_{j \le N-1} d_{\ell j}/d_{\ell-1,j} = 1 + O(\Delta t)$. Finally, again using the estimates above, it is clear that this argument allows for extending this last maximum over the range $1,\ldots,N$. □

LEMMA 4.2. *The matrices* $(I + \theta H_\ell)^{-1}$, D_ℓ *and* D_ℓ^{-1} *are uniformly bounded.*

PROOF. Since both $1-q_{\ell,2}$ and $1+q_{\ell,N-1}$ are bounded and bounded away from zero, the boundedness of D_ℓ^{-1} and D_ℓ follows from a similar argument as was used in 4.1, now by considering

$$\max_j \frac{1}{b_{\ell j}} \prod_2^s \frac{1+q_{\ell,s}}{1-q_{\ell,s-1}}$$

and also the inverse of the minimum of this expression. For both we obtain a bound like $\kappa(1 + O(\Delta x))^{N/2} \le \tilde{\kappa}$, for some uniform constants κ and $\tilde{\kappa}$. Since H_ℓ has nonnegative eigenvalues (see 2.6), it follows that K_ℓ has such eigenvalues, and therefore $\rho((I+\theta K_\ell)^{-1}) \le 1$. Hence $(I+\theta H_\ell)^{-1}$ is uniformly bounded. □

We now rewrite $\Pi_j^i B_\ell$ as

$$\prod_{\ell=j}^{i} B_\ell = (I + \theta H_{i+1})^{-1} \prod_{\ell=j+1}^{i} A_\ell \, (I - (1-\theta)H_j), \tag{4.3}$$

where

$$A_\ell = (I - (1-\theta)H_\ell)(I + \theta H_\ell)^{-1}. \tag{4.4}$$

We then have

THEOREM 4.5. *Let the assumptions of* 4.1. *hold. Then we have stability on a finite interval (case* I*) for* $\Delta t = O(\Delta x)$ *if* θ_i *is such that* $\frac{1}{2} \le \theta \le 1$ *and* $(1-\theta_i) = O(\Delta x)$.

PROOF.

$$\prod_{\ell=j}^{i} B_\ell = (1 + \theta_i H_{i+1})^{-1} D_i^{-1} \cdot \prod_{\ell=j+1}^{i} \{[D_\ell A_\ell D_\ell^{-1}][D_\ell D_{\ell-1}^{-1}]\} \cdot$$

$$\cdot D_{j+1}(I - (1-\theta_j)H_j).$$

The boundedness of ΠB_ℓ follows from 4.1 and 4.2. The boundedness of $I - (1-\theta_j)H_j$ follows from the assumptions. □

COROLLARY 4.6. *Euler Backward is unconditionally stable.*

If we assume a more special behaviour of the coefficients b and c, we can also prove asymptotic stability (case II). One should note that in case I stability was established by estimating $\|D_\ell D_{\ell-1}^{-1}\|$. If we can show that $\|D_\ell D_{\ell-1}^{-1}\| \le 1$ then the result for finite intervals can be obtained in a trivial way. Therefore we only give a sharper bound for $\|D_\ell D_{\ell-1}^{-1}\|$ below

PROPERTY 4.7. *Let for all* x, b(t,x) *be nondecreasing and* c(t,x) *be nonincreasing as a function of* t. *Then* $\|D_\ell D_{\ell-1}^{-1}\| \le 1$.

PROOF. Consider the expression in the proof of 4.1. Obviously $b_{\ell-1,s} \le b_{\ell,s}$ and $c_{\ell-1,s} \ge c_{\ell,s}$. From this it follows that $q_{\ell,s} \le q_{\ell-1,s}$, whence $1 + q_{\ell-1,s} \ge 1 + q_{\ell,s}$ and $1 - q_{\ell-1,s} \le 1 - q_{\ell,s}$. The assertion can now be checked easily. □

REMARK 4.8. The requirement for c in 4.7 is not necessarily of this type only. It can be expected that a similar result will also hold for c(t,x) nondecreasing (and b(t,x) nondecreasing). This can be seen as follows: The requirements imposed on D_ℓ in order to let $D_\ell H_\ell D_\ell^{-1}$ be symmetric can also be used to define the elements of D_ℓ in bottom up ordering, i.e. choose (see (3.7))

$$d_{\ell\,N} = \frac{1}{\sqrt{2\,b_{N1}}}, \quad d_{\ell,N-1} = \sqrt{\frac{1}{b_{\ell,N-1}(1+q_{\ell,N-1})}}.$$

For a thus defined D_ℓ, we then will have $\|D_\ell D_{\ell-1}^{-1}\| \le 1$, like in 4.7, if the same condition for b is satisfied, but the counter part for c (the roles of $1-q_{\cdot\cdot}$ and $1+q_{\cdot\cdot}$ are apparently interchanged).

REMARK 4.9. We do not claim that the results in this section are optimal. In fact we hint in the next section at a much better result. Nevertheless this analysis shows some difficulties when Neumann conditions are discretized this way. Of course under the requirement $\Delta t = O(\Delta x^2)$ one can have stability again,but this is an undesired restriction, cf. 3.10 (b).

5. STABILITY FOR VARIABLE COEFFICIENTS; $p_{ij} > 0$

The stepsize restrictions or θ restrictions encountered in the previous section can be relaxed if the p_{ij} are positive. The matrices H_i are no longer singular (even weakly diagonally dominant). By monotonicity arguments we then may estimate the smallest eigenvalue to get rid of the requirement $\mu \ge 0$ (cf. Figure 3.1). This then is used for an other kind of estimation of $\Pi\, B_\ell$, one which resembles more the way we encountered in the first method of analyzing the stability of the trapezoidal rule in §1.

To simplify things we suppose q to be positive. First we have

LEMMA 5.1. *Define* $\beta_i = \min_j b_{ij}$. *Then the smallest eigenvalue of* H_i *is larger than* $\frac{4}{3}\beta_i\Delta t$.

PROOF. We only consider the case N odd (for N even a similar derivation can be given).
(N.B. We thus have $\frac{(N-1)}{2}\Delta x = \frac{1}{2}$). Now define the vector V_i by

$$V_i = (1+(\tfrac{1}{2}-\Delta x)^2,\ldots,1+(\tfrac{1}{2}-N\Delta x)^2)^T.$$

Let $H_iV_i = W_i$, where W_i has as a jth coordinate

$$\frac{\Delta t}{\Delta x^2} b_{ij}[2\Delta x^2 + 2\Delta x\, q_{ij}\ (1-2j\Delta x)]$$

and as first and last coordinates

$$\frac{\Delta t}{\Delta x^2} b_{im}[p_{im} + (\tfrac{1}{2}-m\Delta x)^2\, p_{im} + 2\Delta x - 6\Delta x^2]$$

for m = 1,N respectively.
From the positivity of H_i^{-1} we then find

$$H_i^{-1}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} \le \frac{1}{2\Delta t\beta_i} H_i^{-1} W_i = \frac{1}{2\Delta t\beta_i} V_i.$$

Hence

$$\rho(H_i^{-1}) \le \|H_i^{-1}\|_\infty \le \frac{1}{2\Delta t\beta_i}\|V_i\|_\infty \le \frac{3}{4}\frac{1}{\Delta t\beta_i}.$$

Since the smallest eigenvalue of H_i equals $(\rho(H_i^{-1}))^{-1}$ the result is proved. □

We now proceed as follows:
Write

$$\prod_{\ell=j}^{i} B_\ell = D_{i+1}^{-1}[\prod_j^i \tilde{A}_\ell D_\ell D_{\ell-1}^{-1}]D_j, \tag{5.2}$$

where $\tilde{A}_\ell$ is short for

$$\tilde{A}_\ell = D_{\ell+1}A_\ell D_{\ell+1}^{-1} = (I + \theta_\ell K_{\ell+1})^{-1})(I - (1-\theta)K_{\ell+1} + E_\ell). \tag{5.3}$$

In (5.3) we have set for E_ℓ the matrix

$$E_\ell = D_{\ell+1}(H_\ell - H_{\ell+1})D_{\ell+1}^{-1}. \tag{5.4}$$

In order to estimate ΠB_ℓ by (5.2) we use the result of §4 for $D_\ell D_{\ell-1}^{-1}$; for $\tilde{A}_\ell$ we can use 5.1 if we have an estimate for E_ℓ. Therefore we first estimate E_ℓ.

LEMMA 5.5. *There exists a* κ_ℓ *such that*

$$\|E_\ell\| \le \kappa_\ell \frac{\Delta t^2}{\Delta x}\cdot(1 + o(1)),$$

where κ_ℓ *of the order of* $\max_{1\le j\le N} |b'_{\ell j}|$, *if* b *is differentiable.*

PROOF. The matrix E_ℓ is tridiagonal and differs from $H_\ell - H_{\ell+1}$ only in that the codiagonal elements are multiplied by $d_{\ell i+1}/d_{\ell i}$ or $d_{\ell i}/d_{\ell i+1}$ (see (3.7)). The latter quantities are of order $1 + O(\Delta x)$. To find $\|E_\ell\|_2$ we should consider

$$\rho^{\frac12}(E_\ell E_\ell^T) \le \|E_\ell\|_1^{\frac12}\|E_\ell\|_\infty^{\frac12}. \quad \square$$

By construction $K_{\ell+1}$ (cf. (3.8)) is symmetric. Let $Q_{\ell+1}$ be the orthogonal matrix and $M_{\ell+1}$ be the diagonal matrix such that

$$Q_{\ell+1}^{-1} K_{\ell+1} Q_{\ell+1} = M_{\ell+1}. \tag{5.6}$$

Then

$$Q_{\ell+1}^{-1} \tilde{A}_\ell Q_{\ell+1} = \left(I+\theta_\ell M_{\ell+1}\right)^{-1}\left(I-(1-\theta_\ell)M_{\ell+1}+Q_{\ell+1}^{-1} E_\ell Q_{\ell+1}\right) \tag{5.7}$$

So

$$\|\tilde{A}_\ell\|_2 \le \max_\mu \left|\frac{1+(1-\theta_\ell)\mu}{1+\theta_\ell\mu}\right| + \max_{\mu'} \frac{\|E_\ell\|_2}{1+\theta_\ell\mu'}. \tag{5.8}$$

This then gives the following interesting result:

PROPERTY 5.9. $\|\tilde{A}_\ell\|_2 \le 1$ *if*

(a) $\frac12 \le \theta_\ell \le 1$ *and* $\Delta t < \frac{4}{3}\frac{\beta_\ell}{\kappa_\ell}\Delta x$,

(b) $0 \le \theta_\ell \le \frac12$ *and* $\frac{\Delta t}{\Delta x^2} \le \frac14 \frac{2-\kappa_\ell\frac{\Delta t^2}{\Delta x}}{1-2\theta_\ell} \max\left\{\min_{2\le j\le N-1}\frac{1}{b_{\ell j}}, \min_{1,N}\frac{1}{b_{\ell j}(1+\frac14 P_{\ell j})}\right\}$.

PROOF. Consider

$$\tilde{g}_\ell(\mu) = \frac{1-(1-\theta_\ell)\mu+\varepsilon_\ell}{1+\theta_\ell\mu} .$$

We can draw a graph like Figure 3.1. In order that $\tilde{g}_\ell(\mu) \leq 1$ we must have $\mu > \varepsilon_\ell$. Taking $\varepsilon_\ell = \|E_\ell\|_2$ and using 5.1 and 5.5 given (a). The other part again follows from the observation that $\tilde{A}_\ell$ can have an eigenvalue -1 at worst if

$$\min_{\mu} \frac{1+(1-\theta_\ell)\mu}{1+\theta_\ell\mu} - \min_{\mu'} \frac{\|E_\ell\|_2}{1+\theta_\ell\mu'} = -1. \quad \square$$

The final result now follows from combination of 4.7 and 5.9, see 5.2:

THEOREM 5.10. *Let for all* x, b(t,x) *be nondecreasing and* c(t,x) *be nonincreasing (nondecreasing) as a function of* t. *Then* (2.4) *is asymptotically stable if in* (5.9) (a) or (b) *the stepsize restriction holds for all* ℓ.

REMARK 5.11. It seems quite likely that a similar analysis can also be given for the $p_{ij} = 0$ case (see §4). This can be seen from the fact that all H_i have the eigenvector $(1,1,\ldots,1)^T$ in common. Let this be e.g. the first column of Q_i (cf. (5.6)). Then the first row and column of $Q_{i+1}^{-1}E_iQ_{i+1}$ must be zero. This then implies that in the estimate for $\|\tilde{A}_\ell\|$ only the positive eigenvalues of H_ℓ will play a role. It is not obvious, however, how one can employ monotonicity or something alike to find a lower bound for them.

REFERENCES

[1] AXELSSON, O., *Error estimates for Galerkin methods for quasilinear parabolic and elliptic differential equations in divergence form* (1977), Num. Math. 28, 1-14.

[2] KEAST, P. & A.R. MITCHELL, *On the instability of the Crank Nicholson formula under derivative boundary conditions* (1966), Computer J. 9, 110-114.

[3] KEAST, P. & A.R. MITCHELL, *Finite difference solution of the Third Boundary Problem in Elliptic and Parabolic Equations* (1967), Numer. Math. 10, 67-75.

[4] STOYAN, G.S., *Monotone Difference Schemes for Diffusion-Convection Problems* (1979), ZAMM 59, 361-372.

[5] VARAH, J.M., *On the stability of boundary conditions for separable difference approximations to parabolic equations* (1977), SIAM J. Numer. Anal. 14, 1114-1125.

[6] WIDLUND, O.B., *Stability of parabolic difference schemes in the maximum norm* (1966), Numer. Math. 8, 186-202.

A COMPARISON OF DISCRETIZATION METHODS THAT ARE USED TO SOLVE THE SHALLOW WATER EQUATIONS

N. PRAAGMAN

REMARK. Only a survey of the material presented at the colloquium is given. Detailed results may be found in [6], [7].

The numerical computation of waterlevels and mean velocities in coastal seas has been an important topic for more than twenty years. During the first period, (1960-1970), mathematical models based on finite differences and regular grids have been developed, ([1], [2], [3]), while later on, (1970-1980), models based on finite element techniques and irregular grids have been built ([4], [5], [6]).

Besides these two possibilities for the discretization in space, i.e. finite differences and finite elements, various discretizations in time may be used (see [6], [8]).

In the present study, ([7]) the importance of the used discretization method is investigated by analysing results of computations. These results are obtained utilizing different combinations of space- and time- discretizations.

Furthermore the influence of the empirical values is studied by varying the value of the Chézy parameter, i.e. varying the bottom friction term.

Below is a summary of the items that are treated in [7]:

- the system of partial differential equations that constitutes the mathematical model is considered; attention is paid to boundary conditions and empirical parameters.
- the discretization in space of the system of partial differential equations is considered; the finite element method and the finite difference method are investigated.
- the discretization in time of the partial differential equations is studied.
- a comparison of numerical results for different Chézy values is given.
- a comparison of numerical results for two different space

discretizations is given.

- a comparison of numerical results for two discretization methods in time is given.

REMARK. In order to give an impression of the results that are treated in [7], time histories of the water elevation in a few locations are shown in the figures 1 and 2. The locations of fig 1 are more or less in the "middle" of the sea while the locations of fig 2 are near to "land-boundaries". It is also noted that, although the observed water-elevations behave qualitatively in the same way as the computed water-elevations, (see fig 2), there is quantitatively a quite large discrepancy between the computed values on one side and the observed values on the other side. This may be due, at least partly, to the fact that in the numerical computations no wind-force was taken into account.

The conclusions of the comparison are (see also [7]):

- the magnitude of the empirical parameter of Chézy, and the discretization of the boundary of the region of interest, play an important role with respect to the accuracy and reliability of the numerical method.
- the numerical method that is used for the time integration has only minor influence.

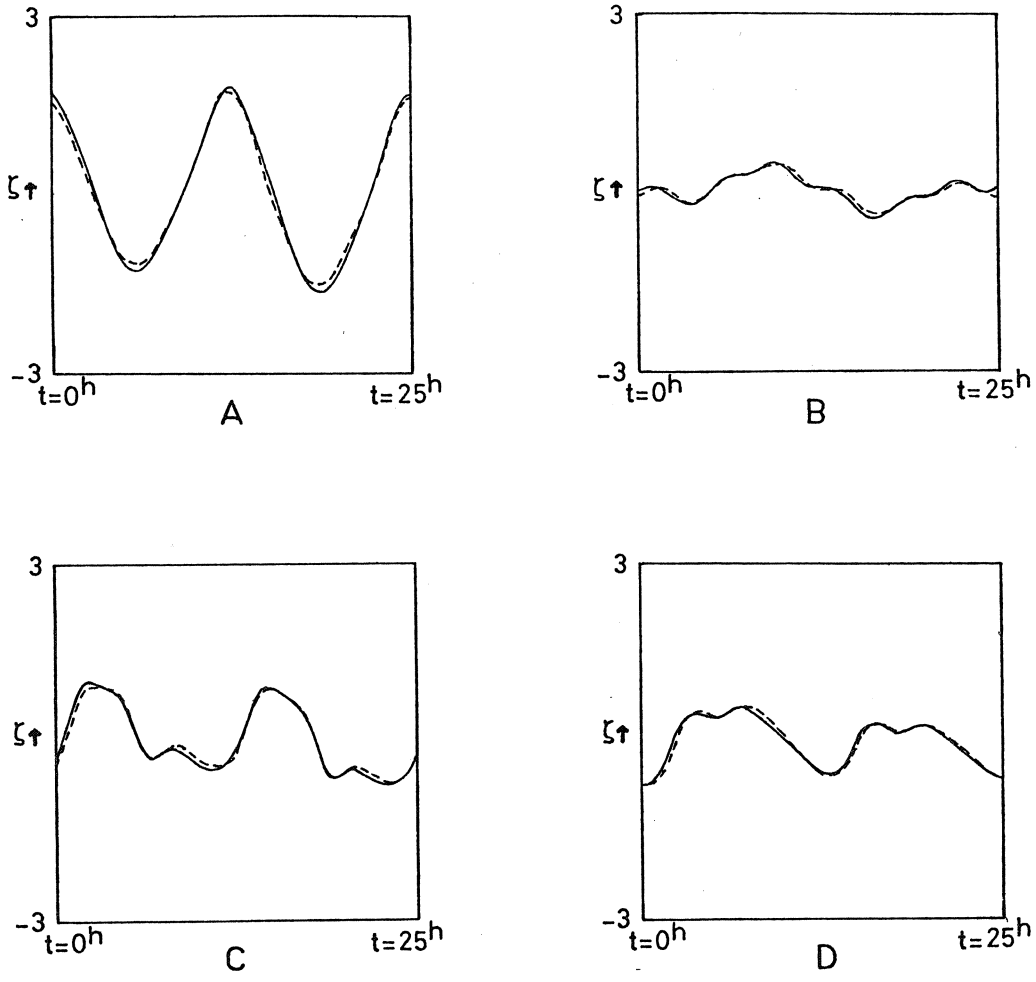
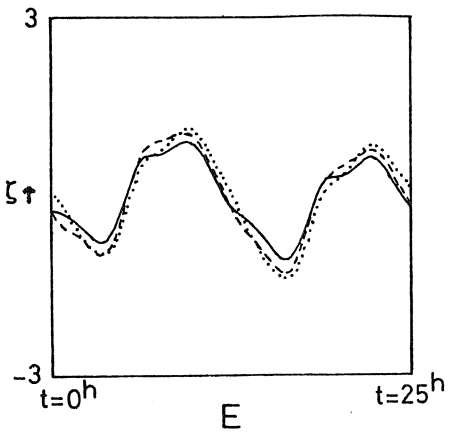
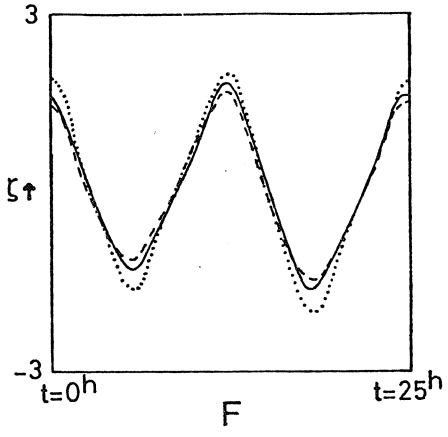


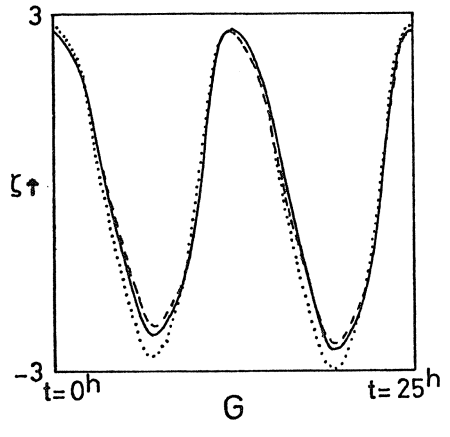
fig. 1 Time histories of the water elevation (in metres) in various locations.

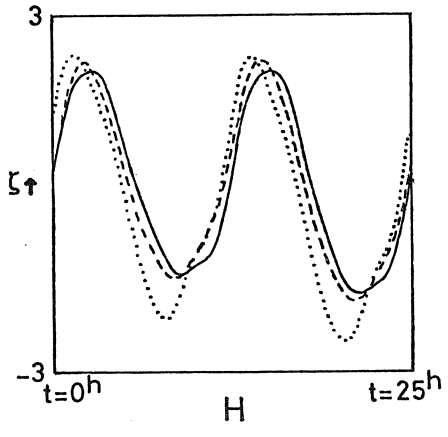
—— finite difference method

--- finite element method









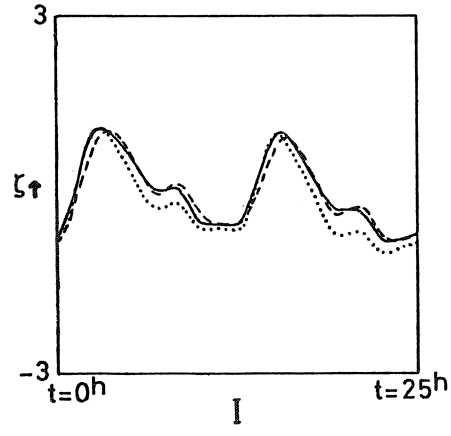


fig. 2 Time histories of the water elevation (in metres) in various locations.

—— finite difference method

--- finite element method

... observed

REFERENCES

[1] FISCHER, G., *Ein numerisches Verfahren zur Errechnung von Windstau und gezeiten in Randmeeren,* Tellus, 11, pp 60-76.

[2] LEENDERTSE, J.J., *Aspects of a computational model for long-period water-wave propagation,* Ph.D., Delft, 1967.

[3] SIELECKI, A., *An energy-conserving difference scheme for the storm surge equations,* Monthly Weather Review, Vol. 96, pp 150-156, 1968.

[4] GROTKOP, G., *Finite element analysis of long period water waves.* Computer methods in applied mechanics and engineering, Vol 2, pp 147-157, 1973.

[5] BREBBIA, C.A. & PARTRIDGE, P.W., *Finite element simulation of water circulation in the North Sea,* Appl. Math. Modelling, Vol 1, pp 101-107, 1976.

[6] PRAAGMAN, N., *A finite element solution of the shallow water equations,* Ph.D., Delft, 1979.

[7] PRAAGMAN, N., *A comparison of discretization methods for the shallow water equations,* NA - report - 33, Delft, 1980.

[8] LAMBERT, J.D., *Computational methods in ordinary differential equations.* John Wiley and Sons, London, 1973.

DISCRETIZATION OF THE CONTINUITY EQUATION FOR THE SOLUTION OF THE NAVIER-STOKES EQUATIONS USING THE FINITE ELEMENT METHOD

A. SEGAL

The Navier-Stokes equations are the equations of motion for a viscous Newtonian fluid. In order to solve these equations we have to add a continuity equation and if necessary a temperature equation and an equation of state.

For an incompressible fluid these equations can be written in the isothermal case as follows:

$$\rho\left(\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j}\right) = -\frac{\partial p}{\partial x_i} + f_i + \mu \Delta u_i \qquad i = 1,2,3 \tag{1}$$

$$\frac{\partial u_i}{\partial x_i} = 0$$

(cartesian coordinates, summation convention used)

with:

ρ the density of the fluid

$\underline{u} = [u_1, u_2, u_3]^T$ the velocity vector

p the pressure

$\underline{f} = [f_1, f_2, f_3]^T$ vector of external forces

μ viscosity

In vector notation the equations (1) become:

$$\frac{\partial \underline{u}}{\partial t} + \underline{u} \cdot \nabla \underline{u} + \frac{1}{\rho} \nabla p - \frac{\mu}{\rho} \Delta \underline{u} = \frac{1}{\rho} \underline{f} \qquad \text{(momentum equations)} \tag{2}$$

$$\operatorname{div} \underline{u} = 0 \qquad \text{(continuity equation)}$$

In this paper we consider the stationary, laminar case, for relatively small Reynolds numbers (Re = $O(10^2) - O(10^3)$). In dimensionless form:

$$
\begin{aligned}
&-\frac{1}{Re}\,\Delta \underline{u} + \underline{u}\cdot\nabla \underline{u} = -\nabla p + \underline{f} \\
&\operatorname{div} \underline{u} = 0.
\end{aligned}
\tag{3}
$$

With Re the Reynolds number.

For simplicity we confine ourselves to Dirichlet boundary conditions, hence

$$
\underline{u}\big|_{\partial\Omega} = \underline{u}_R \quad \text{prescribed.} \tag{4}
$$

One can show that the pressure p is fixed except for an additional constant and that there is a unique velocity $\underline{u}$ when the Reynolds number is small enough.

When solving these equations, either by a finite difference method or a finite element method, two main problems occur.

(i) In the continuity equation only the velocity components appear, not the pressure. This causes severe problems when solving system (3) by a discretization method.

(ii) For large values of the Reynolds number the convective term $\underline{u}\cdot\nabla \underline{u}$ becomes dominant with respect to the diffusive term $\frac{1}{Re}\,\Delta \underline{u}$.

This gives rise to the appearance of boundary layers.

Problem (ii) is the subject of a different chapter (on upwind discretizations of the convection diffusion equation). In order to solve problem (i) there are 3 lines of attack:

(i) Use the stream function-velocity formulation

(ii) Use the stream function formulation

(iii) Solve the equations (3) directly. This is called the primitive variables approach.

Stream function vorticity formulation

In the 2-dimensional space $\mathbb{R}^2$, the continuity equation can exactly be satisfied by the introduction of a stream function ψ according to:

$$
\frac{\partial \psi}{\partial x} = -u_2 \qquad \frac{\partial \psi}{\partial y} = u_1 \qquad (x = x_1;\ y = x_2) \tag{5}
$$

Defining the vorticity ω by:

$$\omega = \frac{\partial u_1}{\partial y} - \frac{\partial u_2}{\partial x} \tag{6}$$

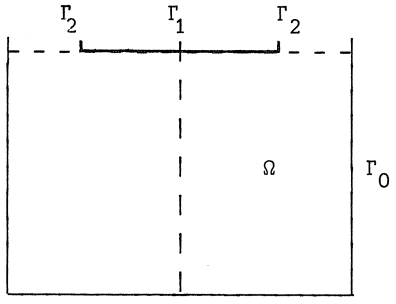
and differentiation of the first equation of motion with respect to x and the second one to y, and subtraction, leads to:

$$\begin{aligned} &\frac{\partial \psi}{\partial y}\frac{\partial \omega}{\partial x} - \frac{\partial \psi}{\partial x}\frac{\partial \omega}{\partial y} - \frac{1}{Re}\Delta\omega = \frac{\partial f_1}{\partial y} - \frac{\partial f_2}{\partial x} \\ &\omega = \Delta\psi \end{aligned} \tag{7}$$

So the pressure has disappeared from the equations.

In general boundary conditions for ψ and not for ω are given. A typical example is the Czochralski crystalpulling process (see LANGLOIS and SHIR [1]), which we will describe briefly.

Consider a rotating cylinder containing some melt. From the centre of the crucible a growing crystal is pulled very gradually. In a simplified model this process can be described by the Navier Stokes equations in cylindrical coordinates. The boundary conditions are given by:



Γ_0 and Γ_1: $\psi = 0$, $\frac{\partial \psi}{\partial n} = 0$

Γ_2: $\psi = 0$, $\frac{\partial^2 \psi}{\partial n^2} = 0$

with Γ_0 the rotating cylinder and
Γ_1 the growing crystal
Γ_2 is the free surface.

The boundary condition along Γ_2 can also be written as:

$$\Gamma_2: \psi = 0,\ \Delta\psi = 0;\ \text{so } \psi = 0 \text{ and } \omega = 0$$

System (7) can be solved using Galerkin's method as follows. We multiply the first equation by a test function $\delta\psi$ and the second one with $\delta\omega$, satisfying the homogeneous boundary conditions:

$$\Gamma_0 \quad \text{and} \quad \Gamma_1: \delta\psi = 0,\ \frac{\partial \delta\psi}{\partial n} = 0 \quad \text{and} \quad \Gamma_2: \delta\psi = 0,\ \delta\omega = 0.$$

So we get:

$$\int_\Omega \{\frac{\partial\psi}{\partial x}\,(\frac{\partial\omega}{\partial x} - \frac{\partial\omega}{\partial y})\,\delta\psi - \frac{1}{Re}\,\Delta\omega\delta\psi\}dx = \int_\Omega (\frac{\partial f_1}{\partial y} - \frac{\partial f_2}{\partial x})\,\delta\psi dx$$

(8)

$$\int_\Omega \{\omega\delta\omega - \Delta\psi\delta\omega\}dx = 0.$$

Partial integration (Gauss theorem) leads to:

$$\int_\Omega \frac{\partial\psi}{\partial x}\,(\frac{\partial\omega}{\partial x} - \frac{\partial\omega}{\partial y})\,\delta\psi dx + \frac{1}{Re}\int_\Omega \nabla\omega\cdot\Delta\delta\psi dx$$

(9)

$$-\frac{1}{Re}\int_\Gamma \frac{\partial\omega}{\partial n}\,\delta\psi ds = \int_\Omega (\frac{\partial f_1}{\partial y} - \frac{\partial f_2}{\partial x})\,\delta\psi dx$$

$$\int_\Omega \{\omega\delta\omega + \nabla\psi\cdot\nabla\delta\omega\}dx - \oint_\Gamma \frac{\partial\omega}{\partial n}\,\delta\omega ds = 0.$$

The surface integrals vanish due to the homogeneous boundary conditions of the test functions. In fact the boundary condition $\frac{\partial\psi}{\partial n} = 0$ appears to be a natural boundary condition. The essential boundary conditions are:

$$\psi = 0, \qquad x \in \Gamma$$

$$\omega = 0, \qquad x \in \Gamma_2.$$

From equation (9) the standard Galerkin discretization immediately follows by approximating:

$$\omega^h = \Sigma\ \omega_j^h(t)\,\phi_j(x)$$

$$\psi^h = \Sigma\ \psi_j^h(t)\,\phi_j(x)$$

and replacing $\delta\omega$ and $\delta\psi$ by ϕ_i. With ϕ_i the standard basis functions.

The equations (9) must be solved by linearizing the system of nonlinear equations. There are several possibilities to linearize these equations, as for example:

$$
\left.
\begin{aligned}
\frac{1}{Re}\int_\Omega \nabla\omega^{k+1}\cdot\nabla\delta\psi dx &= -\int_\Omega \frac{\partial\psi^k}{\partial x}\left(\frac{\partial\omega^k}{\partial x}-\frac{\partial\omega^k}{\partial y}\right)\delta\psi dx \\
&\quad +\int_\Omega \left(\frac{\partial f_1}{\partial y}-\frac{\partial f_2}{\partial x}\right)\delta\psi dx \\
\int_\Omega \nabla\psi^{k+1}\cdot\nabla\delta\omega dx &= -\int_\Omega \omega^k\delta\omega dx
\end{aligned}
\right\} \quad k = 1,2,\ldots \tag{10}
$$

with ψ^1 and ω^1 estimated.

(10) forms the simplest way to linearize the nonlinear equations, however, convergence can only be expected for low Reynolds numbers. Better results can be obtained by taking the convective terms implicitly too. This is for example the case when one applies the classical Newton-Raphson method.

Once the stream function has been calculated the pressure can be computed from the relation

$$
\Delta p = 2\left[\frac{\partial^2\psi}{\partial x^2}\frac{\partial^2\psi}{\partial y^2}-\left(\frac{\partial^2\psi}{\partial x\partial y}\right)^2\right]-\frac{\partial f_1}{\partial x}-\frac{\partial f_2}{\partial y} \tag{11}
$$

This relation can be obtained by taking the divergence of the momentum equations:

$$
-\frac{1}{Re}\Delta\underline{u}+\underline{u}\cdot\nabla\underline{u}+\nabla p=\underline{f} \tag{12}
$$

and using the relation (5).

In order to get boundary conditions for (11) we may take the inner product of (12) and the outer normal along the boundary and use also (5).

Advantages of the stream function-vorticity formulation are that the continuity equation is satisfied exactly and that the number of unknowns is reduced to 2. Furthermore the formulation remains of second order, thus requiring only C^0 continuity of the elements.

Disadvantages are:

- the velocity must be computed by numerical differentiation (inaccurate)
- the computation of the pressure is inaccurate because of the inaccuracy of boundary conditions and right hand side due to numerical differentiations.
- the method can only be applied in $\mathbb{R}^2$.

Stream function formulation

The stream function formulation can easily be derived from (7) by eliminating ω. One obtaines:

$$\frac{\partial\psi}{\partial y}\frac{\partial\Delta\psi}{\partial x} - \frac{\partial\psi}{\partial x}\frac{\partial\Delta\psi}{\partial y} - \frac{1}{Re}\Delta\Delta\psi = \frac{\partial f_1}{\partial y} - \frac{\partial f_2}{\partial x} \tag{13}$$

This formulation has the advantage that only one unknown appears in the equations. However, the equation is of fourth order and so higher order elements (C^1-continuous) are required.

The stream-function-vorticity formulation can be regarded as a mixed formulation of the stream function formulation.

Primitive variables approach

Finite elements formulations involving the primitive variables can be classified as follows:

(i) Integrated formulations in which the velocity components and the pressure are solved for simultaneously.

(ii) Segregated formulations in which the velocity and pressure are uncoupled and hence can be solved for alternatively in an iteration sequence.

(iii) Solenoidal velocity formulations in which the assumed velocity field has zero divergence and hence satisfies the continuity equation exactly.

Integrated formulation

Problems arise with the discretization of the continuity equation. In order to investigate this point one usually considers the linear stationary Stokes equations:

$$\begin{aligned} &-\frac{1}{Re}\Delta\underline{u} + \nabla p = \underline{f} \\ &\qquad\qquad\qquad\qquad x \in \Omega \\ &\operatorname{div}\underline{u} = 0 \\ &\underline{u} = \underline{\alpha} \quad (\text{given} \qquad x \in \Gamma \qquad \underline{u} = (u,v)^T \end{aligned} \tag{14}$$

See for example ARGYRIS and MARECZEK [1], NICKEL, TANNER and CARWELL [1] and TAYLOR and HOOD [1].

One can prove LADYSHENSKAYA [1] that system (14) has exactly one solution when α satisfies $\oint_\Gamma \underline{\alpha}\cdot\underline{n}ds = 0$. This last condition is a necessary one, since from Gauss theorem we get:

$$0 = \int_\Omega \operatorname{div} \underline{u}\, dx = \oint_\Gamma \underline{u}\cdot\underline{n}\, ds = \oint_\Gamma \underline{\alpha}\cdot\underline{n}\, ds \tag{15}$$

The pressure is fixed except for an additional constant.

To solve (14) numerically by the Galerkin method, the momentum equations are multiplied by a test function $\delta\underline{u}$ satisfying

$$\delta\underline{u} = \underline{0} \qquad x \in \Gamma$$

and the continuity equation is multiplied by a test function δp from the space of functions that are given except for an additional constant. So:

$$\begin{aligned} &\int_\Omega (-\frac{1}{Re}\Delta\underline{u} + \nabla p)\cdot\delta\underline{u}\, dx = \int_\Omega \underline{f}\cdot\delta\underline{u}\, dx \\ &\int_\Omega \operatorname{div} \underline{u}\, \delta p dx = 0 \end{aligned} \tag{16}$$

Gauss theorem gives:

$$\begin{aligned} &\int_\Omega \{\frac{1}{Re}\nabla\underline{u}\cdot\nabla\delta\underline{u} - p \operatorname{div} \delta\underline{u}\}dx = \int_\Omega \underline{f}\cdot\delta\underline{u}\, dx \\ &\int_\Omega \delta p \operatorname{div} \underline{u}\, dx = 0 \end{aligned} \tag{17}$$

The first problem that arises is that not every arbitrary element can be used to discretize equation (17). This is caused by the fact that the continuity equation does not contain the pressure as a variable. It is widely accepted (see for example CROUZEIX and RAVIANT [1]) that it is necessary to approximate the pressure with a polynomial one degree less than the polynomial belonging to the velocity approximation. This statement can be made plausible by considering the momentum equations. In these equations second derivatives of the velocity are compared with first derivatives of the pressure. Therefore, in order to get a good balancing, the approximation of the pressure must be one degree less than that of the velocity. When one does not satisfy this requirement one gets singular systems of equations.

However, not all elements with a pressure polynomial of one degree less than the velocity polynomial may be used. For example the simplest element with linear velocities (conforming) and constant pressure yields too many equations for the discretization of the continuity equation (SEGAL [1]). Hence the system is singular. This can be demonstrated easily by the following example:

Consider the element partition of Fig. 1, (i.e. 2 elements). When we solve equation (14) all velocities are prescribed by the boundary conditions. Since the pressure can be computed except for an additional constant one of the pressure components must be prescribed. So in fact we have one unknown pressure and no unknown velocities. As a consequence $\delta\underline{u}$ is not free, and δp has one degree of freedom. So the continuity equation is discretized by one equation, containing zero unknowns, whereas the momentum equations, containing one unknown, are not discretized at all.

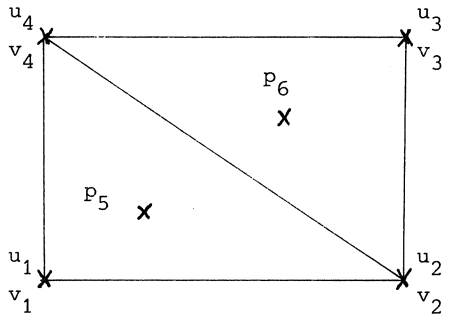


Fig. 1. simple element configuration of 2 elements

Hence the system of equations is singular. The situation becomes even worse when considering more elements. For example in Fig. 2 there are 7 equations to compute 2 unkown velocities and 2 equations to compute 7 unknown pressures:

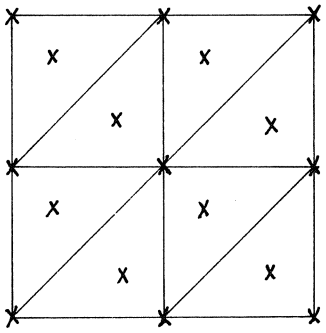


Fig. 2. "8 elements"

In SEGAL [1] some triangular elements that can be used to discretize (17) are treated. Simple elements are for example:

(i) linear non-conforming element:

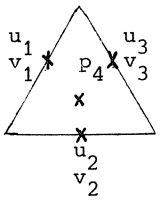


Fig. 4.3

velocity components linear non-conforming.
nodal points: midsides of elements
pressure constant in element.

(ii) conforming quadratic:

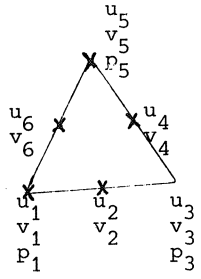


Fig. 4.4

velocity components quadratic and pressure linear. Pressure field as well as velocity field C^0 continuous over the whole domain.

(iii) extended conforming quadratic:

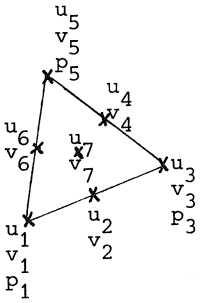


Fig. 4.5

velocity components complete quadratics including one third degree term. Nodal points: vertices, midpoints of the sides as well as the centre of the element. Pressure linear, but not continuous over the whole domain. So in a vertex the value of the pressure may differ from one element to the other.

Also quadrilateral elements may be used, as reported in HUGHES, TAYLOR and LEVY [1].

Summarizing we may state that only a limited number of elements may be used to solve problem (14). One can prove (SEGAL [1]) that the elements mentioned discretize the continuity equation such that exactly one of these discretized equations depends on the others. The dependency is such that

$$(18) \qquad \oint_{\partial\Omega_h} (\underline{u}^h)^T \cdot \underline{n} \, ds = 0$$

is the natural discretization of the compatibility condition (15).

Advantages of the integrated formulation are that the velocity and pressure variables are computed directly, disadvantages are the large number of unknowns and the appearance of a large diagonal block of zeros in the discretization of the continuity equation due to the absence of the pressure variable.

If we number the unknowns for example in the sequence $u_1, u_2, \ldots, u_n, v_1, v_2, \ldots, v_n, p_1, p_2, \ldots, p_m$ then the system of equations can be rewritten as

$$\begin{bmatrix} \Delta^h & 0 & D_1^T \\ 0 & \Delta^h & D_2^T \\ D_1 & D_2 & 0 \end{bmatrix} \begin{bmatrix} \underline{u}^h \\ \underline{v}^h \\ \underline{p}^h \end{bmatrix} = \begin{bmatrix} \underline{f}_1 \\ \underline{f}_2 \\ \underline{0} \end{bmatrix} \tag{19}$$

Direct solution of this system with for example a profile method causes no numerical problems, however the profile becomes very large.

A better numbering is: $u_1, v_1, p_1, u_2, v_2, p_2, \ldots$ etc., however this numbering leads to zeros on the diagonal in the first few rows and therefore partial pivoting must be applied. Hence more computer memory and time is required.

A more sophisticated numbering of the nodal points may remove this problem. In one test run a saving with a factor 2.5 could be obtained.

A disadvantage of the last method is that it is difficult to program.

To overcome these difficulties other methods have been developed. These methods are such that they solve the system of Galerkin equations in a fast way. Hence we must use the same elements as for the integrated formulation.

Segregated formulation

An alternative way to handle the primitive variables is to uncouple the velocity and pressure and adopt a segregated method of solution. Methods to do this are based on the theory of minimizing functionals with constraints.

Therefore again we consider the stationary Stokes equation (14). The variational formulation is given by (17).

Let V be the space of solenoidal vector functions, that is functions $\underline{v}$ satisfiying

$$\operatorname{div} \underline{v} = 0 \tag{20}$$

Clearly, the solution $\underline{u}$ is an element of this space. Now let test functions $\delta\underline{u}$ also satisfy this requirement, so:

$$\text{(21)} \qquad \text{div } \delta\underline{u} = 0$$

Then (17) reduces to:

$$\text{(22)} \qquad \int_\Omega \{\frac{1}{Re} \nabla\underline{u}\cdot\nabla\delta\underline{u}\}dx = \int_\Omega \underline{f}\cdot\delta\underline{u}dx$$

with $\underline{u}$, $\delta\underline{u} \in V$

$$\underline{u}\Big|_\Gamma = \underline{\alpha} \qquad \delta\underline{u}\Big|_\Gamma = \underline{0}$$

(22) is the variational formulation for the Poisson equation, however in the space of solenoidal vector functions. For this equation we can derive an equivalent minimizing problem (see STRANG and FIX [1], MITCHELL and WAIT [1], ZIENKIEWICZ [1], VAN KAN and SEGAL [1])

$$\text{(23)} \qquad \min_{\underline{u}\in V} \int_\Omega \{\frac{1}{2Re} (\nabla\underline{u})^2 - \underline{u}\cdot\underline{f}\}dx$$

with $\underline{u}\Big|_\Gamma = \underline{\alpha}$.

This can be seen as a minimizing problem with the constraint

$$\text{div } \underline{u} = 0.$$

Notice that the pressure has disappeared from these formulas.

To discretize equation (23) we introduce the discretized space V^h defined by elements $\underline{v}^h$ of the finite element solution space satisfying:

$$\text{(24)} \qquad \text{div}^h \underline{v}^h = 0$$

With $\text{div}^h v^h = 0$ we mean: $\int_\Omega \delta p \text{ div } \underline{v}^h dx = 0$, $\forall\delta p$, i.e. the standard discretization of the continuity equation. Compare with (17).

Hence the discretization of (23) becomes:

$$\text{(25)} \qquad \min_{\underline{u}^h\in V^h} \int_\Omega \{\frac{1}{2Re} (\nabla u^h)^2 - \underline{u}^h\cdot\underline{f}\}dx$$

with $\underline{u}^h\big|_\Gamma = \underline{\alpha}$.

From the theory of optimizing with constraints several solution techniques are known for this problem, for example: The penalty function method and the method of Powell and Hestenes (SEGAL [1]).

In this chapter we limit ourselves to the penalty function approach. There are 2 ways in order to apply this method:

(i) The penalty function procedure can be applied to the continuous problem (23) and the result can be discretized.

(ii) The method can be applied to the discretized problem (25).

The penalty function method applied to the discrete problem (25)

The minimizing problem (25) must be solved over the space of functions $\underline{u}^h$ satisfying the constraint:

$$\int_\Omega \delta p \ \mathrm{div}\ \underline{u}^h dx = 0 \quad \forall \delta p \tag{26}$$

By introducing basis functions for the solution and test spaces, (26) can be written as a system of m linear equations with n unknowns ($m \leq n$):

$$L\underline{u}^h = 0 \tag{27}$$

Exactly one of these equations depends on the others.

Discretizing the standard Poisson equation without constraints gives the system of linear equations:

$$\frac{1}{Re} S\underline{u}^h = \underline{f} \tag{28}$$

In the penalty function method we solve:

$$\frac{1}{Re} S\underline{u}^h + \sigma L^T L\underline{u}^h = \underline{f} \tag{29}$$

with $\sigma > 0$ large ($\sigma = O(10^6)$). The matrix $\frac{1}{Re} S + \sigma L^T L$ is positive definite.

One can show (SEGAL [1]) that the solution of (29) converges to the solution of (25) when $\sigma \to \infty$.

For the elements with discontinuous pressure, as for example the linear non-conforming and the extended quadratic element, the computation of the matrix $L^T L$ can be easily carried out. The structure of the matrix is

identical to that of the matrix S. However for the quadratic element with linear continuous pressure this is not the case. The structure of the matrix L^TL essentially differs from that of the matrix S. In fact the band width and the profile get larger and therefore for this type of elements the method must not be applied.

The penalty function method applied to the continuous problem (23)

One easily verifies that, since L is the discretization of the divergence operator $-L^T$ is the discretization of the gradient operator. Hence a continuous translation of the penalty function (29) is:

$$(30) \qquad -\frac{1}{Re}\Delta\underline{u} + \sigma \text{ grad div } \underline{u} = \underline{f}$$

$\sigma > 0$ large. Equation (30) can be discretized in a standard way by the Galerkin method yielding:

$$(31) \qquad \frac{1}{Re} S\underline{u}^h + \sigma A\underline{u}^h = \underline{f}$$

In general the matrix A is not equal to the matrix L^TL.

Of course a necessary condition is that when $A\underline{u}^h = 0$: $L\underline{u}^h = 0$. In SEGAL [2] it is proved that this is indeed the case for cartesian coordinates. However in the case of axial symmetric coordinates this is not true. Indeed practical computations show that this method converges to wrong answers for axial-symmetric coordinates (SEGAL [2]).

Summarizing we may state that the continuous penalty method can be programmed more easily than the discrete one. However the method cannot be applied to axisymmetric problems.

Advantages of a penalty function method versus an integrated solution are:

(i) Easier to program.

(ii) Less unknowns

(iii) No partial pivoting is necessary hence a considerable reduction of computing time and computer memory.

Disadvantages are:

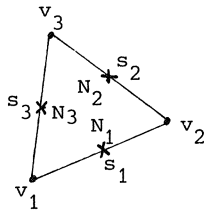
(i) For large Reynolds numbers the convergence for the Navier Stokes equations is slow.

(ii) The computation of the pressure might be inaccurate depending on the method used (see SEGAL [2]).

Solenoidal velocity formulation

A further alternative for the primitive variable approach is the use of solenoidal velocity interpolations, that is ones which satisfy the discretized continuity equations exactly. The first formulation in this field was from FORTIN [1]. However, his method needs complete quartic triangles with 30 degrees of freedom.

GRIFFITHS [1] constructs a solenoidal vector field by introducing a mixed stream function velocity formulation. He is able to do this for example for the linear non-conforming triangle and the extended quadratic. He claims good results with this method, which reduces computation time and memory, compared to the integrated formulation. We shall demonstrate his method using the linear non-conforming element applied to the Stokes equations (14).



The method is based on the fact that the space of vector functions can be split into 2 orthogonal vector spaces D(Ω) and I(Ω) with D(Ω) the space of solenoidal velocity fields and I(Ω) the space of irrotational velocity fields. So each vector $\underline{u}$ therefore can be written as $\underline{u} = \underline{u}_1 + \underline{u}_2$ with $\underline{u}_1 \in D(\Omega)$ and $\underline{u}_2 \in I(\Omega)$. The variational formulation to compute the velocity $\underline{u}$ and the pressure p can be written as:

$\underline{u} \in D(\Omega) \rightarrow \operatorname{div} \underline{u} = 0$

$\underline{u} \in I(\Omega) \rightarrow \operatorname{curl} \underline{u} = 0.$

$$(32) \qquad \int_\Omega \frac{1}{Re} \nabla\underline{u}\cdot\nabla\delta\underline{u}dx = \int_\Omega \underline{f}\cdot\delta\underline{u}dx \qquad \underline{u},\ \forall\delta\underline{u} \in D(\Omega)$$

with $\underline{u}\big|_\Gamma = \underline{\alpha}$ and $\delta\underline{u}\big|_\Gamma = \underline{0}$ and

$$(33) \qquad \int_\Omega p \operatorname{div} \delta\underline{u}dx = \frac{1}{Re} \nabla\underline{u}\cdot\nabla\delta\underline{u}dx - \int_\Omega f\cdot\delta\underline{u}dx \qquad \forall\delta\underline{u} \in I(\Omega)$$

with $\delta\underline{u}\big|_\Gamma = 0$. See (17) and (22).

With (32) we can conpute $\underline{u}$, provided we are able to construct soleniodal vector functions. Given $\underline{u}$ we can compute p from (33) (except for an additive constant), provided we are able to construct irrotational vector functions.

The velocity computation is completely uncoupled from the pressure computation. For example in time dependent problems, when we are interested in the pressure for only a limited number of points of time this is an important feature.

The space $D(\Omega)$ is characterized by the existence of a stream function defined by:

$$\frac{\partial \psi}{\partial x} = -u_2 \qquad \frac{\partial \psi}{\partial y} = u_1 \tag{34}$$

(Compare with (5)).

Now consider the discretization of the continuity equation div $\underline{u} = 0$ given by (17)

$$\int_{\Omega} \text{div } \underline{u}^h \delta p \, dx = 0 \qquad \forall \delta p \tag{35}$$

Since δp is a constant for all elements (35) can be written as:

$$\int_{e} \text{div } \underline{u}^h dx = 0 \qquad \forall e \tag{36}$$

Since $\underline{u}^h$ is linear in each element this implies

$$\text{div } \underline{u}^h = 0 \qquad \text{in each element.} \tag{37}$$

So in each element there exists a stream function ψ^h defined by:

$$\underline{u}^h = (\psi^h_y, -\psi^h_x)^T \tag{38}$$

The idea is now to introduce new nodal parameters ψ^h_i in each vertex through the relations:

$$\int_{s_\ell} \underline{u}^h \cdot \underline{n}_\ell ds = \psi^h_{\ell+1} - \psi^h_\ell \qquad \ell = 1,2,3 \tag{39}$$

$(\psi^h_4 \equiv \psi^h_1)$

with s_ℓ the ℓ'th side of the triangle and $\underline{n}$ the outward directed normal on s_ℓ.

As suggested by the notation ψ^h_ℓ provides an approximation to the stream function ψ at vertex v_ℓ. The stream function ψ^h can be computed except for

an additional constant. This is in accordance with the fact that one of the equations (16) depends on the others.

The method is now to eliminate 3 nodal velocities in each element and to replace them by the stream function parameters using relation (39).

Therefore the approximation $\underline{u}^h$ is written in the following form:

$$\underline{u}^h\big|_e = \sum_{j=1}^{3} u_t^h(N_j)\underline{\xi}_j(x,y) + u_n^h(N_j)\eta(x,y) \tag{40}$$

with N_j the midpoint of side s_j, $u_t^h(N_j)$ the tangential component of $\underline{u}^h$ in N_j and $u_n^h(N_j)$ the normal component of $\underline{u}^h$ in N_j. $\underline{\xi}_j$ and $\underline{\eta}_j$ are linear vector basis functions corresponding to these parameters. These basis functions can be computed easily.

One immediately verifies that (39) can be written as:

$$\int_{s_\ell} \underline{u}^h \cdot \underline{n}_\ell ds = \psi_{\ell+1}^h - \psi_\ell^h = L_\ell u_n^h(N_\ell) \qquad \ell = 1,2,3 \tag{41}$$

where L_ℓ denotes the length of side s_ℓ.

With the aid of (41) the normal components of $\underline{u}^h$ are eliminated and (40) can be written as:

$$\underline{u}^h\big|_e = \sum_{j=1}^{3} u_t^h(N_j)\underline{\xi}_j(x,y) + \psi_j^h\left[\frac{1}{L_{j-1}}\underline{\eta}_{j-1}(x,y) - \frac{1}{L_j}\underline{\eta}_j(x,y)\right] \tag{42}$$

with $\underline{u}^h \in D^h(\Omega)$ the discretization of $D(\Omega)$.

The eliminated unknowns are used to approximate the space $I(\Omega)$. It follows that, if

$$\underline{u}^h\big|_e = \sum_{j=1}^{3} u_n^h(N_j)\underline{\eta}_j(x,y) \tag{43}$$

then $\underline{u}^h \in I^h(\Omega)$, the approximation of $I(\Omega)$. The implementation of the divergence-free basis functions leads to the algebraic system

$$\hat{A}\hat{\underline{u}} = \hat{\underline{F}} \tag{44}$$

with

$$(\hat{A})_{ij} = \mu \int_\Omega \nabla\hat{\underline{\phi}}_j \cdot \nabla\hat{\underline{\phi}}_j dx$$

$$(\hat{F})_i = \int_\Omega \underline{f}\cdot\underline{\phi}_i \, dx$$

$\{\hat{\phi}_i\}$ are the basis functions of the space $D^h(\Omega)$ and $\hat{\underline{u}}$ contains the nodal parameters for both velocity (tangential components only) and stream function.

From (41) it follows that there exists a matrix R such that:

$$\underline{u}_v^h = R\hat{\underline{u}} \tag{45}$$

with $\underline{u}_v^h$ the (2N×1) vector representation of $\underline{u}^h$, i.e. the vector containing the values of $\underline{u}^h$ in the nodal points.

Now the matrix $\hat{A}$ can be written as $R^T A R$ and the vector $\hat{F}$ as $R^T F$ with:

$$(A)_{ij} = \mu \int_\Omega \nabla\underline{\phi}_j \cdot \nabla\underline{\phi}_j \, dx$$

$$(F)_i = \int_\Omega \underline{f}\cdot\underline{\phi}_i \, dx$$

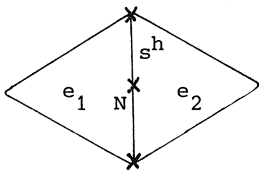
with $\{\underline{\phi}_i\}$ the classical basis functions without constraints. So in fact we can compute the element matrix A^e and the element vector F^e with a standard element package, and compute the element matrix $\hat{A}^e$ and the element vector $\hat{F}^e$ by the transformations:

$$\hat{A}^e = R^{e^T} A^e R^e \qquad \hat{F}^e = R^{e^T} F^e \tag{46}$$

with R^e given by relation (41).

REMARK. One of the stream function values must be given, for example: $\psi_1 = 0$.

The computation of the pressure from (33) and (43) is straight forward.



Consider two adjacent elements e_1 and e_2 with a common side s^h of length L and unit vector $\underline{n}$ directed out of e_1 and into e_2.

Applying equation (33) with $\delta\underline{u}$ is chosen as the basis function associated with the mid point N of s^h in the direction of $\underline{n}$ gives:

$$(47) \qquad (p_1^h - p_2^h)L = \int_\Omega \left(\frac{1}{Re}\,\nabla \underline{u}^h \cdot \nabla \delta \underline{u} - \underline{f} \cdot \delta \underline{u}\right) dx$$

where p_1^h and p_2^h are the constant pressures on e_1 and e_2 respectively. Thus, given the pressure on any one element of the mesh, the pressure on remaining elements can be computed explicitly one element at a time through (47).

REFERENCES

[1] ARGYRIS, J.H. & G. MARECZEK, *Finite element analysis of the slow incompressible viscous fluid motion,* Ingenieur-Archive, 43, 1974, p. 92-109.

[1] CROUZEIX, M. & P.A. RAVIANT, *Conforming and non-conforming finite element methods for solving the stationary Stokes equations,* Rev. Française Automat. Informat. Recherche Opérationelle, 7, 1973, p. 33-76.

[1] CUVELIER, C., N. PRAAGMAN & A. SEGAL, *A survey of finite element methods in fluid mechanics,* Report NA-26. Delft University of Technology, 1979.

[1] FORTIN, M., *Calcul Numérique des écoulements des fluides de Bingham et des fluides newtonien incompressible par la méthode des éléments finis,* Thèse de doctorat, Université de Paris Vi, 1972.

[1] GRIFFITHS, D.F., *Finite elements for incompressible flow,* Math. Meth. in Applied Science, 1, 1979, p. 16-31.

[1] KAN, J. VAN & A. SEGAL, *Dictaat numerieke analyse* CII/BIII, Technische Hogeschool Delft, onderafdeling der Wiskunde, Delft, Nederland, (Dutch).

[1] LADYSHENSKAYA, O.A., *The mathematical theory of viscous incompressible flow,* Gordon and Breach, New York (1969), (English translation).

[1] LANGLOIS, W.E. & C.C. SHIR, *Digital simulation of flow patterns in the Czochraliski crystal-pulling process,* Comp. Meth. in Meth. and Eng, 12, 1977, p. 145-152.

[1] MITCHELL, A.R. & R. WAIT, *The finite element method in partial differential equations,* John Wiley & Sons, Chichester, 1977.

[1] NICKEL, R.E., R.I. TANNER & J. CARWELL, *The solution of viscous, incompressible jet and free-surface flow using finite element methods,* J. Fluid Mechanics, 65, 1974, p. 189-206.

[1] SEGAL, A., *On the numerical solution of Stokes equations using the finite element method,* Comp. Meth. in Appl. Mech. and Engng., 19, 1979, p. 165-185.

[2] SEGAL, A., *The numerical solution of Navier-Stokes equations by the finite element method,* Report NA-28, Delft University of Technology, (to appear).

[1] STRANG, G. & G.J. FIX, *An analysis of the finite element method,* Prentice Hall Inc., Englewood Cliffs, N.J., 1973.

[1] TAYLOR, C. & P. HOOD, *A numerical solution of the Navier-Stokes equations using the finite element technique,* Comp. and Fluids, 1, 1973, p. 73-100.

[1] ZIENKIEWICZ, O.C., *The finite element method in engineering science,* McGraw-Hill, London, 1971.

ON UPWIND DISCRETIZATIONS OF THE CONVECTION DIFFUSION EQUATION

A. SEGAL

1. INTRODUCTION

The purpose of our research is the numerical solution of the Navier-Stokes equations, which in the stationary, isothermal, incompressible case can be written as:

$$-\frac{1}{Re}\Delta\underline{u} + (\underline{u}\cdot\nabla)\underline{u} + \nabla p = \underline{f}$$

$$\operatorname{div}\underline{u} = 0$$

with $\underline{u}$ the velocity vector, p the pressure and Re the Reynolds number. One of the problems that arises when solving these equations for large Reynolds numbers is the presence of boundary layers where the solution has steep gradients.

In order to analyse the numerical problems related with boundary layers we consider the convection diffusion equation as a model problem:

$$-\varepsilon\Delta\phi + \underline{u}\cdot\nabla\phi = f, \qquad x \in \Omega$$

$$\phi\big|_{\partial\Omega} \qquad \text{given}$$

with ε small and $\underline{u} = O(1)$.

When solving the convection diffusion equation we must keep in mind the characteristics that the solution of the Navier-Stokes equations generally satisfy, i.e.

- in the internal region the solution is more or less smooth.
- along some boundaries steep gradients may occur (boundary layers).

Furthermore, in general the region under consideration is not as simple as for example a rectangle.

The investigation of the convection diffusion equation for small values of the parameter ε has been the subject of many papers. See for example HEMKER [1], PEARSON [2,3], IL'IN [4], HEINRICH et al. [5,6], GRIFFITHS [7], CHRISTIE et al. [8], AXELSSON and GUSTAFSSON [9], and CHIEN [10].

It is the aim of this paper to compare the so-called upwind schemes given in some of these papers with standard central difference schemes and finite element schemes, both for direct solution methods and for iterative solution methods.

It is shown that in the context of Navier-Stokes equations for not too large Reynolds numbers central difference schemes together with mesh refinement are preferable to upwind schemes. When the Reynolds number increases the best results can be achieved by a combination of a central difference scheme and some upwind scheme using a so-called defect correction method. When an iterative method is used as for example a preconditioned unsymmetric conjugate gradient method this last method also gives the best results for smaller Reynolds numbers.

2. BOUNDARY LAYERS

The problems mentioned in Chapter 1 will be investigated using the simple one dimensional convection diffusion equation as a test case:

$$(2.1)\qquad \begin{cases} -\varepsilon \dfrac{d^2\phi}{dx^2} + u \dfrac{d\phi}{dx} = 0 & x \in (0,1) \\ \phi(0) = 0 \quad \phi(1) = 1 \end{cases}$$

An example of an exact solution of (2.1) has been plotted in Fig. 2.1. Equation (2.1) can be discretized either by a finite difference method (fdm) or a finite element method (fem).

Central differences with constant mesh size give the following scheme:

$$(2.2)\qquad \begin{aligned} &\frac{1}{h^2}\{-(\varepsilon+\tfrac{uh}{2})y_{i-1} + 2\varepsilon y_i - (\varepsilon-\tfrac{uh}{2})y_{i+1}\} = 0 && i = 1,2,\ldots,n \\ &y_0 = 0 \quad y_{n+1} = 1 \end{aligned}$$

with y_i the numerical approximation of $\phi(x_i)$.

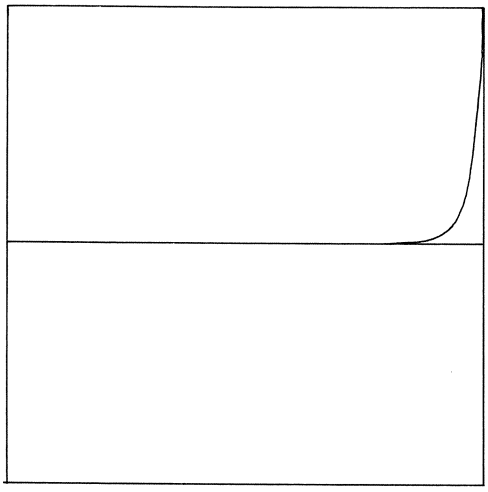


Fig. 2.1. exact solution of (2.1)
$\varepsilon = 0.025$; $u = 1$

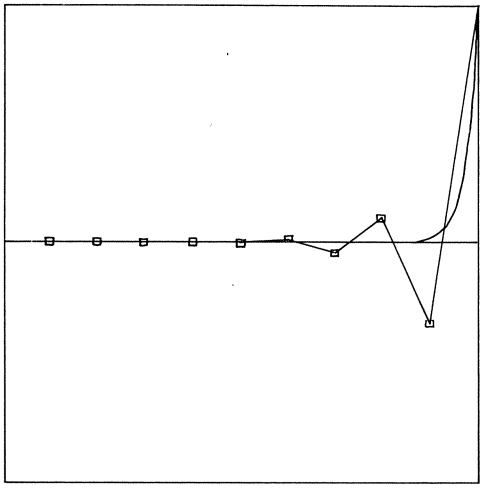


Fig. 2.2. numerical solution of (2.1) using central differences
$\varepsilon = 0.025$; $u = 1$; $h = 0.1$
——— exact solution
—□— numerical solution

Fig. 2.2 shows that the numerical solution oscillates around the exact solution in the case $\varepsilon = 0.025$, $u = 1$ and $h = 0.1$. In the case $h = 0.025$ the solution appears to be monotone and the exact solution cannot be distinguished from the numerical one in the plot.

In order to explain this phenomenon we consider the following set of difference equations:

$$cy_{i-1} + by_i + ay_{i+1} = 0 \qquad (2.3)$$
$$y_0 = 0 \quad y_{n+1} = 1$$

with $b > 0$, $c < 0$ and

$$a + b + c = 0.$$

Note that (2.2) is a special case of (2.3).

THEOREM 2.4. *The solution of the difference equations* (2.3) *is monotone iff* $a \leq 0$.

PROOF. See SEGAL [11].

REMARK. In fact the theorem states that in order to have monotone solutions it is necessary and sufficient that the matrix A corresponding to (2.3) is diagonally dominant.

$$A = \begin{bmatrix} b & a & & & \\ c & b & a & & 0 \\ & c & b & a & \\ & & & & \\ & 0 & & c & b & a \\ & & & & c & b \end{bmatrix}$$

Application of Theorem (2.4) to the central difference scheme (2.2) yields the well-known condition:

$$\frac{uh}{2\varepsilon} \leq 1 \tag{2.5}$$

The quantity $\frac{uh}{\varepsilon}$ is called Péclet number by physicists; when ε is the viscosity coefficient it is called the Reynolds number. For small values of ε the condition (2.5) is a severe restriction on the mesh size. Therefore one often tries to find other difference schemes in order to have diagonally dominant matrices independently of the mesh size h.

Upwind differencing

The simplest method to avoid oscillations is to approximate $\frac{d\phi}{dx}$ by backward differences if $u > 0$, and by forward differences if $u < 0$. In the literature this method is known by the name "upwind differencing" or "upstream differencing" (see ROACHE [13]). In this way we get the following system of difference equations:

$$\begin{aligned} &\frac{1}{h^2}\{-(\varepsilon+uh)y_{i-1} + (2\varepsilon+uh)y_i - \varepsilon y_{i+1}\} = 0 \qquad i = 1,2,\ldots,n \\ &y_0 = 0 \quad y_{n+1} = 1 \end{aligned} \tag{2.6}$$

The corresponding matrix is diagonally dominant, and due to Theorem (2.4), the solution is monotone.

In order to analyse the performance of this method we consider the

truncation error:

$$e_i = \frac{1}{h^2}\{-(\varepsilon+uh)\phi_{i-1} + (2\varepsilon+uh)\phi_i - \varepsilon\phi_{i+1}\} + \varepsilon\frac{d^2\phi_i}{dx^2} - u\frac{d\phi_i}{dx}$$

(2.7)

$$= -\frac{uh}{2}\frac{d^2\phi_i}{dx^2} + O(h^2).$$

So in fact instead of the d.e. (2.4) we approximately solve the d.e.:

$$-(\varepsilon + \frac{uh}{2})\frac{d^2\phi}{dx^2} + u\frac{d\phi}{dx} = 0$$

(2.8)

$$\phi(0) = 0 \quad \phi(1) = 1$$

The term $-\frac{uh}{2}\frac{d^2\phi}{dx^2}$ is called artificial viscosity.

When one discretizes equation (2.8) with central differences, equation (2.6) arises. Results of (2.6) are plotted in Fig. (2.3) and Fig. (2.4).

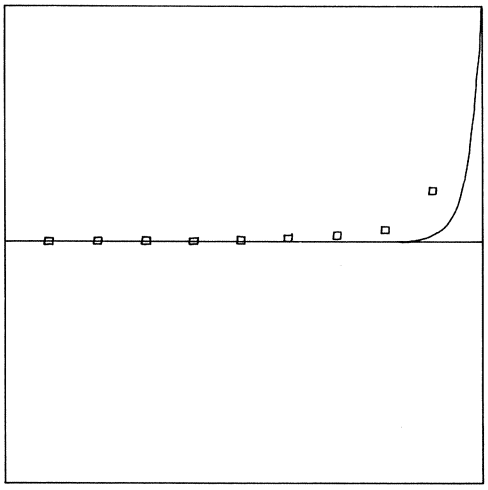


Fig. 2.3. $\varepsilon = 0.025$, $u = 1$, $h = 0.1$

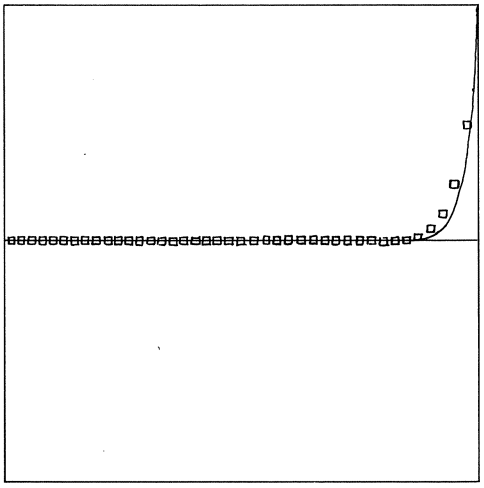


Fig. 2.4. $\varepsilon = 0.025$, $u = 1$, $h = 0.025$

Solution of (2.1) using the backward difference scheme (2.6)

—— exact solution

□ numerical solution

We see that the solution is monotone, however due to the artificial viscosity the numerical solution is far less "steep" than the exact solution. The artificial viscosity has a so-called "smoothing" effect on the solution.

REMARK. The boundary layer has thickness $O(\varepsilon)$. The thickness of the numerical boundary layer is $O(\varepsilon + \frac{uh}{2})$. We only get the correct boundary layer thickness if $\frac{uh}{2} < \varepsilon$ which leads to the undesirable criterion (2.5).

In the literature more accurate upwind schemes have been derived, see for example [9], [14] and [4]. The most accurate scheme with respect to equation (2.1) is the so-called IL'IN scheme [4,1,10]. This method supplies equation (2.4) with an artificial viscosity α such that the solution of:

$$(2.9) \qquad \frac{1}{h^2}\{-(\varepsilon+\alpha+\frac{uh}{2})y_{i-1} + (2\varepsilon+2\alpha)y_i - (\varepsilon+\alpha-\frac{uh}{2})y_{i+1}\} = 0$$

is exact in the nodal points.

One easily verifies that this is the case when

$$(2.10) \qquad \begin{aligned} \alpha &= -\varepsilon + \frac{\varepsilon}{G} \\ G &= \frac{2}{R}\{1 - \frac{2(e^R-1)}{e^{2R}-1}\} \\ R &= \frac{uh}{\varepsilon} \end{aligned}$$

The Il'in scheme is of course the most accurate one for the equation (2.1). However, when variable coefficients are used its accuracy decreases, although a good approximation remains possible as long as the coefficients vary not too much. Il'ins method can be extended to more dimensional rectangular regions, however, for general meshes with local refinements no generalization is known.

One can show ([11]) that in order to get a diagonally dominant matrix it is necessary to introduce an artificial viscosity $\alpha \frac{d^2u}{dx^2}$ with:

$$(2.11) \qquad \begin{aligned} &\alpha \geq \frac{uh}{2} - \varepsilon \qquad \text{for } \varepsilon \text{ small} \\ &\text{and} \\ &\alpha \geq 0 \end{aligned}$$

The finite element method

The oscillations noticed for the central difference scheme is not specific for a fdm. In applications of the fem the same problems occur. (See [5-8, 15,17-21].)

This is something one should expect, since the FEM is a tool to construct finite difference equations. For example, the Galerkin method together with linear elements applied to equation (2.1) gives the central difference scheme (2.2). So exactly the same results can be expected. Higher order elements too give oscillations [19,20,21].

In the literature one has tried to construct finite element models that gave upwind differencing schemes. This approach is described in [5-8]. The method used is the so-called Petrov-Galerkin method (see GRIFFITH [6]), i.e. the test functions differ from the basis functions. This method will be illustrated using an example of ZIENKIEWICZ [23].

EXAMPLE. Consider the convection-diffusion equation (2.1)

$$-\varepsilon \frac{d^2u}{dx^2} + v \frac{du}{dx} = f \qquad x \in (0,1). \tag{2.12}$$

We use piecewise linear basis functions (shape functions) ϕ_i and test functions (weighting functions) w_i, $i = 1,2,\ldots,n$. The contribution to one element of the basis functions is given by:

$$\begin{aligned} \phi_{i-1}(x) &= \lambda_1(x) = \tfrac{1}{2}(1-\zeta) \\ \phi_i(x) &= \lambda_2(x) = \tfrac{1}{2}(1+\zeta) \end{aligned} \qquad x_{i-1} \le x \le x_i \tag{2.13}$$

with

$$\zeta = \frac{2x-(x_i+x_{i-1})}{x_i-x_{i-1}}$$

and the contribution to the weighting function is chosen according to:

$$\begin{aligned} w_{i-1}(x) &= \lambda_1(x) + \alpha F(x) \\ w_i(x) &= \lambda_2(x) + \alpha F(x) \end{aligned} \tag{2.14}$$

with α an arbitrary parameter and $F(\zeta)$ a quadratic function given by

$$F(\zeta) = -\frac{3}{4}(1-\zeta)(1+\zeta)$$

The basis functions, weights and modifying function F are shown in Fig. 2.5. The set of finite element equations now becomes

$$(2.15) \qquad \int_{\Omega} w_i[-\varepsilon \frac{d^2u^h}{dx^2} + v \frac{du^h}{dx}]dx = \int_{\Omega} fw_i dx \qquad i = 1,2,\ldots,n$$

with

$$u^h = \sum_{j=0}^{n+1} u_j \phi_j$$

In the homogeneous case ($f \equiv 0$) we obtain for equidistant mesh size the following set of linear equations.

$$(2.16) \qquad -[\varepsilon + \frac{vh}{2}(\alpha-1)]u_{i+1} + (2\varepsilon + vh\alpha)u_i - [\varepsilon + \frac{vh}{2}(\alpha+1)]u_{i-1} = 0 \quad i = 1,2,\ldots,n$$

So we see that an artificial viscosity of

$$(2.17) \qquad \frac{vh}{2}\alpha$$

has been introduced.

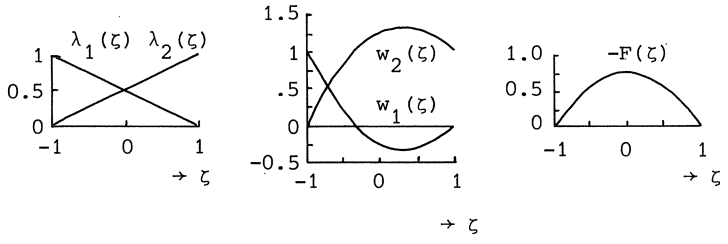


Fig. 2.5. One dimensional basis function, weights and modifying function for linear elements.

In fact it would be easier to introduce artificial viscosity straightforward than to use the Petrov Galerkin method to get an upwind scheme.

The introduction of an artificial viscosity can easily be performed on the level of finite element matrices. Consider for example the element matrix of the convective-diffusion equations (2.1)

$$(2.18) \qquad s^{e_i} = \frac{1}{h_i}\begin{bmatrix} \varepsilon - \frac{vh_i}{2} & -\varepsilon + \frac{vh_i}{2} \\ -\varepsilon - \frac{vh_i}{2} & \varepsilon + \frac{vh_i}{2} \end{bmatrix}$$

with $h_i = x_i - x_{i-1}$ the length of the i^{th} element.

The introduction of an artificial viscosity α replaces this matrix by

$$(2.19) \qquad s^{e_i} = \frac{1}{h_i}\begin{bmatrix} \varepsilon + \alpha - \frac{vh_i}{2} & -\varepsilon - \alpha + \frac{vh_i}{2} \\ -\varepsilon - \alpha - \frac{vh_i}{2} & \varepsilon + \alpha + \frac{vh_i}{2} \end{bmatrix}$$

In order that the matrix s^{e_i} be semi positive definite we must choose

$$\varepsilon + \alpha - \frac{vh_i}{2} \geq 0, \quad \text{so} \quad \alpha \geq \frac{vh_i}{2} - \varepsilon.$$

This is exactly the condition (2.11)

3. SMOOTH SOLUTIONS

The problems noticed in Chapter 2 have only been derived for the homogeneous equation (2.1). The question arises whether these problems arise for each differential equation of convection diffusion type or not.

Therefore we consider two special non-homogeneous cases.

$$(3.1) \qquad -\varepsilon \frac{d^2u}{dx^2} + v \frac{du}{dx} = v(1-2x) + 2\varepsilon$$

$$u(0) = 0 \qquad u(1) = 0$$

The exact solution of (3.1) is given by $u(x) = x(1-x)$.

The central difference scheme solves (3.1) exactly independent of the mesh size h, since the truncation error is identical zero. So whether the matrix is diagonally domainant or not, no oscillations occur. A less trivial

example is the following:

$$(3.2) \qquad -\varepsilon \frac{d^2u}{dx^2} + v \frac{du}{dx} = \varepsilon\pi^2 \sin(\pi x) + v\pi \cos(\pi x)$$

$$u(0) = 0 \qquad u(1) = 0$$

The exact solution of (3.2) is given by $u(x) = \sin(\pi x)$.

In Fig. 3.1 and 3.2 the solution of (3.2) has been plotted computed with central differences resp. the Il'in scheme. In all cases $h = 0.1$, $\varepsilon = 0.025$ and $v = 1$ has been chosen, so the central difference matrix is not diagonally dominant.

These figures show that the central difference scheme is very accurate; the Il'in method introduces some damping, in fact it tries to create a boundary layer. Moreover although the central difference matrix is not diagonally dominant, no oscillations occur. (See [11].)

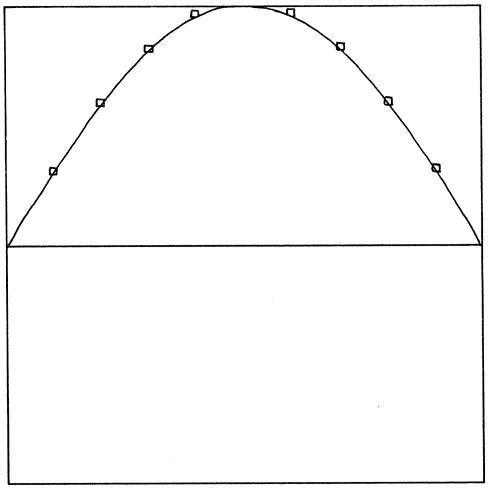


Fig. 3.1. Central differences

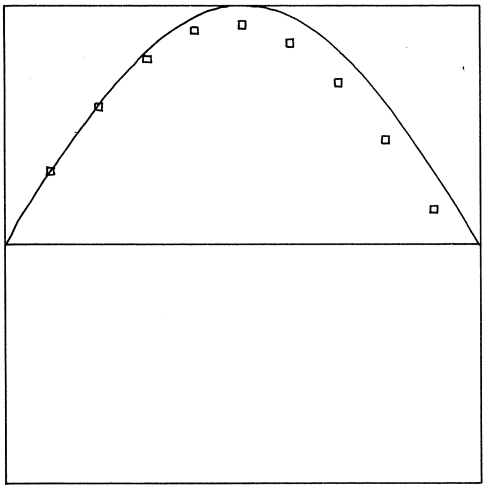


Fig. 3.2. The Il'in scheme

—— exact solution

□ numerical solution

$h = 0.01$; $\varepsilon = 0.025$; $v = 1$

The results of this chapter are in agreement with the following theorems proved in SEGAL [12].

THEOREM 3.1. *Let ϕ be the solution of the convection diffusion equation:*

$$-\varepsilon \frac{d^2\phi}{dx^2} + u \frac{d\phi}{dx} = f(x) \qquad x \in (0,1)$$

(3.3)

$$\phi(0) = \phi_0 \qquad \phi(1) = \phi_1$$

Let ϕ be smooth.

Then the C^∞ error e_c of the central difference scheme with constant mesh size satisfies:

$$e_c = O(h^2) + O(h^4/\varepsilon)$$

THEOREM 3.2. *Let ϕ be a smooth solution of the convection diffusion equation (3.3). Let ϕ_h be the solution of an upwind scheme with an artificial viscosity $-\tilde{\varepsilon}\, d^2\phi/dx^2$ and $\tilde{\varepsilon}$ such that the difference matrix is diagonally dominant. Let $\tilde{\varepsilon}$ be $O(h)$, then the C^∞ error $|\phi_h-\phi|$ satisfies:*

$$|\phi_h-\phi| = O(h)$$

So the error of the Il'in scheme is of order h.

THEOREM 3.3. *Let ϕ be a smooth solution of the convection diffusion equation (3.3). Let ϕ_h be the solution of a so-called modified upwind scheme, i.e. a scheme with an artificial viscosity $-\tilde{\varepsilon}\, d^2\phi/dx^2$ with*

$$\tilde{\varepsilon} = O(h^2)$$

Then the C^∞ error $|\phi_h-\phi|$ satisfies

$$|\phi_h-\phi| = O(h^2).$$

REMARK. The modified upwind scheme is not an upwind scheme in the classical sense since the difference matrix is not diagonally dominant when ε small with respect to h. Compare with condition (2.11).

These theorems show that for smooth solutions:

(i) If ε is not too small, i.e. $\varepsilon \geq O(h^2)$, then the central difference

scheme and the modified upwind scheme give the most accurate results.

(ii) If ε is small then the most accurate results can be achieved by the modified upwind scheme.

Moreover from the computations given in Table 3.1 we may conclude that oscillations are not caused by the non-diagonal dominancy of the difference matrix, but are due to the presence of boundary layers. So in the case of boundary layers we may expect better results when mesh refinement is applied.

4. NON-EQUIDISTANT MESH SIZES

Since it seems that the truncation error plays an important role in the question whether or not there are oscillations, the equation (2.1) has been solved with several graded meshes.

The non-equidistant central difference scheme can be written as:

$$(4.1)\qquad \frac{2}{h_i+h_{i-1}}\{-(\varepsilon/h_{i-1}+\frac{u}{2})u_{i-1} + (\varepsilon/h_{i-1}+\varepsilon/h_i)y_i - (\varepsilon/h_i-\frac{u}{2})y_{i+1}\} = 0$$

(4.1)

$$y_0 = 0 \qquad y_{n+1} = 1 \qquad\qquad i = 1,2,\ldots,n$$

with $h_j = x_{j+1} - x_j$.

The following three cases are computed:

(i) $\varepsilon = 0.025$, $u = 1$, $n = 9$

$x_1 = 0$, $x_2 = .5$, $x_3 = .8$, $x_4 = .85$, $x_5 = .88$, $x_6 = .91$, $x_7 = .93$, $x_8 = .95$, $x_9 = .97$, $x_{10} = .99$, $x_{11} = 1$.

The results are given in the following table:

TABLE 4.2. Results of the non-equidistant central difference scheme, for the solution of (2.1).

X-COOR	exact sol.	num. solut.	difference
+.5000	+.0000	-.0000	+.206'-08
+.8000	+.0003	-.0000	+.335'-03
+.8500	+.0025	+.0014	+.107'-02
+.8800	+.0082	+.0056	+.261'-02
+.9100	+.0273	+.0225	+.483'-02
+.9300	+.0608	+.0525	+.833'-02
+.9500	+.1353	+.1224	+.129'-01
+.9700	+.3012	+.2857	+.155'-01
+.9900	+.6703	+.6667	+.365'-02

$\varepsilon = 0.025$, $u = 1$, $n = 9$

Notice that the solution is negative in the points x_2 and x_3, however its value is very small ($<10^{-5}$). The error has the same sign over the whole region so no oscillations occur. The matrix is not diagonally dominant except in the boundary layer. The methods described in Chapter 2 give comparable results, however, without negative values. The overall errors are printed in Table (4.3).

TABLE 4.3. Overall errors for the non-equidistant mesh

	central diff.	backward dif.	Il'in
ℓ_2 error	$.520_{10}{}^{-3}$	$.397_{10}{}^{-1}$	$.446_{10}{}^{-3}$
ℓ_∞ error	$.155_{10}{}^{-1}$	$-.116_{10}{}^{0}$	$-103_{10}{}^{-1}$

$u = 1$, $\varepsilon = 0.025$, $n = 9$

This table shows that the Il'in scheme is slightly better than the central difference scheme.

(ii) $\varepsilon = 0.01$, $u = 1$, $n = 20$

The region is divided into 2 subregions:

$[0,0.96]$ and $[0.96,1]$

In each subregion an equidistant mesh is chosen, with

$h_1 = 0.096$ and $h_2 = 0.004$

Since the matrix corersponding to the first region is non-diagonally dominant and the mesh size is equidistant, in this region oscillations must occur for the central difference scheme (Theorem 2.4).

Results are printed in Table (4.4).

TABLE 4.4. Results of the non-equidistant central difference scheme, for the solution of (2.1)

X-COOR	exact sol.	num. solut.	difference
+.0960	+.0000	-.0006	+.639'-03
+.1920	+.0000	+.0003	-.336'-03
+.2880	+.0000	-.0012	+.115'-02
+.3840	+.0000	+.0011	-.112'-02
+.4800	+.0000	-.0023	+.235'-02
+.5760	+.0000	+.0029	-.294'-02
+.6720	+.0000	-.0051	+.513'-02
+.7680	+.0000	+.0072	-.719'-02
+.8640	+.0000	-.0116	+.116'-01
+.9600	+.0183	+.0171	+.122'-02
+.9640	+.0273	+.0258	+.156'-02
+.9680	+.0408	+.0388	+.199'-02
+.9720	+.0608	+.0583	+.252'-02
+.9760	+.0907	+.0876	+.316'-02
+.9800	+.1353	+.1315	+.387'-02
+.9840	+.2019	+.1973	+.457'-02
+.9880	+.3012	+.2961	+.508'-02
+.9920	+.4493	+.4443	+.502'-02
+.9960	+.6703	+.6666	+.374'-02

$\varepsilon = 0.01$, $u = 1$, $n = 19$

Indeed oscillations occur, however, the amplitude is small (<0.012).

In Table (4.5) the errors of the methods of Chapter 2 are compared.

TABLE 4.5. Overall errors for the non-equidistant mesh.

	centr. diff.	backw. diff.	Il'in
ℓ_2 error	$.355_{10}{}^{-3}$	$.394_{10}{}^{-1}$	$.158_{10}{}^{-1}$
ℓ_∞ error	$.116_{10}{}^{-1}$	$-.796_{10}{}^{-1}$	$-.472_{10}{}^{-1}$

$\varepsilon = 0.01$, $u = 1$, $n = 19$

We see that in this case the central difference scheme is much better than the Il'in scheme.

(iii) As (ii), but with the following subdivision:

$$[0,0.9] \quad \text{and} \quad [0.9,1]$$

Thus $h_1 = 0.09$ and $h_2 = 0.01$.

Results of the central difference scheme are given in Table (4.6)

TABLE 4.6. Results of the non-equidistant central difference scheme, for the solution of (2.1)

X-COOR	exact sol.	num. solut.	difference
+.0900	+.0000	-.0000	+.474'-06
+.1800	+.0000	+.0000	-.271'-06
+.2700	+.0000	-.0000	+.900'-06
+.3600	+.0000	+.0000	-.940'-06
+.4500	+.0000	-.0000	+.195'-05
+.5400	+.0000	+.0000	-.259'-05
+.6300	+.0000	-.0000	+.455'-05
+.7200	+.0000	+.0000	-.667'-05
+.8100	+.0000	-.0000	+.110'-04
+.9000	+.0000	+.0000	+.286'-04
+.9100	+.0001	+.0001	+.728'-04
+.9200	+.0003	+.0002	+.183'-03
+.9300	+.0009	+.0005	+.455'-03
+.9400	+.0025	+.0014	+.111'-02
+.9500	+.0067	+.0041	+.262'-02
+.9600	+.0183	+.0123	+.597'-02
+.9700	+.0498	+.0370	+.128'-01
+.9800	+.1353	+.1111	+.242'-01
+.9900	+.3679	+.3333	+.345'-01

$\varepsilon = 0.01$, u = 1, n = 19

The oscillations in this case have a much smaller amplitude, however the overall accuracy is also smaller because of the poor approximation of the solution in the boundary layer.

In Table (4.7) the errors of the methods of Chapter 2 are compared.

TABLE 4.7. Overall errors for the non-equidistant mesh

	centr. diff.	backw. diff.	Il'in
ℓ_2 error	$.199_{10}-2$	$.395_{10}-1$	$.223_{10}-6$
ℓ_∞ error	$.345_{10}-1$	$-.132_{10}0$	$-.158_{10}-3$

$\varepsilon = 0.01$, u = 1, n = 19

So when there are few points within the boundary layer Il'ins scheme turns out to be the best.

Concluding remarks

Since in practical problems either a right hand side may be present, or the boundary layer (in more dimensional cases) is not so simple as in

this one dimensional example, mesh refinement is preferable to upwind differencing.

However, when the perturbation parameter ε becomes too small the central difference scheme becomes inaccurate and some upwind differencing of $O(h^2)$, compare with Theorems (3.1) and (3.3), may be necessary.

When we want to solve the system of equations iteratively, it may be necessary to make the matrix diagonally dominant. See Chapter 5.

5. THE CONDITION OF THE DISCRETIZED EQUATIONS AND ITS CONSEQUENCES FOR ITERATIVE SOLUTION METHODS

One can show (SEGAL [12]) that when a central difference scheme or a finite element method is used, the condition of the difference matrix is of order $(1/\varepsilon)$. Hence for small values of ε this may cause problems.

In order to investigate this matter we have computed numerical solutions for the following two-dimensional convection diffusion equation.

$$
\begin{aligned}
&-\varepsilon\Delta\phi + u\frac{\partial\phi}{\partial x} = f \qquad (x,y) \in \Omega \\
&\text{with} \\
&\phi = 0 \qquad x \in \partial\Omega
\end{aligned}
\tag{5.1}
$$

$$\Omega = (0,1) \times (0,1)$$

$$f(x,y) = 2\pi^2\varepsilon\sin(\pi x)\sin(\pi y) + u\pi\cos(\pi x)\sin(\pi y)$$

The exact solution of (5.1) is given by

$$\phi(x,y) = \sin(\pi x)\sin(\pi y)$$

Equation (5.1) has been solved by a central difference scheme with constant mesh size in both directions. ($h = \Delta x = \Delta y$). One can prove [12], that the error of the central difference scheme is of order h^2 and for ε small enough, independent of ε.

In Table 5.1 results of the direct method (Gaussian elimination without pivoting) for various values of ε and h has been given. The computations have been performed on an IBM 370/158 computer with an accuracy of 16 decimal places.

TABLE 5.1. ℓ_∞ error of the solution of (5.1) with a central difference scheme (u = 1).

h= Δx= Δy	$\varepsilon = 10^{-7}$	$\varepsilon = 10^{-8}$	$\varepsilon = 10^{-9}$	$\varepsilon = 10^{-10}$	$\varepsilon = 10^{-11}$	$\varepsilon = 10^{-12}$
h = .1	$.84_{10}{}^{-3}$	$.84_{10}{}^{-3}$	$.80_{10}{}^{-3}$	$.24_{10}{}^{-2}$	.44	$.18_{10}{}^{1}$
h = .05	$.71_{10}{}^{-4}$	$.71_{10}{}^{-4}$	$.63_{10}{}^{-4}$	$.89_{10}{}^{-4}$	$.12_{10}{}^{-1}$	$.79_{10}{}^{-1}$

The table shows that for $\varepsilon \geq 10^{-8}$ the condition of the matrix does not influence the results, whereas for $\varepsilon \leq 10^{-9}$ it does. Moreover when h decreases the condition of the matrix increases since then the matrix becomes "more diagonally domainant". In fact the value of $\varepsilon = 10^{-8}$ is so small that for practical problems the condition of the matrix does not disturb the results.

In Table 5.2 results of a special iterative method (a version of a non-symmetric conjugate grafient method (IDR) developed by SONNEVELD [26]) and in Table 5.3 results for a preconditioned variant of this method (PIDR) have been given for various values of ε and h.

TABLE 5.2. Number of iterations to solve equation (2.1) by 2 iterative methods.

	h	ε = 1	ε = 0.1	ε = 0.01	ε = 0.001
IDR	0.1	14	24	56	-
	0.05	32	47	78	-
PIDR	0.1	8	8	7	-
	0.05	15	16	10	86

u = 1, central difference scheme (h = Δx = Δy), required accuracy IDR : 10^{-3}
PIDR: 10^{-4}

The number of unknown is equal to 81 for h = 0.05. So from a practical point of view a number of iterations of 56 for h = 0.1 means no convergence.

The condition for diagonal dominancy is $\Delta x \leq \frac{2\varepsilon}{u}$, so Table 2.2 indicates that the IDR method converges as long as the matrix is diagonally dominant, whereas for the PIDR method this condition may be violated a little.

In Table 5.3 the number of iterations of the PIDR method as been given for various values of h and ε.

Table 5.3 shows that the PIDR method converges rapidly as long as $\varepsilon \geq 0.04\,\Delta x$, for smaller values of ε no convergence can be guaranteed.

TABLE 5.3. Number of iterations to solve equation (5.1) by the PIDR method and central differences.

	$\varepsilon = 0.01h$	$\varepsilon = 0.08h$	$\varepsilon = 0.06h$	$\varepsilon = 0.04h$	$\varepsilon = 0.02h$	$\varepsilon = 0.01h$
$h = 0.1$	8	8	11	15	31	-
$h = 0.05$	10	11	14	26	133	-

Required accuracy: 10^{-4}

$h = \Delta x = \Delta y$.

The question arises whether there are iterative methods to overcome this difficulty, that is whether there are iterative methods that allow for small values of ε. In Chapter 6 it is shown that the defect correction method combined with for example PIDR is such a method.

6. THE DEFECT CORRECTION METHOD

The defect correction method has been the subject of many investigations [27,28,29,30]. The method may be used to improve the accuracy of a difference scheme by a higher order scheme, or to estimate the error of such a scheme. Moreover it may be used to approximate the solution of a higher order scheme even when the higher order scheme is unstable [30].

In this chapter we study the effect of the defect correction process applied to the upwind schemes using a central difference scheme as correction.

We must solve equation (3.1) which will formally be denoted by:

$$L\phi = f \tag{6.1}$$

This problem can be solved by an upwind scheme denoted by:

$$L_h\phi_h = f_h \tag{6.2}$$

or alternatively by a central difference scheme:

$$L_h'\phi_h' = f_h'. \tag{6.3}$$

Now we define the following defect correction process:

(i) Solve $L_h\phi_h = f_h$ (upwind scheme)

(ii) Correct ϕ_n with the aid of the central difference scheme $L_h'\phi_h' = f_h'$ in order to get a new approximation $\bar{\phi}_h$:

$$\bar{\phi}_h = \phi_h - L_h^{-1}[L_h'\phi_h - f_h'] \tag{6.4}$$

(6.4) may be improved by repeated iterations.

In [12] the following theorem has been proved:

THEOREM 6.1. *Let* $L\phi = f$ *be the one dimensional convection-diffusion equation* (3.1) *with viscosity parameter* ε. *Let* $L_h'\phi_h' = f_h'$ *be the central difference discretization, and* $L_h\phi_h = f_h$ *be the upwind difference discretization. Let the upwind discretization be such that it is identical to the central difference scheme applied to* (3.1) *with a viscosity parameter* $\tilde{\varepsilon}$. *Then the iterated defect correction method* (6.4) *always converges with a rate of convergence of approximately:*

$$\frac{|\tilde{\varepsilon}-\varepsilon|}{\tilde{\varepsilon}}$$

When $\varepsilon << \tilde{\varepsilon}$ *this means: the rate of convergence is approximately* 1.

Hence the defect correction method must *not* be used as an iteration process.

Another theorem proved in [12] states the following:

THEOREM 6.2. *Let* $L\phi = f$ *be the one-dimensional convection diffusion equation* (3.1) *with viscosity parameter* ε. *Let* $L_h'\phi_h' = f_h'$ *be the central difference discretization, and let* $L_h'\phi_h' = f_h'$ *be an upwind discretization, both with constant mesh size* h. *Let the upwind discretization be such that it is identical to the central difference scheme applied to* (3.1) *with a viscosity parameter* $\tilde{\varepsilon}$. *Let* ϕ *be smooth.*

Then the defect correction method (6.4) (without iteration) has an error $|\phi_h-\phi|$ that is of $O(h^2)$ in almost all points except for the last few points where its error is of the same order as the upwind scheme applied.

Hence when $\tilde{\varepsilon}$ is of order h, then the error in the last few points is of order h, is $\tilde{\varepsilon}$ of order h^2 then the overall error is of order h^2.

So in fact the defect correction method does not improve the order of

accuracy in comparison with the upwind scheme. However, numerical results given in [12] show that the actual error is smaller than that of the upwind schemes.

Since in practical problems mesh refinement in the boundary layers will be applied one can expect better results in that case.

Numerical computations given in [12] show that for example in a $O(\sqrt{\varepsilon})$ boundary layer, which is of practical interest, the accuracy of the defect correction process when a mesh refinement is applied, is significantly better than that of the upwind scheme.

7. CONCLUSIONS

The aim of the research done in this paper has been the investigation of the solution of the Navier Stokes equations for large Reynolds numbers.

In general the solution of the Navier Stokes equations consists of 2 parts. In the neighbourhood of the boundary steep gradients may occur (boundary layers) and outside this region the solution in general is smooth.

In this report the aspect of the numerical problems due to the presence of boundary layers has been investigated. In order to do so the convection diffusion equation has been used as a model problem.

It has been shown that the solution in the boundary layer best can be approximated by central differences or alternatively finite elements, with small step sizes. The mesh sizes must be chosen such that the matrix with respect to the points in the boundary layer is locally diagonally dominant.

Outside the boundary layer the solution is smooth and the accuracy of the methods used depends on the diffusion parameter ε. When $h \leq O(\sqrt{\varepsilon})$ the central difference scheme and hence the finite element method is of order h^2, and is therefore the most accurate; when ε becomes smaller an upwind scheme will become more accurate. It has been shown that for small values of ε an upwind scheme with diffusion coefficient $\tilde{\varepsilon}$ of order h^2 gives the most accurate results.

In some special problems as for example the convection-diffusion equation with homogeneous right hand side and constant mesh size a special adapted scheme like the Il'in scheme gives the best answers. However in more general problems, like the Navier Stokes equations, we cannot expect to derive such schemes. In fact the error estimates for the smooth solutions show that the Il'in scheme in this case does not give the best answers.

Furthermore it has been shown that the defect correction process yields more accurate results than an upwind scheme and - for small values of ε - a central difference scheme. The defect correction process applied with a diagonal dominant matrix yields an accuracy of order h^2 in the internal points and an accuracy of order h in the points in the neighbourhood of the boundary. In practical applications where we must have a refinement of the mesh near the boundary we might expect an accuracy of order h^2 over the whole region. Otherwise it is always possible to get an accuracy of order h^2 by choosing $\tilde{\varepsilon}$ of order h^2.

One of the advantages of the defect correction method is that we can use iterative methods as for example the IDR method as long as the matrix corresponding to the upwind scheme is diagonally dominant.

Better results may be expected when the preconditioned IDR method is used, since this method allows for somewhat smaller values of $\tilde{\varepsilon}$. So our conclusion is that this last method together with mesh refining in the boundary layer is the most promising attack of the Navier Stokes equations for large Reynolds numbers.

REFERENCES

[1] HEMKER, P.W., *A numerical study of stiff two-point boundary problems,* Thesis, Mathematisch Centrum, Amsterdam (1977).

[2] PEARSON, C.E., *On a differential equation of boundary layer type,* J. Math. Phys. 47 (1968) pp. 134-154.

[3] PEARSON, C.E., *On nonlinear ordinary differential equations of boundary layer type,* J. Math. Phys. 47 (1968) pp. 351-358.

[4] IL'IN, A.M., *Differencing scheme for a differential equation with a small parameter affecting the highest derivative,* Math. Notes Acad. Sc. USSR, 6 (1969) pp. 596-602.

[5] HEINRICH, J.C., P.S. HUYAKHORN, O.C. ZIENKIEWICZ & A.R. MITCHELL, *An upwind finite element scheme for two-dimensional convective transport equations,* Int. J. Num. Meth. Engng, 11 (1977) pp. 131-143.

[6] GRIFFITHS, D.F. & J. LORENZ, *An analysis of the Petrov-Galerkin finite element method applied to a model problem,* Dept. of Math. and Stat., University of Calgary, Alberta, Canada, research paper no. 334 (February, 1977).

[7] HEINRICH, J.C. & O.C. ZIENKIEWICZ, *Quadratic finite element schemes for two-dimensional convective transport problems*, Int. J. Num. Meth. Engng, 11 (1977) pp. 1831-1844.

[8] CHRISTIE, I., D.F. GRIFFITHS, A.R. MITCHELL & O.C. ZIENKIEWICZ, *Finite element methods for second order equations, with significant first derivatives*, Int. J. Num. Meth. Engng, 10 (1976) pp. 1389-1396.

[9] AXELSSON, O. & I. GUSTAFSSON, *A modified upwind scheme for convective transport equations, and the use of a conjugate gradient, method for the solution of nonsymmetric systems of equations*, Journal of the Institute of Mathematics and its Applications, 23 (1979) pp. 321-337.

[10] CHIEN, J.C., *A general finite difference formulation with application to Navier-Stokes equations*, Computers and Fluids, 5 (1977) pp. 15-31.

[11] SEGAL, A., *On the need for upwind differencing for elliptic singular perturbation problems*, Part 1, Report NA-27, Technical University Delft, 1980.

[12] SEGAL, A., *On the need for upwind differencing for elliptic singular perturbation problems*, Part 2, Report NA-35, Technical University Delft, 1980.

[13] ROACHE, P.J., *Computational fluid dynamics*, Hermosa Publishers, Albuquerque, New Mexico (1972).

[14] SAMARSKII, A.A., *Monotonic difference schemes for elliptic and parabolic equations in the case of a non-selfadjoint elliptic operator*, Z. Vycisl. Mat. i Mat. Fiz., 5 (1965) pp. 548-581.

[15] SMITH, I.A., *Integration in time of diffusion and diffusion-convection equations*, In [16], pp. 1.3-1.20.

[16] GRAY, W.G. & G.F. PINDER (eds.), *Finite elements in water resources*, First conference, Princeton University U.S.A., July 1976, Pentech Press, London.

[17] MERCER, J.W. & C.R. FAUST, *The application of finite element techniques to immiscible flow in porous media*, In [16], pp. 1.21-1.58.

[18] MELLI, P., *An application of the Galerkin method to the Eulerian-Lagrangian treatment of time dependent advection and diffusion of air pollutants*, In [16], pp. 1.59-1.70.

[19] GENUCHTEN, M.TH. VAN, *On the accuracy and efficiency of several numerical schemes for solving the convective-dispersion equation,* In [16], pp. 1.71-1.90.

[20] EHLIG, C., *Comparison of numerical methods for solution of the diffusion-convection equation in one and two dimensions,* In [16], pp. 1.91-1.102.

[21] LAM, D.C.L., *Comparison of finite element and finite difference methods for nearshore advection-diffusion transport models,* In [16], pp. 1.115-1.130.

[22] MITCHELL, A.R. & R. WAIT, *The finite element method in partial differential equations,* John Wiley, Chichester (1977).

[23] ZIENKIEWICZ, O.C. & J.C. HEINRICH, *The finite element method and convection problems in fluid mechanics,* In [24], pp. 1-22.

[24] GALLAGHER, R.H., et al. (eds.), *Finite elements in fluids,* Vol. 3, Wiley, London (1968).

[25] GARTLING, D.K., *Some comments on the paper by Heinrich, Huyakorn, Zienkiewicz and Mitchell,* Int. J. for Num. Methods in Engng, 12 (1978), 1, pp. 187-190.

[26] WESSELING, P. & P. SONNEVELD, *Numerical experiments with a multiple grid and a preconditioned Lanczos type method,* Lecture Notes in Math. 771, *Approximation methods for Navier Stokes equations,* (ed.) R. Rautmann, Berlin 1980, p. 543-562.

[27] FRANK, R., *The method of iterated defect-correction and its application to two-point boundary value problems,* Part I: Numer. Math. 25, (1976) pp. 409-419; Part II: Numer. Math. 27, (1977) pp. 407-420.

[28] FRANK, R. & C.W. UEBERHUBER, *Iterated defect correction for differential equations,* Part I: Theoretical results, Computing 20 (1978) pp. 207-228.

[29] STETTER, H.J., *The defect correction principle and discretization methods,* Numer. Math. 29 (1978) pp. 425-443.

[30] HACKBUSCH, W., *Bemerkungen zur iterienten defektkorrectur und zu ihrer kombination mit mehrgitterverfahren,* Report 79-13, Universität zu Köln, Mathematisches Institut, September 1979.

ON ITERATED DEFECT CORRECTION AND THE LOD-METHOD FOR PARABOLIC EQUATIONS

J.G. VERWER

1. INTRODUCTION

This contribution is concerned with the numerical solution of the initial boundary value problem for multispace dimensional parabolic partial differential equations. Our purpose is to discuss several theoretical and computational aspects of the locally one-dimensional (LOD) splitting method [13]. This method is usually applied as a time-integration method. By making use of the notion of iterated defect correction [5,10,12], it will be shown that the LOD-method can also be used as an iterative method for the solution of those systems of linear algebraic equations which arise in the application of fully implicit time-integration formulas. In this note we shall report some first investigations aimed at the development of an efficient iterative LOD-technique.

We shall follow the method of lines approach, that is, we shall consider partial differential equations of which the space differential operators have already been discretized (large systems of particular ordinary differential equations). Let the partial differential equation be given in the non-linear form

$$(1.1) \qquad u_t = \sum_{k=1}^{d} F_k(t,x_1,\ldots,x_d,u,u_{x_k},u_{x_kx_k}).$$

Let (1.1) be defined in the set $(0,T] \times \Omega$, where Ω denotes a bounded and path-connected region in the d-dimensional $(x_1,\ldots,x_d)$ - space with boundary $\partial\Omega$. For simplicity we assume that $\partial\Omega$ is always parallel to the coordinate axes and that, for $(t,x_1,\ldots,x_d) \in (0,T] \times \partial\Omega$, a boundary condition of the form

$$(1.2)\qquad a(t,x_1,\ldots,x_d)u + b(t,x_1,\ldots,x_d)u_n = c(t,x_1,\ldots,x_d)$$

is prescribed. Here u_n denotes the normal derivative. Further we assume that at $t = 0$ an initial function on Ω is given. It shall then be clear that by replacing u_{x_k}, $u_{x_kx_k}$ and u_n (on a suitable set of grid points) by finite difference expressions, the initial-boundary value problem is converted into an initial value problem for a system of ordinary differential equations, i.e.,

$$(1.3)\qquad \begin{aligned} &\vec{y}' = \vec{f}(t,\vec{y}),\ t \in (0,T],\ \vec{y}(0) = \vec{y}_0,\\ &\vec{f}(t,\vec{y}) = \sum_{k=1}^{d} \vec{f}_k(t,\vec{y}). \end{aligned}$$

The so-called k-th splitting function $\vec{f}_k$ is assumed to approximate, on the grid covering Ω, the one-dimensional differential operator $F_k[6]$. Further, each component of $\vec{y}$ represents an approximation to the exact solution u at a certain grid point $(x_1,\ldots,x_d) \in \Omega$ for all $t \in [0,T]$.

In the following it is assumed that all Jacobians $J_k(t,\vec{y}) = \partial\vec{f}_k(t,\vec{y})/\partial\vec{y}$ can be reordered to tridiagonal matrices (3-point finite differences). Further it is assumed that all their eigenvalues are situated in a narrow strip along the negative axis of the complex plane (parabolic equation). Finally we suppose that the exact solution $\vec{y}$ of (1.3) is sufficiently smooth.

2. THE LOD-INTEGRATION METHOD

The LOD-integration method for (1.3) is defined by [6]

$$(2.1)\qquad \begin{aligned} &\vec{y}_{(0)} = \vec{y}_n,\\ &\vec{y}_{(k)} = \vec{y}_{(k-1)} + \tau\vec{f}_k(t_{n+1},\vec{y}_{(k)}),\ k = 1(1)d,\\ &\vec{y}_{n+1} = \vec{y}_{(d)}. \end{aligned}$$

In this one-step integration formula $\tau = t_{n+1} - t_n$ denotes the time step and $\vec{y}_n$ denotes the LOD-approximation to the exact solution $\vec{y}(t)$ of (1.3) at time $t = t_n$. It is easy to see that the order of consistency of (2.1) is equal to 1. One integration step with (2.1) involves the solution of d systems of non-linear algebraic or transcendental equations. Because the Jacobian matrices $J_k(t,\vec{y})$ are tridiagonal, this can be achieved very

effectively by means of some Newton-type iteration. Moreover, the order of (2.1) is only equal to 1 and thus it is not necessary to solve these systems very accurately (one iteration step).

Method (2.1) belongs to a wide class of related splitting integration methods [6]. The basic idea of splitting, for a d dimensional problem, is to solve a sequence of d simple (one-dimensional) problems in such a way that the resulting scheme possesses unconditional stability properties. In fact, a nice property of (2.1) is that high frequency components are strongly damped [13]. A drawback of (2.1) is its low order of accuracy.

In [12] it has been shown that by means of the notion of defect correction [3,5,10] the LOD-integration method can be used to approximate numerical solutions defined by certain collocation schemes. The most simple scheme among these is backward Euler. For this scheme we shall shortly describe the defect correction process given in [12]. In section 3 we then shall concentrate on the iterative LOD-method for particular systems of linear algebraic equations.

Let G_L be the non-linear operator associated to (2.1), i.e.

$$\vec{y}_{n+1} = G_L \vec{y}_n. \tag{2.1'}$$

Let the non-linear operator equation

$$F\vec{v} = \vec{y}_n \tag{2.2}$$

define the backward Euler scheme, i.e.

$$F\vec{v} = \vec{v} - \tau \vec{f}(t_{n+1}, \vec{v}). \tag{2.2'}$$

The iterative process given in [12] (section 2, the case m=1) can then shortly be formulated as the defect correction process (DCPA-process, see [10])

$$\begin{aligned} &\vec{y}^{(0)}_{n+1} = G_L \vec{y}_n, \\ &\vec{y}^{(j+1)}_{n+1} = (I - G_L F)\vec{y}^{(j)}_{n+1} + G_L \vec{y}_n, \qquad j = 0,1,\ldots . \end{aligned} \tag{2.3}$$

Because each application of F and G_L involves an evaluation of the non-linear vector function $\vec{f}(t,\vec{y})$, each iteration with (2.3) costs 2 $\vec{f}$-evalua-

tions and at least d tridiagonal forward-backward substitutions. Therefore, it will be certainly more effective to solve (2.2) with a Newton-type method and to apply, per Newton iteration, a similar process as (2.3) to the resulting system of linear algebraic equations. An additional advantage of this approach is that without any modification we are allowed to consider other types of time-integration formulas, e.g. the higher order backward differentiation formulas, Rosenbrock formulas, and semi-explicit Runge-Kutta formulas [8].

3. THE ITERATIVE LOD-METHOD

Consider the linear system

$$(3.1) \qquad F\vec{v} = \vec{b},\ F = I-\hat{\tau}J,\ J = \sum_{k=1}^{d} J_k,$$

where J_k now denotes the Jacobian matrix of $\vec{f}_k(t,\vec{y})$ evaluated at some point. The scalar $\hat{\tau} > 0$ stands for the product of a time step τ and some constant determined by the integration formula under consideration. The vector $\vec{b}$ is also determined by the integration formula and changes per Newton-iteration.

Define the approximate inverse

$$(3.2) \qquad G_L = \prod_{k=1}^{d} (I-\hat{\tau}J_{d-k+1})^{-1}.$$

Recall that each matrix $I - \hat{\tau}J_k$ is non-singular and tridiagonal. The iterative LOD-method for (3.1) is then defined by

$$(3.3) \qquad \vec{v}^{(j+1)} = (I-G_L F)\vec{v}^{(j)} + G_L\vec{b}, \quad j = 0,1,\ldots .$$

Each application of (3.3) requires one complete matrix-vector operation and d tridiagonal forward-backward substitutions.

REMARK 3.1. If in (2.2') the non-linear function $\vec{f}(t,\vec{v})$ is replaced by the linear expression $J\vec{v}$, the DCPA-processes (2.3) and (3.3) become identical by setting $\vec{b} = \vec{y}_n$ and $\hat{\tau} = \tau$. □

REMARK 3.2. A similar DCPA-formulation as (3.3), where k = 2, has been suggested in [7], remark 3.2. The formulation there appears in a somewhat different setting and is not connected with the LOD-method. □

4. CONVERGENCE PROPERTIES OF THE ITERATIVE LOD-METHOD

Let $\vec{\varepsilon}^{(j)} = \vec{v} - \vec{v}^{(j)}$ be the j-th iteration error for (3.3), thus

$$\vec{\varepsilon}^{(j+1)} = (I-G_L F)\vec{\varepsilon}^{(j)}. \tag{4.1}$$

Using Fourier analysis, we shall investigate the convergence properties of (3.3) for the parabolic equation

$$u_t = \sum_{k=1}^{d} a_k u_{x_k x_k}, \quad a_k > 0, \ a_k \text{ constant}, \tag{4.2}$$

defined on Ω: $-\pi < x_k < \pi$, $k = 1,\ldots,d$, and having 2π-periodic boundary conditions in each direction. Herewith it is supposed that, for each k, the operator $\partial^2/\partial x_k^2$ has been discretized on a uniform grid by the second order finite difference operator $h_k^{-2}\delta_k^2$. Now let m be a d-dimensional vector whose components $m_1,\ldots,m_d$ are grid indices. For each grid point equation (4.1) can then be formally rewritten as

$$\varepsilon_m^{(j+1)} = [E - \prod_{k=1}^{d} (E-\hat{\tau}a_k h_k^{-2}\delta_k^2)^{-1}(E-\hat{\tau}\sum_{k=1}^{d} a_k h_k^{-2}\delta_k^2)]\varepsilon_m^{(j)}, \tag{4.3}$$

where E denotes the identity operator. Next, substitution of

$$e^{i[\omega_1 x_{m_1}+\ldots+\omega_d x_{m_d}]}, \qquad \omega_k \in \mathbb{Z}\setminus\{0\}, \tag{4.4}$$

for $\varepsilon_m^{(j)}$ into (4.3), gives the amplification factor

$$\psi_L = 1 - \prod_{k=1}^{d} (1-z_k)^{-1}(1-\sum_{k=1}^{d} z_k), \tag{4.5}$$

where

$$z_k = 2\hat{\tau}a_k h_k^{-2}(\cos\theta_k - 1), \quad -\pi < \theta_k = \omega_k h_k < \pi. \tag{4.6}$$

It is immediate that always

$$0 < \psi_L < 1. \tag{4.7}$$

More precisely, if all $z_k \to -\infty$ then $\psi_L \to 1$, and if all $z_k \to 0$ then $\psi_L \to 0$.

Consequently, low frequency components are better damped than high frequency components. This is remarkable because (all known) standard iterative methods show a reversed damping behaviour. In the next section we shall therefore combine the iterative LOD-method with such a standard method. Of course, the aim of this is to obtain amplification factors which all, in modulus, are significantly smaller than 1.

REMARK 4.1. Usually the lower harmonics are dominant in the initial error vector $\vec{\varepsilon}^{(0)}$ occurring in (4.1). This means that in the initial phase of the process the convergence is determined by the small eigenvalues of $I-G_L F$. However, the speed of convergence decreases with the number of iterations because the error vectors $\vec{\varepsilon}^{(j)}$ tend to lie in subspaces spanned by the dominant eigenvectors. For these eigenvectors the amplification factors are close to 1. For an illustrative experiment, see [12] table 3.1. □

5. AN ITERATIVE PROCESS BASED ON THE LOD-METHOD AND THE LINE JACOBI METHOD

For every iterative process it is desirable that over the whole frequency range the amplification factors are significantly smaller than 1. Let us therefore try to combine our LOD-method (3.2)-(3.3) with one or more alternative iterative methods which possess good damping properties for the high frequency components. For that purpose we have several possibilities, e.g. the point Gaus-Seidel method, the line Gaus-Seidel and line Jacobi method [2], and also iterative methods which are based on an incomplete factorization of the matrix F[4,9]. Here we shall consider the line Jacobi method. For convenience of presentation we now confine ourselves to d = 2, i.e. two space dimensions. For reasons of symmetry, line Jacobi will be applied in both directions.

Let G_{L1} and G_{L2} be two approximate inverses of the matrix F in (3.1). Let, for i = 1 and i = 2,

$$\vec{v}^{(j+1)} = (I-G_{Li}F)\vec{v}^{(j)} + G_{Li}\vec{b} , \tag{5.1}$$

be the corresponding DCPA-process. Consider the non-stationary defect correction process

$$
\begin{aligned}
\vec{v}^{(j+1/3)} &= (I-G_{L1}F)\vec{v}^{(j)} + G_{L1}\vec{b}\ , \\
\vec{v}^{(j+2/3)} &= (I-G_{L2}F)\vec{v}^{(j+1/3)} + G_{L2}\vec{b}\ , \\
\vec{v}^{(j+1)} &= (I-G_{L}F)\vec{v}^{(j+2/3)} + G_{L}\vec{b}\ ,
\end{aligned}
\tag{5.2}
$$

which may be interpreted as a new DCPA-process, say

$$
\vec{v}^{(j+1)} = (I-GF)\vec{v}^{(j)} + G\vec{b}, \tag{5.3}
$$

where G denotes a new approximate inverse. Now let G_{L1} and G_{L2} denote the approximate inverses defined by the line Jacobi procedure along x_1-grid lines and x_2-grid lines, respectively. The corresponding amplification factors, say ψ_{L1} and ψ_{L2}, are then given by

$$
\psi_{L1} = \frac{2\hat{\tau}a_2h_2^{-2}\cos\theta_2}{1 - 2\hat{\tau}a_1h_1^{-2}(\cos\theta_1-1) + 2\hat{\tau}a_2h_2^{-2}}\ , \tag{5.4}
$$

$$
\psi_{L2} = \frac{2\hat{\tau}a_1h_1^{-2}\cos\theta_1}{1 - 2\hat{\tau}a_2h_2^{-2}(\cos\theta_2-1) + 2\hat{\tau}a_1h_1^{-2}}\ . \tag{5.5}
$$

REMARK 5.1. Line Jacobi along x_1-lines should not be applied if $a_2 >> a_1$ or $h_2 << h_1$, and vice versa. In the following we therefore assume that $a_1 \simeq a_2$ and $h_1 \simeq h_2$. If these quantities differ greatly in magnitude it is implicitly assumed that the wrong direction will not be used. □

The amplification factor for the iterative process (5.2) is given by

$$
\psi = \psi_L\psi_{L1}\psi_{L2}\ , \qquad 0 < |\psi| < 1. \tag{5.6}
$$

Introduce the abbreviations $\alpha_k = 2\hat{\tau}a_kh_k^{-2}$ and $\xi_k = \cos\theta_k$. Then

$$(5.6') \qquad \psi = \frac{\alpha_1^2\alpha_2^2\xi_1(\xi_1-1)\xi_2(\xi_2-1)}{(1-\alpha_1\xi_1+\alpha_1)(1-\alpha_2\xi_2+\alpha_2)(1-\alpha_1\xi_1+\alpha_1+\alpha_2)(1-\alpha_2\xi_2+\alpha_1+\alpha_2)} \ .$$

Let us interpret ψ as a function of continuous (ξ_1,ξ_2)-values, where $-1 \le \xi_1,\xi_2 \le 1$. Further, for simplicity, let $\alpha_1 = \alpha_2 = \alpha$ be a given parameter. We are interested in the behaviour of ψ_{max} and ψ_{min} as $\alpha \to \infty$. Observe that ψ is symmetric. The derivatives $\partial\psi/\partial\xi_1$ and $\partial\psi/\partial\xi_2$ vanish if ξ_1 and ξ_2 satisfy

$$(5.7) \qquad (2\alpha^2+2\alpha)\xi^2-2(2\alpha^2+3\alpha+1)\xi+(2\alpha^2+3\alpha+1) = 0.$$

In the interior of the square $-1 \le \xi_1,\xi_2 \le 1$ the function ψ possesses one extreme value, viz. at the point

$$(5.8) \qquad (\xi_1,\xi_2) = \left(\frac{\beta-\sqrt{\beta}}{2\alpha}, \frac{\beta-\sqrt{\beta}}{2\alpha}\right), \qquad \beta = 2\alpha + 1.$$

A second extreme value occurs at the boundary at the point

$$(5.9) \qquad (\xi_1,\xi_2) = \left(-1, \frac{\beta-\sqrt{\beta}}{2\alpha}\right).$$

An inspection of fig. 5.1 reveals that ψ is maximal and positive at the point (5.8) and minimal and negative at the point (5.9). We have

$$\psi_{max} = \left(\frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}\right)^4 ,$$

$$(5.10)$$

$$\psi_{min} = -\frac{(1-\sqrt{\beta})^4}{\beta(3\beta-1)} .$$

Hence, $|\psi|_{max} = \max(\psi_{max}, -\psi_{min})$ is given by

$$(5.11) \qquad |\psi|_{max} = \max\left\{\left(\frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}\right)^4, \frac{(1-\sqrt{\beta})^4}{\beta(3\beta-1)}\right\} .$$

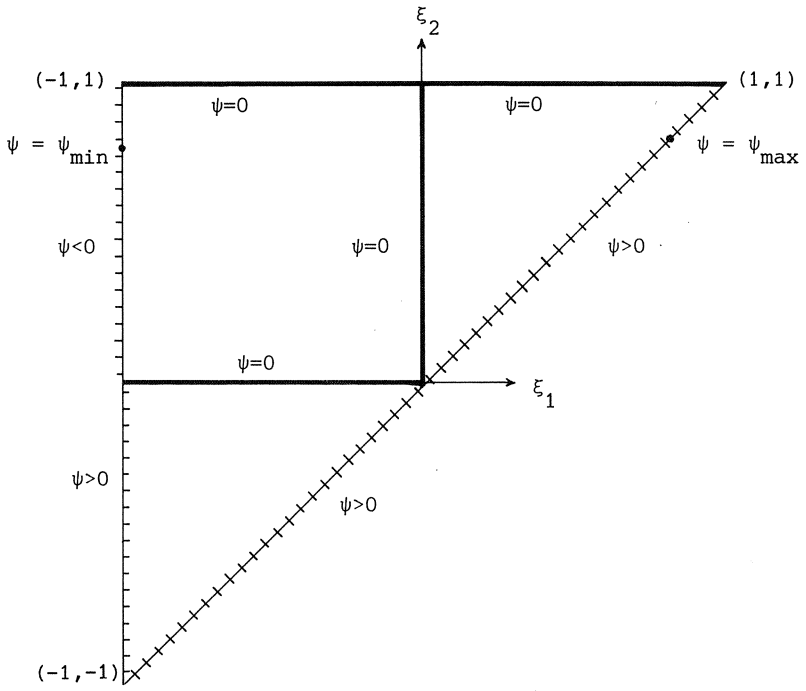


Fig. 5.1 The behaviour of the amplification factor (5.6').

For large values of α

$$|\psi|_{max} = \left(\frac{1-\sqrt{2\alpha+1}}{1+\sqrt{2\alpha+1}}\right)^4 \to 1 \text{ as } \alpha \to \infty. \tag{5.12}$$

Further, if $\alpha \to \infty$, the extremal point (5.8) approaches the point (1,1). We thus see that for the iterative method (5.2) "lower" frequency components determine the rate of convergence. Recall that in the derivation of (5.12) we assumed ξ_1 and ξ_2 to be continuous. In case of small grid distances h_1 and h_2, it is allowed to make such a simplifying assumption.

Let us now compare the maximal amplification factors of the LOD-method (3.3, d = 2), the combined method (5.2), and the symmetric line Jacobi method. Let us suppose that $a_1 = a_2 = 1$ and $h_1 = h_2 = h$. Consequently, $\alpha_1 = \alpha_2 = \alpha$ where

(5.13) $\alpha = 2\hat{\tau}h^{-2}.$

Using (4.5), (5.4), (5.5), (5.12) and (5.13), we find (α being large)

(5.14) $$|\psi_L|_{max} = \frac{4\hat{\tau}^2 h^{-4}(\cos(\pi-h)-1)^2}{(1+2\hat{\tau}h^{-2}-2\hat{\tau}h^{-2}\cos(\pi-h))^2},$$

(5.15) $$|\psi_{L1}\psi_{L2}|_{max} = \frac{4\hat{\tau}^2 h^{-4}\cos^2 h}{(1+4\hat{\tau}h^{-2}-2\hat{\tau}h^{-2}\cos h)^2},$$

(5.16) $$|\psi|_{max} = \left(\frac{1-\sqrt{1+4\hat{\tau}h^{-2}}}{1+\sqrt{1+4\hat{\tau}h^{-2}}}\right)^4.$$

Next it is supposed that

(5.17) $\alpha = 2\hat{\tau}h^{-2} \to \infty, \quad$ as $\hat{\tau}, h \to 0.$

This condition is natural when discussing implicit time-integration methods. We thus arrive at

(5.14') $$|\psi_L|_{max} \sim 1 - \frac{1}{\alpha} \sim 1 - \frac{1}{2}\hat{\tau}^{-1}h^2, \text{ as } \hat{\tau}, h \to 0, \alpha \to \infty,$$

(5.15') $$|\psi_{L1}\psi_{L2}|_{max} \sim 1 - \frac{2}{\alpha} \sim 1 - \hat{\tau}^{-1}h^2, \text{ as } \hat{\tau}, h \to 0, \alpha \to \infty,$$

(5.16') $$|\psi|_{max} \sim 1 - \frac{8}{\sqrt{2\alpha}} \sim 1 - \frac{4}{\sqrt{\hat{\tau}}}h, \text{ as } \hat{\tau}, h \to 0, \alpha \to \infty.$$

From these results we can conclude that, for the model problem, the asymptotic rate of convergence of our new method (5.2) is much better than the asymptotic convergence rates of both underlying iterative schemes. To be more precise, the numbers of decimal digits gained in each iteration step, are given by

(5.14") $$-^{10}\log|\psi_L|_{max} \sim \frac{\hat{\tau}^{-1}h^2}{2\ln 10},$$

(5.15") $$-^{10}\log|\psi_{L1L2}|_{max} \sim \frac{\hat{\tau}^{-1}h^2}{\ln 10},$$

(5.16") $$-^{10}\log|\psi|_{max} \sim \frac{8(2\alpha)^{-\frac{1}{2}}}{\ln 10} \sim \frac{4\hat{\tau}^{-\frac{1}{2}}h}{\ln 10}.$$

It should be noted that the operation count of method (5.2) is rather high. Each iterative step requires, approximately, 1 matrix-vector multiplication, 1 tridiagonal forward-backward substitution per grid line, and 1 line relaxation sweep in both directions. If we count 3 operations per grid point for a tridiagonal forward-backward substitution, we thus arrive at the operation counts:

method (5.2)	22
method (3.3), d = 2	12
symmetric line Jacobi	10

6. RICHARDSON ACCELERATION

From the asymptotic derivations (see e.g. [2], Section 21 for an explanation) it seems justified to conclude that by combining the iterative LOD-method (3.3) with the symmetric line Jacobi scheme the gain in convergence is worthwhile. If we set $\hat{\tau} = O(h)$, we even have (see (5.16'))

$$|\psi|_{max} \sim 1 - 4\sqrt{h}, \quad \text{as } h \to 0. \tag{6.1}$$

In theory, and hopefully also in practice, it is possible to obtain a much faster convergence by applying Richardson acceleration. Consider the matrix GF occurring in equation (5.3). For our model problem all eigenvalues of this matrix are positive and its spectral condition number P is given by (see (5.6'), (5.10))

$$P = \frac{(1-\psi)_{max}}{(1-\psi)_{min}} = \frac{1-\psi_{min}}{1-\psi_{max}} \sim \frac{1}{6}\sqrt{2\alpha}, \quad \text{as } \alpha \to \infty. \tag{6.2}$$

Richardson acceleration is obtained if in the iteration equation

$$\vec{v}^{(j+1)} = (I-\omega_j GF)\vec{v}^{(j)} + \omega_j G\vec{b}, \quad j = 0,1,\ldots \tag{6.3}$$

the acceleration parameters are chosen in such a way that the j-th error equation is given by

$$\vec{\varepsilon}^{(j)} = R_j(GF)\vec{\varepsilon}^{(0)}, \tag{6.4}$$

where $R_j(z)$ is the shifted Chebyshev polynomial

$$(6.5) \qquad R_j(z) = \frac{T_j(\frac{b+a-2z}{b-a})}{T_j(\frac{b+a}{b-a})} .$$

Here $a > 0$ denotes a lower estimate of the minimal eigenvalue of GF, and b denotes an upper estimate of the maximal eigenvalue. A close upperbound for the modulus of the maximal eigenvalue of $R_j(GF)$ is given by

$$(6.6) \qquad [T_j(\nu)]^{-1} = \frac{2}{(\nu+\sqrt{\nu^2-1})^j + (\nu-\sqrt{\nu^2-1})^j} , \quad \nu = \frac{b+a}{b-a} .$$

By comparing this maximal eigenvalue with $|\psi|_{max}$ given by (5.16), we get an indication on the gain in convergence. Note that for our model problem $\nu = (P+1)/(P-1)$.

For $\varepsilon = 10^{-4}$, 10^{-6} and $\alpha \in [10, 10^4]$ we computed the minimal integers j_1 and j_2 satisfying (see fig. 6.1)

$$(6.7) \qquad |\psi|_{max}^{j_1} \leq \varepsilon, \quad [T_{j_2}(\frac{P+1}{P-1})]^{-1} \leq \varepsilon.$$

These integers serve as a measure for the number of iterations necessary to decrease the relative error (in some norm) with a number equal to ε. Note that $\alpha = 2\hat{\tau}h^{-2} = 10^4$ is rather large. We see from fig. 6.1 that the gain in convergence by use of Richardson's method is considerable.

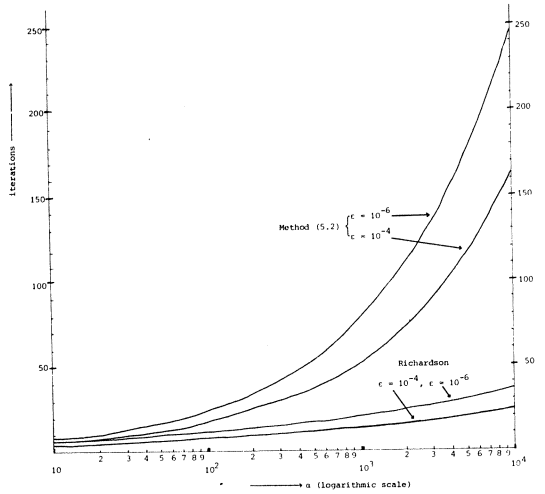


Fig. 6.1

In fact it is well known, see e.g. [1], that for the integers j satisfying

$$j \geq \frac{1}{2}\sqrt{\frac{b}{a}} \ln \frac{2}{\varepsilon}, \qquad \varepsilon > 0, \tag{6.8}$$

the inequality

$$[T_j(\nu)]^{-1} \leq \varepsilon \tag{6.7'}$$

holds. For the case we analyzed, condition (6.8) reads

$$j \geq 288^{-\frac{1}{4}} \alpha^{\frac{1}{4}} \ln \frac{2}{\varepsilon}, \tag{6.8'}$$

provided α is sufficiently large. Substituting $\alpha = 2\hat{\tau}h^{-2}$ yields

$$j \geq 12^{-\frac{1}{2}} h^{-\frac{1}{2}} \hat{\tau}^{\frac{1}{4}} \ln \frac{2}{\varepsilon}. \tag{6.8"}$$

Consequently, for fixed ε and $\hat{\tau}$, the number of iterations varies as $h^{-1/2}$ as $h \to 0$. For our model problem, with ε and $\hat{\tau}$ fixed, the number of iterations is therefore of the same order of magnitude as for the preconditioned conjugate gradient method given in [9] (see e.g. their results from fig. 5).

For the sake of completeness we still observe that in general equation (6.3) will cause the Richardson acceleration to be numerically unstable. This disadvantage may be eliminated by making use of two-step Chebyshev recursions. For clarifying discussions on analytical and practical aspects of Richardson's method, see e.g. [1,2].

7. SOME FIRST NUMERICAL RESULTS

In this section we present results of some first numerical experiments. The main purpose of the experiments was to test the methods in a non-model situation, that is, to find out whether in a non-model situation the theoretical gain in convergence of the new method comes through. By way of comparison we applied the LOD-method (3.3), the symmetric line Jacobi method, and the combination (5.2).

In the experiments we confined ourselves to solving just one linear system (no time-stepping) of type (3.1), i.e.

$$(7.1) \qquad (I-\hat{\tau}J)\vec{v} = 0,$$

J being the test matrix suggested by VARGA [11], appendix B, equation B4 (see also [9]). J represents a discrete approximation to a linear, 2-dimensional elliptic operator having piecewise constant coefficients. This operator certainly does not satisfy the restrictions made in section 4. Because it would go too far to give a precise definition of J, we only state that we used a uniform grid with grid distance h = 0.1. The exact solution of (7.1) is given by $\vec{v} = 0$. To prevent a fast convergence by coincident, in all experiments an initial starting vector was chosen with its components random between 0 and 2. Note that in an integration process of a parabolic equation we always have a good, usually smooth, initial starting vector to our disposal.

For $\hat{\tau} = 0.1$ the results have been plotted in fig. 7.1. We can conclude that for the present practical, non-standard problem the gain in convergence of the combination (5.2) certainly comes through. Both underlying methods fail to solve the problem.

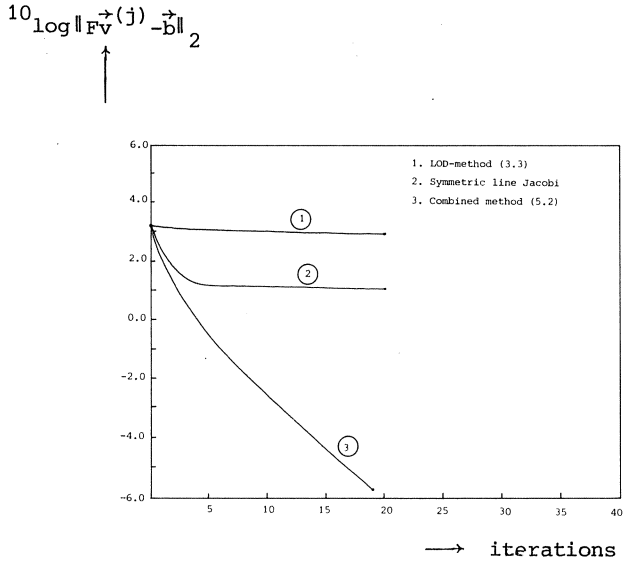


Figure 7.1 Results for equation (7.1). Remember that 1 iteration of the combined method is approximately twice as expensive as 1 iteration of the underlying methods.

8. CONCLUDING REMARKS

In this contribution we left several questions unanswered which could be subject to further investigation. First it seems worthwhile to investigate a combination of method (3.3) and an iterative incomplete factorization method, say

$$\vec{v}^{(j+1)} = (I-C^{-1}F)\vec{v}^{(j)} + C^{-1}\vec{b}, \tag{8.1}$$

C being the incomplete factorization of F. Of particular interest seem the modified incomplete factorizations proposed by GUSTAFSSON [4]. For certain classes of elliptic problems he has shown that these factorizations give already a condition number of $O(h^{-1})$, $h \to 0$. By combining (8.1) and (3.3) we expect an interesting improvement (cf. combination (5.2)). Such a combination should then be accelerated by use of Richardson's method.

Further it is of interest to investigate the use of such combinations also for elliptic problems. In fact, it can be easily shown that for 2 space dimensions method (3.3) is a special case of the iterative method of Peaceman and Rachford for elliptic matrix problems. Suppose we are given such a problem, say

$$F\vec{v} = \vec{b}, \quad F = F_1 + F_2. \tag{8.2}$$

By substitution of the relations $\hat{\tau}J_1 = \frac{1}{2}I - F_1$, $\hat{\tau}J_2 = \frac{1}{2}I - F_2$ into equation (3.2), where d = 2, there results the approximate inverse

$$G = (\tfrac{1}{2}I + F_2)^{-1}\,(\tfrac{1}{2}I + F_1)^{-1}. \tag{8.3}$$

Next, substitution of (8.3) into equation (3.3) yields the Peaceman-Rachford scheme (see e.g. [2])

$$\begin{aligned} v^{(j+1/2)} &= v^{(j)} - \mu[F_1\vec{v}^{(j+1/2)} + F_2\vec{v}^{(j)} - \vec{b}] \\ \vec{v}^{(j+1)} &= \vec{v}^{(j+1/2)} - \mu[F_1\vec{v}^{(j+1/2)} + F_2\vec{v}^{(j+1)} - \vec{b}], \end{aligned} \tag{8.4}$$

where $\mu = 2$. So, we could also have started the investigations with method (8.4) for the special matrices $F_1 = \frac{1}{2}I - \hat{\tau}J_1$, $F_2 = \frac{1}{2}I - \hat{\tau}J_2$, $F = F_1 + F_2 =$

$I - \hat{\tau}J$. It should be mentioned that in the literature on elliptic matrix problems methods of type (8.4) are usually not recommended because of their disappointing results for non-model problems. Our numerical experiment reported in section 7 confirms this (method (3.3) fails). The combination (5.2), however, performs satisfactorily on the example problem.

ACKNOWLEDGEMENT. The author wishes to acknowledge Mrs. M. Louter Nool for her programming assistance and for correcting some of the formulas.

REFERENCES

[1] AXELSSON, O., *Solution of linear systems of equations: Iterative methods,* Lecture Notes in Mathematics 572, pp. 1-51, Springer-Verlag Berlin-Heidelberg, 1977.

[2] FORSYTHE, G.E. & W.R. WASOW, *Finite difference methods for partial differential equations,* John Wiley & Sons, New York, 1960.

[3] FRANK, R. & C.W. UEBERHUBER, *Iterated defect correction for the efficient solution of stiff systems of ordinary differential equations,* BIT 17, 146-159, 1977.

[4] GUSTAFSSON, I., *On incomplete factorization,* (these lecture notes).

[5] HEMKER, P.W., *Introduction to multigrid methods,* (these lecture notes).

[6] HOUWEN, P.J. VAN DER & J.G. VERWER, *One-step splitting methods for semi-discrete parabolic equations,* Computing 22, 291-309, 1979.

[7] HOUWEN, P.J. VAN DER, *Multistep splitting methods for non-linear initial value problems,* (these lecture notes).

[8] LAMBERT, J.D., *Computational methods in ordinary differential equations,* John Wiley & Sons, London, 1973.

[9] MEIJERINK, J.A. & H.A. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric* M-*matrix,* Math. Comp. 31, 148-162, 1977.

[10] STETTER, H.J., *The defect correction principle and discretization methods,* Num. Math. 29, 425-443, 1978.

[11] VARGA, R.S., *Matrix iterative analysis,* Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1962.

[12] VERWER, J.G., *The application of iterated defect correction to the LOD method for parabolic equations*, BIT 19, 384-394, 1979.

[13] YANENKO, N.N., *The method of fractional steps*, Springer-Verlag, Berlin, 1971.

ON A CLASS OF EXPLICIT THREE-STEP RUNGE-KUTTA METHODS WITH EXTENDED REAL STABILITY INTERVALS

J.G. VERWER

REMARK. Only a summary of the colloquium lecture is given. The complete contribution has been reported in [7].

The lecture deals with the numerical solution of the initial value problem for large systems of ordinary differential equations written in the explicit, autonomous form

$$y' = f(y),$$

and possessing the property that the eigenvalues of the Jacobian matrix $J(y) = \partial f(y)/\partial y$ are situated in a long narrow strip along the negative axis of the complex plane. Such systems frequently arise when discretizing the space variables of initial-boundary value problems for parabolic partial differential equations [4]. We shall focus our attention to these semi-discrete parabolic problems. For our discussion it is not necessary to define a particular class of parabolic problems or to specify the method of semi-discretization. Our only restriction is the location of the eigenvalues of the Jacobian matrix $J(y)$. Further it is always assumed that the problem is sufficiently smooth.

In considering the application of explicit integration methods to semi-discrete parabolic equations one must weight up an important advantage against an important disadvantage. Their advantage, when compared with implicit or partly implicit methods (see e.g. [4,2]), is that they do not require the solution of large and complicated systems of nonlinear algebraic or transcendental equations (more dimensional problems) and, consequently, that they can be easily applied to large problem classes. Their disadvantage, as is well known, is the conditional stability. Fortunately it is possible to reduce this disadvantage considerably by using so-called stabilized Runge-Kutta methods (cf. [1], section 2.7). Such a method uses a relatively

large number of f(y)-equations per integration step, say m, the majority of which serves to enlarge the real stability boundary. As this boundary increases quadratically with m it certainly pays to employ such a stabilized method instead of a standard explicit one (see [4]). In fact, stabilized Runge-Kutta methods usually become more efficient as the degree m, that is the number of f(y)-equations per integration step, increases.

We shall present a new class of stabilized, 3-step Runge-Kutta methods containing the 3-step methods earlier reported in [5,6]. These last methods, however, have two disadvantages. The first is that the integration parameters are not known in closed form. The second, and the most severe one, is that they are internally unstable. That is, within one single integration step they exhibit a severe accumulation of rounding errors, even when satisfying the condition of absolute stability. Because of the fact that this accumulation can easily influence the local accuracy it is desirable to develop internally stable methods. Recently, VAN DER HOUWEN & SOMMEIJER [3] reported high degree, one-step Runge-Kutta methods which are indeed internally stable for all values of m. They obtained internal stability after identifying each intermediate stability function with a Chebyshev polynomial via a stable, two-step Chebyshev recursion. Following this idea we develop a class of internally stable, 3-step Runge-Kutta methods. Our absolute stability boundaries are approximately three times larger than the boundaries reported by VAN DER HOUWEN & SOMMEIJER [3]. We also discuss a linearization of the new formulas. By linearization we can replace m-1 f(y)-evaluations by m-1 multiplications of a Jacobian matrix with intermediate vectors. This means that for many problems the linearized schemes will be less expensive. The lecture is concluded with a numerical example.

REFERENCES

[1] HOUWEN, P.J. VAN DER, *Construction of integration formulas for initial value problems*, North-Holland Publishing Company, Amsterdam, 1977.

[2] HOUWEN, P.J. VAN DER & J.G. VERWER, *One-step splitting methods for semi-discrete parabolic equations*, Computing 22, 191-309, 1979.

[3] HOUWEN, P.J. VAN DER & B.P. SOMMEIJER, *On the internal stability of explicit, m-stage Runge-Kutta methods for large values of* m, ZAMM, to appear.

[4] RICHTMYER, R.D. & K.W. MORTON, *Difference methods for initial-value problems,* Interscience, New York, 1969.

[5] VERWER, J.G., *A class of stabilized three-step Runge-Kutta methods for the numerical integration of parabolic equations,* J. of Comp. and Appl. Math. 3, 155-166, 1977.

[6] VERWER, J.G., *An implementation of a class of stabilized, explicit methods for the time integration of parabolic equations,* TOMS 6, 188-205, 1980.

[7] VERWER, J.G., *On a class of explicit three-step Runge-Kutta methods with extended real stability intervals,* Report NW 77/79, Mathematical Centre, Amsterdam, prepublication, 1979.

UITGAVEN IN DE SERIE MC SYLLABUS

Onderstaande uitgaven zijn verkrijgbaar bij het Mathematisch Centrum, 2e Boerhaavestraat 49 te Amsterdam-1005, tel. 020-947272.

MCS 1.1 F. GÖBEL & J. VAN DE LUNE, *Leergang Besliskunde, deel 1: Wiskundige basiskennis*, 1965. ISBN 90 6196 014 2.

MCS 1.2 J. HEMELRIJK & J. KRIENS, *Leergang Besliskunde, deel 2: Kansberekening*, 1965. ISBN 90 6196 015 0.

MCS 1.3 J. HEMELRIJK & J. KRIENS, *Leergang Besliskunde, deel 3: Statistiek*, 1966. ISBN 90 6196 016 9.

MCS 1.4 G. DE LEVE & W. MOLENAAR, *Leergang Besliskunde, deel 4: Markovketens en wachttijden*, 1966. ISBN 90 6196 017 7.

MCS 1.5 J. KRIENS & G. DE LEVE, *Leergang Besliskunde, deel 5: Inleiding tot de mathematische besliskunde*, 1966. ISBN 90 6196 018 5.

MCS 1.6a B. DORHOUT & J. KRIENS, *Leergang Besliskunde, deel 6a: Wiskundige programmering 1*, 1968. ISBN 90 6196 032 0.

MCS 1.6b B. DORHOUT, J. KRIENS & J.TH. VAN LIESHOUT, *Leergang Besliskunde, deel 6b: Wiskundige programmering 2*, 1977. ISBN 90 6196 150 5.

MCS 1.7a G. DE LEVE, *Leergang Besliskunde, deel 7a: Dynamische programmering 1*, 1968. ISBN 90 6196 033 9.

MCS 1.7b G. DE LEVE & H.C. TIJMS, *Leergang Besliskunde, deel 7b: Dynamische programmering 2*, 1970. ISBN 90 6196 055 X.

MCS 1.7c G. DE LEVE & H.C. TIJMS, *Leergang Besliskunde, deel 7c: Dynamische programmering 3*, 1971. ISBN 90 6196 066 5.

MCS 1.8 J. KRIENS, F. GÖBEL & W. MOLENAAR, *Leergang Besliskunde, deel 8: Minimaxmethode, netwerkplanning, simulatie*, 1968. ISBN 90 6196 034 7.

MCS 2.1 G.J.R. FÖRCH, P.J. VAN DER HOUWEN & R.P. VAN DE RIET, *Colloquium Stabiliteit van differentieschema's, deel 1*, 1967. ISBN 90 6196 023 1.

MCS 2.2 L. DEKKER, T.J. DEKKER, P.J. VAN DER HOUWEN & M.N. SPIJKER, *Colloquium Stabiliteit van differentieschema's, deel 2*, 1968. ISBN 90 6196 035 5.

MCS 3.1 H.A. LAUWERIER, *Randwaardeproblemen, deel 1*, 1967. ISBN 90 6196 024 X.

MCS 3.2 H.A. LAUWERIER, *Randwaardeproblemen, deel 2*, 1968. ISBN 90 6196 036 3.

MCS 3.3 H.A. LAUWERIER, *Randwaardeproblemen, deel 3*, 1968. ISBN 90 6196 043 6.

MCS 4 H.A. LAUWERIER, *Representaties van groepen*, 1968. ISBN 90 6196 037 1.

MCS 5 J.H. VAN LINT, J.J. SEIDEL & P.C. BAAYEN, *Colloquium Discrete wiskunde*, 1968.
ISBN 90 6196 044 4.

MCS 6 K.K. KOKSMA, *Cursus ALGOL 60*, 1969. ISBN 90 6196 045 2.

MCS 7.1 *Colloquium Moderne rekenmachines, deel 1*, 1969. ISBN 90 6196 046 0.

MCS 7.2 *Colloquium Moderne rekenmachines, deel 2*, 1969. ISBN 90 6196 047 9.

MCS 8 H. BAVINCK & J. GRASMAN, *Relaxatietrillingen*, 1969.
ISBN 90 6196 056 8.

MCS 9.1 T.M.T. COOLEN, G.J.R. FÖRCH, E.M. DE JAGER & H.G.J. PIJLS, *Elliptische differentiaalvergelijkingen, deel 1*, 1970.
ISBN 90 6196 048 7.

MCS 9.2 W.P. VAN DEN BRINK, T.M.T. COOLEN, B. DIJKHUIS, P.P.N. DE GROEN, P.J. VAN DER HOUWEN, E.M. DE JAGER, N.M. TEMME & R.J. DE VOGELAERE, *Colloquium Elliptische differentiaalvergelijkingen, deel 2*, 1970.
ISBN 90 6196 049 5.

MCS 10 J. FABIUS & W.R. VAN ZWET, *Grondbegrippen van de waarschijnlijkheidsrekening*, 1970. ISBN 90 6196 057 6.

MCS 11 H. BART, M.A. KAASHOEK, H.G.J. PIJLS, W.J. DE SCHIPPER & J. DE VRIES, *Colloquium Halfalgebra's en positieve operatoren*, 1971.
ISBN 90 6196 067 3.

MCS 12 T.J. DEKKER, *Numerieke algebra*, 1971. ISBN 90 6196 068 1.

MCS 13 F.E.J. KRUSEMAN ARETZ, *Programmeren voor rekenautomaten; De MC ALGOL 60 vertaler voor de EL X8*, 1971. ISBN 90 6196 069 X.

MCS 14 H. BAVINCK, W. GAUTSCHI & G.M. WILLEMS, *Colloquium Approximatietheorie*, 1971. ISBN 90 6196 070 3.

MCS 15.1 T.J. DEKKER, P.W. HEMKER & P.J. VAN DER HOUWEN, *Colloquium Stijve differentiaalvergelijkingen, deel 1*, 1972. ISBN 90 6196 078 9.

MCS 15.2 P.A. BEENTJES, K. DEKKER, H.C. HEMKER, S.P.N. VAN KAMPEN & G.M. WILLEMS, *Colloquium Stijve differentiaalvergelijkingen, deel 2*, 1973. ISBN 90 6196 079 7.

MCS 15.3 P.A. BEENTJES, K. DEKKER, P.W. HEMKER & M. VAN VELDHUIZEN, *Colloquium Stijve differentiaalvergelijkingen, deel 3*, 1975.
ISBN 90 6196 118 1.

MCS 16.1 L. GEURTS, *Cursus Programmeren, deel 1: De elementen van het programmeren*, 1973. ISBN 90 6196 080 0.

MCS 16.2 L. GEURTS, *Cursus Programmeren, deel 2: De programmeertaal ALGOL 60*, 1973. ISBN 90 6196 087 8.

MCS 17.1 P.S. STOBBE, *Lineaire algebra, deel 1*, 1974. ISBN 90 6196 090 8.

MCS 17.2 P.S. STOBBE, *Lineaire algebra, deel 2*, 1974. ISBN 90 6196 091 6.

MCS 17.3 N.M. TEMME, *Lineaire algebra , deel 3*, 1976. ISBN 90 6196 123 8.

MCS 18 F. VAN DER BLIJ, H. FREUDENTHAL, J.J. DE IONGH, J.J. SEIDEL & A. VAN WIJNGAARDEN, *Een kwart eeuw wiskunde 1946-1971, Syllabus van de Vakantiecursus 1971*, 1974. ISBN 90 6196 092 4.

MCS 19 A. HORDIJK, R. POTHARST & J.Th. RUNNENBURG, *Optimaal stoppen van Markovketens*, 1974. ISBN 90 6196 093 2.

MCS 20 T.M.T. COOLEN, P.W. HEMKER, P.J. VAN DER HOUWEN & E. SLAGT, *ALGOL 60 procedures voor begin- en randwaardeproblemen*, 1976. ISBN 90 6196 094 0.

MCS 21 J.W. DE BAKKER (red.), *Colloquium Programmacorrectheid*, 1975. ISBN 90 6196 103 3.

MCS 22 R. HELMERS, F.H. RUYMGAART, M.C.A. VAN ZUYLEN & J. OOSTERHOFF, *Asymptotische methoden in de toetsingstheorie; Toepassingen van naburigheid*, 1976. ISBN 90 6196 104 1.

MCS 23.1 J.W. DE ROEVER (red.), *Colloquium Onderwerpen uit de biomathematica, deel 1*, 1976. ISBN 90 6196 105 X.

MCS 23.2 J.W. DE ROEVER (red.), *Colloquium Onderwerpen uit de biomathematica, deel 2*, 1976. ISBN 90 6196 115 7.

MCS 24.1 P.J. VAN DER HOUWEN, *Numerieke integratie van differentiaalvergelijkingen, deel 1: Eenstapsmethoden*, 1974. ISBN 90 6196 106 8.

MCS 25 *Colloquium Structuur van programmeertalen*, 1976. ISBN 90 6196 116 5.

MCS 26.1 N.M. TEMME (ed.), *Nonlinear analysis, volume 1*, 1976. ISBN 90 6196 117 3.

MCS 26.2 N.M. TEMME (ed.), *Nonlinear analysis, volume 2*, 1976. ISBN 90 6196 121 1.

MCS 27 M. BAKKER, P.W. HEMKER, P.J. VAN DER HOUWEN, S.J. POLAK & M. VAN VELDHUIZEN, *Colloquium Discretiseringsmethoden*, 1976. ISBN 90 6196 124 6.

MCS 28 O. DIEKMANN, N.M. TEMME (EDS), *Nonlinear Diffusion Problems*, 1976. ISBN 90 6196 126 2.

MCS 29.1 J.C.P. BUS (red.), *Colloquium Numerieke programmatuur, deel 1A, deel 1B*, 1976. ISBN 90 6196 128 9.

MCS 29.2 H.J.J. TE RIELE (red.), *Colloquium Numerieke programmatuur, deel 2*, 1976. ISBN 144 0.

* MCS 30 P. GROENEBOOM, R. HELMERS, J. OOSTERHOFF & R. POTHARST, *Efficiency begrippen in de statistiek*, . ISBN 90 6196 149 1.

MCS 31 J.H. VAN LINT (red.), *Inleiding in de coderingstheorie*, 1976. ISBN 90 6196 136 X.

MCS 32 L. GEURTS (red.), *Colloquium Bedrijfssystemen*, 1976. ISBN 90 6196 137 8.

MCS 33 P.J. VAN DER HOUWEN, *Differentieschema's voor de berekening van waterstanden in zeeën en rivieren*, 1977. ISBN 90 6196 138 6.

MCS 34 J. HEMELRIJK, *Oriënterende cursus mathematische statistiek*, ISBN 90 6196 139 4.

MCS 35 P.J.W. TEN HAGEN (red.), *Colloquium Computer Graphics*, 1977. ISBN 90 6196 142 4.

MCS 36 J.M. AARTS, J. DE VRIES, *Colloquium Topologische Dynamische Systemen*, 1977. ISBN 90 6196 143 2.

MCS 37 J.C. van Vliet (red.), *Colloquium Capita Datastructuren*, 1978. ISBN 90 6196 159 9.

MCS 38.1 T.H. KOORNWINDER (ED.), *Representations of locally compact groups with applications*, 1979. ISBN 90 6196 161 0.

MCS 38.2 T.H. KOORNWINDER (ED.), *Representations of locally compact groups with applications*, 1979. ISBN 90 6196 181 5.

MCS 39 O.J. VRIEZE & G.L. WANROOY, *Colloquium Stochastische Spelen*, 1978. ISBN 90 6196 167 X.

MCS 40 J. VAN TIEL, *Convexe Analyse*, 1979. ISBN 90 6196 187 4.

MCS 41 H.J.J. TE RIELE (ED.), *Colloquium Numerical Treatment of Integral Equations*, 1979. ISBN 90 6196 189 0.

MCS 42 J.C. VAN VLIET (RED.), *Colloquium Capita Implementatie van Programmeertalen*, 1980. ISBN 90 6196 191 2.

MCS 43 A.M. COHEN & H.A. WILBRINK, *Eindige groepen (Een inleidende cursus)*, 1980. ISBN 90 6196 203 X

MCS 44 J.G. VERWER (ED.), *Numerical solution of partial differential equations*, 1980. ISBN 90 6196 205 6.

MCS 45 P. KLINT (red.), *Colloquium hogere programmeertalen en computerarchitectuur*, 1980. ISBN 90 6196 206 4.

De met een * gemerkte uitgaven moeten nog verschijnen.