PRECONDITIONED CONJUGATE GRADIENTS AND MIXED-HYBRID FINITE ELEMENTS FOR THE SOLUTION OF POTENTIAL FLOW PROBLEMS

Typing: Angelina de Wit
Cover design: Frank Goddijn

PRECONDITIONED CONJUGATE GRADIENTS AND MIXED-HYBRID FINITE ELEMENTS FOR THE SOLUTION OF POTENTIAL FLOW PROBLEMS

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Delft, op gezag van de
Rector magnificus, Prof.drs. P.A. Schenck,
in het openbaar te verdedigen ten overstaan van een
commissie aangewezen door het College van Dekanen
op donderdag 12 april 1990 te 14.00 uur

door

ENRIQUE FRANCISCO KAASSCHIETER
geboren op 27 augustus 1960 te Buenos Aires
doctorandus in de wiskunde

Dit proefschrift is goedgekeurd door de promotor:
Prof.dr. H.A. van der Vorst

Aan mijn grootouders,
mijn ouders
en Ingeborg

Foreword.

This thesis comprises four essentially self-contained papers, preceded by a general introduction written primarily for the reader who is not specialized in groundwater hydraulics, finite element methods or preconditioned conjugate gradient methods.

The four papers that form the core of the thesis are:

I. A practical termination criterion for the conjugate gradient method,
by E.F. Kaasschieter.
BIT 28 (1988), pp. 308-322.

II. Preconditioned conjugate gradients for solving singular systems,
by E.F. Kaasschieter.
Journal of Computational and Applied Mathematics 24 (1988),
pp. 265-275.

III. A general finite element preconditioning for the conjugate gradient method,
by E.F. Kaasschieter.
To appear in BIT.

IV. Mixed-hybrid finite elements and streamline computation for the potential flow problem,
by E.F. Kaasschieter and A.J.M. Huijben.
Report PN 90-02-A of the TNO Institute of Applied Geoscience, Delft.

Contents.

GENERAL INTRODUCTION

The subject of this thesis are difficulties that may arise when solving potential flow problems numerically. Potential flow problems are fundamental in several fields of mathematical physics, e.g. heat conduction and electrostatics, but the main inspiration for this thesis comes from the field of groundwater hydraulics.

Therefore, in section 1 of this introduction the derivation of the potential flow problem in groundwater hydraulics is explained briefly (for details see, e.g., [5]). In section 2 the numerical solution of potential flow problems is discussed. Real-world problems give rise to several complications, some of which are treated in this thesis. The introduction concludes with a brief description of its contents.

1. Groundwater flow.

Subsoil generally consists of granular material with pores in between. Below a certain depth there is a saturated zone in which all pores are completely filled with water. This zone is bounded from below by impervious bedrock.

The saturated zone can be subdivided into aquifers and aquicludes. An aquifer is a geological formation that contains water and permits it to move through under ordinary field conditions. An aquiclude is a formation that may contain water, but is incapable of transmitting it under ordinary field conditions.

In an aquifer, groundwater is usually in motion. Groundwater motion occurs at very low velocities. However, because of the large cross-sectional areas through which this motion takes place, large quantities of water are transported.

The flow of groundwater takes place through the interconnected pores. When dealing with this flow, the microscopic flow patterns inside individual pores will be ignored and a fictitious average flow is considered. For this, the continuum approach is employed, i.e. it is assumed that all variables and parameters have their average meaning in a porous medium regarded as a continuum.

1.1. The continuity equation

The specific discharge $\boldsymbol{q}$ $[LT^{-1}]$ is defined as the volume of water flowing per unit time through a unit cross-sectional area normal to the direction of flow. Assume that the flow is stationary or that both fluid and porous medium are incompressible. Consider a volume V inside the flow domain. By the law of mass conservation, in the absence of sources or sinks the total outflow through the surface ∂V of V is equal to zero, i.e.

$$(1.1) \qquad \int_{\partial V} \boldsymbol{n} \cdot \boldsymbol{q} \, ds = 0,$$

where $\boldsymbol{n}$ is the outward normal to ∂V. Using Gauss's law, it follows from (1.1) that

$$(1.2) \qquad \int_{V} \nabla \cdot \boldsymbol{q} \, dx = 0.$$

Since (1.2) holds for every volume V inside the flow domain, the continuity equation follows, i.e.

$$(1.3) \qquad \nabla \cdot \boldsymbol{q} = 0.$$

1.2. Darcy's law

Inside an aquifer, one can measure the piezometric head (potential) ϕ $[L]$. This is usually done by constructing an observation well. At a certain depth water enters the well through a permeable filter. The piezometric head at this depth is defined as the height of the water level in the observation well. Of course, this height is measured with respect to some datum level. For a fluid with a constant specific weight, at a certain point in the flow domain it holds that

$$(1.4) \qquad \phi = z + p/\gamma,$$

where z is the height of this point, p is the pressure at this point and γ is the specific weight of water.

The relation between the specific discharge $\boldsymbol{q}$ and the piezometric head ϕ is given by Darcy's law, i.e.

$$(1.5) \qquad \boldsymbol{q} = -\boldsymbol{K}\nabla\phi.$$

Here $\boldsymbol{K}$ $[LT^{-1}]$ is the second rank tensor of hydraulic conductivity (for a detailed discussion on second rank tensors see, e.g., [13: section 1.5]). The tensor $\boldsymbol{K}$ is symmetric, thus only six distinct entries are needed to fully define the hydraulic conductivity. $\boldsymbol{K}$ expresses the ease with which a fluid is transported through a porous medium. Therefore, it depends on both solid and fluid properties. If the hydraulic conductivity at a certain point is independent of the direction, then the medium is said to be isotropic at that point. In this case,

$$(1.6) \qquad K = kI,$$

where k is a scalar and I is the unit tensor.

1.3. The potential flow problem

The continuity equation and Darcy's law contain no information related to any specific case of flow through a porous medium. The supplementary information that together with equations (1.3) and (1.5) defines an individual problem should include specifications of:

(i) the geometry of the domain Ω in which the flow under consideration takes place;

(ii) values of the tensor of hydraulic conductivity inside Ω;

(iii) statements on how the fluid in Ω interacts with its surroundings, i.e. boundary conditions.

Let $\partial\Omega_D$ and $\partial\Omega_N$ be portions of the boundary $\partial\Omega$ of Ω. We consider two types of boundary conditions:

* Boundary of prescribed potential, i.e.

$$\phi = g_D \quad \text{on } \partial\Omega_D, \tag{1.7}$$

where g_D is a known function. Condition (1.7) is called a Dirichlet boundary condition.

* Boundary of prescribed flux, i.e.

$$\boldsymbol{n}\cdot\boldsymbol{q} = g_N \quad \text{on } \partial\Omega_N, \tag{1.8}$$

where $\boldsymbol{n}$ is the outward normal to $\partial\Omega$ and g_N is a known function. Condition (1.8) is called a Neumann boundary condition. Combining (1.3), (1.5), (1.7) and (1.8), we obtain the divergence-free potential flow problem:

$$\boxed{\begin{array}{l} \nabla\cdot\boldsymbol{q} = 0,\ \boldsymbol{q} = -\boldsymbol{K}\nabla\phi \ \text{ in } \Omega, \\ \phi = g_D \ \text{ on } \partial\Omega_D,\ \boldsymbol{n}\cdot\boldsymbol{q} = g_N \ \text{ on } \partial\Omega_N. \end{array}} \tag{1.9}$$

This problem is a well-posed problem, i.e. the solution exists, is unique and depends continuously on the data. Note that if $\partial\Omega_N = \partial\Omega$, then ϕ is only unique up to a constant.

1.4. Sources and sinks

Often, injection or production wells are in the flow domain Ω. The discharge of a well is denoted by Q $[L^3T^{-1}]$. The permeable filter of an injection or production well acts as a source or sink. Consider a volume V around such a source or sink. By the law of mass conservation and Gauss's law, we have

(1.10) $$\int_V \nabla\cdot\boldsymbol{q}\,dx = -Q.$$

It is convenient to introduce a source function f $[T^{-1}]$, such that

(1.11) $$\int_V f\,dx = -Q$$

for all volumes V around the source or sink. It follows immediately from (1.10) and (1.11) that

(1.12) $$\nabla\cdot\boldsymbol{q} = f.$$

If we replace (1.3) by (1.12) in problem (1.9), we obtain

(1.13) $$\boxed{\begin{array}{l}\nabla\cdot\boldsymbol{q} = f,\ \boldsymbol{q} = -\boldsymbol{K}\nabla\phi \ \text{ in } \Omega,\\ \phi = g_D \ \text{ on } \partial\Omega_D,\ \boldsymbol{n}\cdot\boldsymbol{q} = g_N \ \text{ on } \partial\Omega_N.\end{array}}$$

This well-posed problem is referred to as the general potential flow problem.

1.5. The velocity

Let $\boldsymbol{v}$ $[LT^{-1}]$ be the average velocity of the fluid, then

(1.14) $$\boldsymbol{v} = \boldsymbol{q}/n,$$

where n [-] is the porosity. Consider a certain point in the flow domain and a small representative volume V around it (see [5: section 2-5]). Then the porosity n in this point is equal to the ratio of the interconnected pore space in V and the volume of V. Of course, $0 < n < 1$.

2. Numerical solution of the potential flow problem.

Only in special cases can an exact solution of a potential flow problem (see (1.13)) be derived. Therefore, numerical methods are the major tool for solving such problems as encountered in practice. Generally, numerical complications will arise because of

(i) the irregularity of the shape of the flow domain under consideration;
(ii) the large spatial variation in the hydraulic conductivity, usually with jumps of several orders of magnitude along irregularly shaped internal boundaries;
(iii) the small vertical scale (tens of metres) vs. the large horizontal scale (kilometres) of the flow domain;
(iv) the very small size of wells vs. the large size of the flow domain.

2.1. The conforming finite element method

The finite element method is a very powerful tool for determining an approximation of the solution of a real-world potential flow problem. Using this method the flow domain Ω is subdivided into a small number of subdomains called finite elements. Each subdomain has a simple geometrical shape, e.g. a tetrahedron or a block. In each subdomain the solution is approximated by a polynomial function. This piecewise polynomial approximation has to fulfil certain continuity conditions along the interelement boundaries.

Using (1.5), the potential flow problem (1.13) can be rewritten into the elliptic boundary value problem:

$$\boxed{\begin{array}{l} -\nabla\cdot(K\nabla\phi) = f \ \text{ in } \Omega, \\ \phi = g_D \ \text{ on } \partial\Omega_D, \ -\boldsymbol{n}\cdot(K\nabla\phi) = g_N \ \text{ on } \partial\Omega_N. \end{array}} \tag{2.1}$$

The conforming finite element method (see, e.g., [6], [3], [15]) determines a piecewise polynomial approximation ϕ_h of the solution ϕ of (2.1), where ϕ_h has to be continuous along the interelement boundaries.

Using Green's formula, (2.1) is transformed into a variational problem. The approximation ϕ_h is determined as the unique solution of the associate discrete variational problem. Now, ϕ_h is written as a linear combination of the form

$$(2.2) \qquad \phi_h(x) = \sum_{i=1}^{n} \phi_i \, \psi_i(x), \; x \in \Omega,$$

where ψ_i, $i = 1,\ldots,n$, are the global basis functions of the finite-dimensional space in which the approximation ϕ_h is sought. The coefficients ϕ_i are taken to be the values of ϕ_h at the global nodes x_i, $i = 1,\ldots,n$, within Ω. Define $\Phi = (\phi_1,\ldots,\phi_n)^T$, then the approximation ϕ_h is such that the vector Φ is the solution of the system of linear equations

$$(2.3) \qquad A\Phi = F.$$

Here, $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $F \in \mathbb{R}^n$.

2.2. The preconditioned conjugate gradient method

The coefficient matrix A generally is large, sparse and ill-conditioned. A direct method for solving (2.3), e.g. using a Cholesky decomposition (see [8: section 5.2]), results in factors of A that are substantially less sparse. A vast amount of computer storage is needed to store the entries of these factors. Moreover, the computation of these factors results in many floating point operations.

Both drawbacks can be circumvented by the use of iterative methods for solving the system (2.3) (see [10]). Starting with a first guess $\Phi_0 \in \mathbb{R}^n$ successive approximations $\Phi_1,\Phi_2,\ldots$ of the solution Φ are computed. These approximations must converge to Φ.

In [11] the conjugate gradient method is introduced to solve a system of linear equations with a symmetric positive definite matrix. Unfortunately the conjugate gradient method converges rather slowly for ill-conditioned matrices. An important way around this difficulty is to precondition A (see, e.g., [8: section 10.3], [3: section 1.4]).

This refers to finding a nonsingular matrix C, such that $\tilde{A} = C^{-1}AC^{-T}$ has a more favourable distribution of its eigenvalues than the original matrix A (for details on the rate of convergence of the conjugate gradient method see [4], [16]). We can then apply the conjugate gradient method (with improved convergence properties) to the transformed system

$$\tilde{A}\tilde{\Phi} = \tilde{F}, \tag{2.4}$$

where $\tilde{\Phi} = C^T\Phi$ and $\tilde{F} = C^{-1}F$. After transforming the iterates we obtain the preconditioned conjugate gradient method with respect to the preconditioning matrix $M = CC^T$.

A variety of choices for the preconditioning matrix M has been discussed in the literature (for surveys see [2], [7]). Popular methods for computing M are to use an incomplete Cholesky decomposition (see [12]) or a modified incomplete Cholesky decomposition (see [9]).

2.3. The mixed-hybrid finite element method

Although the conforming finite element method is very appropriate to determine an accurate approximation of the solution ϕ of the potential flow problem (1.13), it is not always suitable for obtaining an accurate approximation of the specific discharge $\boldsymbol{q} = -\boldsymbol{K}\nabla\phi$. Using the conforming finite element method, the piecewise polynomial approximation of ϕ is differentiated in each finite element and multiplied by the tensor $\boldsymbol{K}$ to obtain an approximation of $\boldsymbol{q}$. In solving tough real-world problems, an inaccurate specific discharge results from this approach, i.e. the approximation thus obtained does not fulfil the continuity equation (1.12) accurately.

An accurate approximation of q can be determined by the mixed finite element method (see [14]). This method starts from the original problem (1.13).

The mixed finite element method determines piecewise polynomial approximations q_h and ϕ_h of the solutions q and ϕ of (1.13), where the normal component of q_h has to be continuous across the interelement boundaries.

Eventually, a large system of linear equations is obtained. The choice of a numerical method to solve this system is restricted by the fact that its matrix is indefinite. This drawback can be circumvented by an implementation technique called hybridization, which leads to a symmetric positive definite system of linear equations (see [1]). Since this system is sparse, it can be solved efficiently by the preconditioned conjugate gradient method.

3. Outline of the thesis.

In this thesis various aspects concerning finite element methods and preconditioned conjugate gradient methods will be discussed.

Chapter I deals with the conjugate gradient method for the iterative solution of a system of linear equations $Ax = b$. It is shown how the smallest active eigenvalue of A can be cheaply approximated, and the usefulness of this approximation for a practical termination criterion for the conjugate gradient method is ascertained. It is proved that this termination criterion is reliable in many relevant situations.

In chapter II the preconditioned conjugate gradient method is used to solve the system of linear equations $Ax = b$, where A is a singular matrix. The method diverges if b is not exactly in the range of A. If the null space of A is explicitly known, then this divergence can be avoided by subtracting from b its orthogonal projection onto the null space. As well as analysing this subtraction, conditions necessary for the existence of the incomplete Cholesky decomposition are given. Finally, the theory is applied to the discretized potential flow problem with Neumann boundary conditions.

Discretizing a symmetric elliptic boundary value problem by a finite element method results in a system of linear equations with a symmetric positive definite matrix. In chapter III a preconditioning matrix is proposed that can be constructed for all finite element methods if a mild condition for the node numbering is fulfilled. Such a numbering can be constructed by a variant of the Cuthill-McKee algorithm.

In chapter IV the lowest order mixed-hybrid finite element method is discussed in detail for general potential flow problems. The elementwise computation of streamlines and residence times is presented.

REFERENCES

[1] Arnold, D.N. and F. Brezzi, *Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates*, Mathematical Modelling and Numerical Analysis 19 (1985), pp. 7-32.

[2] Axelsson, O., *A survey of preconditioned iterative methods for linear systems of algebraic equations*, BIT 25 (1985), pp. 166-187.

[3] Axelsson, O. and V.A. Barker, *Finite Element Solution of Boundary Value Problems*, Academic Press, New York (1984).

[4] Axelsson, O. and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numerische Mathematik 48 (1986), pp. 499-523.

[5] Bear, J., *Hydraulics of Groundwater*, McGraw-Hill, New York (1979).

[6] Ciarlet, P.G., *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam (1978).

[7] Concus, P., G.H. Golub and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM Journal of Scientific and Statistical Computation 6 (1985), pp. 220-252.

[8] Golub, G.H. and C.F. van Loan, *Matrix Computations*, North Oxford Academic, Oxford (1983).

[9] Gustafsson, I., *A class of first order factorization methods*, BIT 18 (1978), pp. 142-156.

[10] Hageman, L.A. and D.M. Young, *Applied Iterative Methods*, Academic Press, New York (1981).

[11] Hestenes, M.R. and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards 49 (1952), pp. 409-436.

[12] Meijerink, J.A. and H.A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Mathematics of Computation 31 (1977), pp. 148-162.

[13] Morse, P.M. and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, New York (1953).

[14] Raviart, P.-A. and J.-M. Thomas, *A mixed finite element method for 2-nd order elliptic problems*, in: I. Galligani and E. Magenes (eds.), *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Mathematics 606, Springer, Berlin (1977), pp. 292-315.

[15] Schwarz, H.R., *The Finite Element Method*, Academic Press, New York (1988).

[16] van der Sluis, A. and H.A. van der Vorst, *The rate of convergence of conjugate gradients*, Numerische Mathematik 48 (1986), pp. 543-560.

CHAPTER I

A PRACTICAL TERMINATION CRITERION FOR THE CONJUGATE GRADIENT METHOD

E. F. KAASSCHIETER *

Department of Mathematics and Informatics, Delft University of Technology, P.O. Box 356, 2600 AJ Delft, The Netherlands

Abstract.

The conjugate gradient method for the iterative solution of a set of linear equations $Ax = b$ is essentially equivalent to the Lanczos method, which implies that approximations to certain eigenvalues of A can be obtained at low cost. In this paper it is shown how the smallest "active" eigenvalue of A can be cheaply approximated, and the usefulness of this approximation for a practical termination criterion for the conjugate gradient method is studied. It is proved that this termination criterion is reliable in many relevant situations.

AMS(MOS) Classifications: 65F10, 65F50.

1. The conjugate gradient method.

In [4] the conjugate gradient method (*cg*-method) is introduced to solve a set of linear equations

$$Ax = b, \tag{1.1}$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite and $b \in \mathbb{R}^n$.

Starting with a vector $x_0 \in \mathbb{R}^n$ successive approximations $x_1, x_2, \ldots$ are computed in this method according to

ALGORITHM 1:

```
$r_0 := b - Ax_0$
for $i = 0, 1, \ldots$
    if $r_i = 0$ then stop
    $\beta_{i-1} := r_i^T r_i / r_{i-1}^T r_{i-1}$    $(\beta_{-1} := 0)$
    $p_i := r_i + \beta_{i-1} p_{i-1}$    $(p_0 := r_0)$
    $\alpha_i := r_i^T r_i / p_i^T A p_i$
    $x_{i+1} := x_i + \alpha_i p_i$
    $r_{i+1} := r_i - \alpha_i A p_i$.
```

* Present Address: TNO-DGV Institute of Applied Geoscience, P.O. Box 285, 2600 AG Delft, The Netherlands.

Received April 1986. Revised October 1987.

For the basic relations in the *cg*-method, see section 10.2 of [1], section IV-2 of [3] and section 5 of [4]. It follows from theorem IV-4.1 of [3], that the *cg*-method terminates in $N \leq n$ steps, where N is the smallest integer, such that the vectors $r_0, Ar_0, \ldots, A^N r_0$ are linearly dependent.

For further reference the following theorem is proved.

Theorem 1: *After $N \leq n$ iterations of the cg-method (N as previously defined)*

$$\operatorname{span}\{r_0, Ar_0, \ldots, A^{N-1} r_0\} = \operatorname{span}\{u_1, \ldots, u_N\},$$

where $u_1, \ldots, u_N$ are normalized eigenvectors of A corresponding to eigenvalues $0 < \mu_1 < \ldots < \mu_N$[*]).

Proof: $A^N r_0$ is a linear combination of $r_0, Ar_0, \ldots, A^{N-1} r_0$.

Hence $\operatorname{span}\{r_0, Ar_0, \ldots, A^{N-1} r_0\}$ is an invariant subspace of A with dimension N. From this the theorem follows directly. ■

In practice exact termination, i.e. $r_N = 0$, is prevented because of rounding errors. Moreover, an approximation to the solution of (1.1), obtained long before exact termination should occur, is often sufficient. Therefore the following (relative) termination criterion is chosen:

$$\|x - x_i\| / \|x\| \leq \varepsilon, \tag{1.2}$$

where $\varepsilon > 0$ is a preordained accuracy and $x \neq 0$[**]).

Note that it is also possible to choose the absolute termination criterion $\|x - x_i\| \leq \varepsilon$. The derivation and the analysis of a practical absolute termination criterion are obvious from the following presentation.

Unfortunately it is impossible to determine $\|x\|$ and $\|x - x_i\|$ cheaply, because the solution x is unknown. To get rid of $\|x\|$ the following theorem can be used:

Theorem 2. *If $\|x - x_i\| \leq \|x_i\|\varepsilon/(1+\varepsilon)$, then $\|x - x_i\| \leq \|x\|\varepsilon$.*

Proof. If $x - x_i = 0$, then the assertion follows directly. Therefore it is assumed that $x - x_i \neq 0$.

According to the triangle inequality we have:

$$\|x_i\| \leq \|x\| + \|x - x_i\|.$$

[*]) The linear subspace $K^j(A, r_0) = \operatorname{span}\{r_0, Ar_0, \ldots, A^{j-1} r_0\}$ is called the jth Krylov subspace of A with respect to r_0.

[**]) In the following, $\|x\|$ is written for the Euclidean norm of a vector $x \in \mathbf{R}^n$.

From this it follows that

$$\frac{\|x\|}{\|x+x_i\|} \geq \frac{\|x_i\| - \|x-x_i\|}{\|x-x_i\|} = \frac{\|x_i\|}{\|x-x_i\|} - 1 \geq \frac{1+\varepsilon}{\varepsilon} - 1 = \frac{1}{\varepsilon}. \qquad \blacksquare$$

Note that $\varepsilon/(1+\varepsilon) \approx \varepsilon$ for small ε.

On the other hand, after i iterations of the *cg*-method:

$$x - x_i = A^{-1}(b - Ax_i) = A^{-1}r_i$$

and then, because of theorem 1:

$$\|x - x_i\| = \|A^{-1}r_i\| \leq \|r_i\|/\mu_1.$$

The criterion (1.2) can therefore be replaced by

$$\|r_i\| \leq \mu_1\|x_i\|\varepsilon/(1+\varepsilon) \tag{1.3}$$

from which (1.2) follows. Hence it is necessary to determine the smallest "active" eigenvalue μ_1 of A (see theorem 1 and [6], section 2.2).

Since the *cg*-method is essentially equivalent to the Lanczos method, after i iterations it is possible to obtain an approximation $\mu_1^{(i)}$ to μ_1 from the iteration constants $\alpha_0, \alpha_1, \ldots, \alpha_{i-1}$ and $\beta_0\beta_1, \ldots, \beta_{i-2}$ (see [5], chapter 7). The idea of approximating $\mu_1^{(i)}$ from these constants has also been proposed in section 7.5 of [2]. The next section contains a more elegant and cheaper algorithm for the identification of $\mu_1^{(i)}$. The approximation is updated after each iteration of the *cg*-method.

In the last section it will be shown that the termination criterion, introduced in section 3, is reliable in many relevant situations. A very modest degree of convergence of $\mu_1^{(i)}$ leads to a strict (and computable) upper bound for $\|x - x_i\|/\|x\|$.

2. An approximation for the smallest active eigenvalue.

If the *cg*-method passes through $i \leq N$ iterations, then

$$Ar_j = -\frac{\beta_{j-1}}{\alpha_{j-1}}r_{j-1} + \left(\frac{1}{\alpha_j} + \frac{\beta_{j-1}}{\alpha_{j-1}}\right)r_j - \frac{1}{\alpha_j}r_{j+1} \quad \text{for } j = 1(1)i-1. \tag{2.1}$$

Let $R_i = [r_0, r_1, \ldots, r_{i-1}] \in \mathbb{R}^{n \times i}$, then (2.1) can be written in matrix form:

$$AR_i = R_iT_i - \frac{1}{\alpha_{i-1}}r_ie_i^T, \tag{2.2}$$

where $e_i = (0, \ldots, 0, 1)^T$ is the ith unit vector of dimension i and

$$(2.3) \qquad T_i = \begin{bmatrix} 1/\alpha_0 & -\beta_0/\alpha_0 & & \\ -1/\alpha_0 & 1/\alpha_1+\beta_0/\alpha_0 & \ddots & \\ & -1/\alpha_1 & \ddots & -\beta_{i-1}/\alpha_{i-2} \\ & & -1/\alpha_{i-2} & 1/\alpha_{i-1}+\beta_{i-2}/\alpha_{i-2} \end{bmatrix}$$

From the equivalence of the cg-method and the Lanczos method the eigenvalues of T_i have the following property:

Theorem 3: (*Strict interlacing property*). *For $i = 1(1)N$ the matrix $T_i \in \mathbb{R}^{i\times i}$ has the real eigenvalues* *) $\mu_1^{(i)}, \ldots, \mu_i^{(i)}$, *such that*

$$0 < \mu_1 \leq \mu_1^{(i+1)} < \mu_1^{(i)} < \mu_2^{(i+1)} < \ldots < \mu_i^{(i+1)} < \mu_i^{(i)} < \mu_{i+1}^{(i+1)} \leq \mu_N$$

for $i = 1(1)N-1$.

Proof: See [7], corollary 6.2. ■

For $i = 1(1)N$ the approximation $\mu_1^{(i)}$ of μ_1 is obtained as the smallest eigenvalue of T_i or, equivalently, as the smallest root of the normalized characteristic polynomial φ_i defined by

$$(2.4) \qquad \varphi_i(x) = \det(T_i - xI)/\det T_i \quad \text{for } x \in \mathbb{R}.$$

The value of $\varphi_i(x)$, for any given x, can be determined recursively by the formula

$$(2.5) \qquad \boxed{\begin{aligned} &\varphi_0(x) := 1, \\ &\varphi_1(x) := 1 - \alpha_0 x, \\ &\varphi_{j+1}(x) := (1 + \alpha_j\beta_{j-1}/\alpha_{j-1} - \alpha_j x)\varphi_j(x) - (\alpha_j\beta_{j-1}/\alpha_{j-1})\varphi_{j-1}(x) \\ &\hfill \text{for } j = 1(1)i-1. \end{aligned}}$$

From the formal analogy of (2.1) and (2.5) and the determination of r_i according to algorithm 1 it follows that for $i = 1(1)N$ the polynomial φ_i can be determined recursively by the formula

$$(2.6) \qquad \boxed{\begin{aligned} &\psi_0(x) := 1, \quad \varphi_1(x) := 1 - \alpha_0 x, \\ &\psi_j(x) := \varphi_j(x) + \beta_{j-1}\psi_{j-1}(x), \quad \varphi_{j+1}(x) := \varphi_j(x) - \alpha_j x \psi_j(x) \\ &\hfill \text{for } j = 1(1)i-1. \end{aligned}}$$

*) The eigenvalues $\mu_1^{(i)}, \ldots, \mu_i^{(i)}$ are called the Ritz values of A with respect to $K^i(A, r_0)$.

Note that the computation of $\phi_i(x)$, for a given $x \in \mathbb{R}$, according to (2.6) is cheaper than when using (2.5).

For $i = 2(1)N$ the numbers $\mu_1^{(i)}$ can be determined by, e.g., employing the bisection method to the normalized characteristic polynomial ϕ_i (see [1], section 8.5). The aim is to find an approximation $\tilde{\mu}_1^{(i)}$ of $\mu_1^{(i)}$ with a relative error smaller than u, i.e.

$$(2.7) \qquad |\tilde{\mu}_1^{(i)} - \mu_1^{(i)}|/\mu_1^{(i)} < u,$$

where $u > 0$ is a preordained tolerance.

Since $\phi_i(0) = 1$ the following variant may be constructed on the bisection method:

Algorithm 2:

```
$z := \tilde{\mu}_1^{(i-1)}$
**if** $(\phi_j(z) > 0$ for $j = 1(1)i)$ **then** $\tilde{\mu}_1^{(i)} := z$ **else**
  $y := 0$
  **while** $z - y > uy$
    $x := (y+z)/2$
    **if** $(\phi_j(x) > 0$ for $j = 1(1)i)$ **then** $y := x$ **else** $z := x$
$\tilde{\mu}_1^{(i)} := x.$
```

In the following analysis of this algorithm it is proved by induction that the approximation $\tilde{\mu}_1^{(i)}$, computed by algorithm 2, to $\mu_1^{(i)}$ always satisfies (2.7).

Assume that, before applying algorithm 2, an approximation $\tilde{\mu}_1^{(i-1)} > 0$ to $\mu_1^{(i-1)}$ is known, such that $|\tilde{\mu}_1^{(i-1)} - \mu_1^{(i-1)}|/\tilde{\mu}_1^{(i-1)} \le u$. Two situations are then possible:

$$(2.8a) \qquad 0 < \mu_1^{(i)} \le \tilde{\mu}_1^{(i-1)} \quad \text{or}$$

$$(2.8b) \qquad 0 < \tilde{\mu}_1^{(i-1)} < \mu_1^{(i)}.$$

In the case of (2.8a) the following two situations can arise after using algorithm 2:

I. An approximation $\tilde{\mu}_1^{(i)}$ to $\mu_1^{(i)}$ and a number z are obtained, such that

$$0 < \tilde{\mu}_1^{(i)} < \mu_1^{(i)} \le z \le \tilde{\mu}_1^{(i-1)} \quad \text{and } z - \tilde{\mu}_1^{(i)} \le u\tilde{\mu}_1^{(i)}.$$

Hence $0 < \mu_1^{(i)} - \tilde{\mu}_1^{(i)} \le z - \tilde{\mu}_1^{(i)} \le u\tilde{\mu}_1^{(i)} < u\mu_1^{(i)}$.

II. An approximation $\tilde{\mu}_1^{(i)}$ to $\mu_1^{(i)}$ and a number y are obtained, such that

$$0 \le y < \mu_1^{(i)} \le \tilde{\mu}_1^{(i)} < \tilde{\mu}_1^{(i-1)} \quad \text{and } \tilde{\mu}_1^{(i)} - y \le uy.$$

Hence $0 \le \tilde{\mu}_1^{(i)} - \mu_1^{(i)} < \tilde{\mu}_1^{(i)} - y \le uy < u\mu_1^{(i)} \le u\tilde{\mu}_1^{(i)}$.

In the case of (2.8b) the approximation $\tilde{\mu}_1^{(i)} = \tilde{\mu}_1^{(i-1)}$ to $\mu_1^{(i)}$ is obtained as a result of algorithm 2. It follows that

$$0 < \mu_1^{(i)} - \tilde{\mu}_1^{(i)} = \mu_1^{(i)} - \tilde{\mu}_1^{(i-1)} < \mu_1^{(i-1)} - \tilde{\mu}_1^{(i-1)} \le u\tilde{\mu}_1^{(i-1)} = u\tilde{\mu}_1^{(i)} < u\mu_1^{(i.)}.$$

Hence it follows, that after using algorithm 2 an approximation $\tilde{\mu}_1^{(i)}$ to $\mu_1^{(i)}$ is obtained, such that $|\tilde{\mu}_1^{(i)} - \mu_1^{(i)}|/\mu_1^{(i)} < u$ and $|\tilde{\mu}_1^{(i)} - \mu_1^{(i)}|/\tilde{\mu}_1^{(i)} \le u$.

This completes the analysis.

3. A practical termination criterion.

In the previous section it was shown how, after $i \le N$ iterations of the *cg*-method, the smallest active eigenvalue μ_1 of A is approximated by $\tilde{\mu}_1^{(i)}$. Hence it is tempting to replace criterion (1.3) by

$$\|r_i\| \le \tilde{\mu}_1^{(i)}\|x_i\|\varepsilon/(1+\varepsilon). \tag{3.1}$$

In order to obtain a first approximation $\tilde{\mu}_1^{(1)}$ to μ_1 it is anyway necessary to carry out the first iteration of the *cg*-method. A combination of algorithms 1 and 2, which includes the termination criterion (3.1) is given by

```
ALGORITHM 3:
    $r_0 := b - Ax_0$                                  }
    $p_0 := r_0$                                       }
    $\alpha_0 := r_0^T r_0 / p_0^T A p_0$              }
    $x_1 := x_0 + \alpha_0 p_0$                        }  (first iteration of algorithm 1)
    $r_1 := r_0 - \alpha_0 A p_0$                      }
    $\tilde{\mu}_1^{(1)} := 1/\alpha_0$                }
    for $i = 1(1)n-1$
      if $\|r_i\| \le \tilde{\mu}_1^{(i)}\|x_i\|\varepsilon/(1+\varepsilon)$ then stop   }
      $\beta_{i-1} := r_i^T r_i / r_{i-1}^T r_{i-1}$   }
      $p_i := r_i + \beta_{i-1} p_{i-1}$               }
      $\alpha_i := r_i^T r_i / p_i^T A p_i$            }  (iteration of algorithm 1)
      $x_{i+1} := x_i + \alpha_i p_i$                  }
      $r_{i+1} := r_i - \alpha_i A p_i$                }
      $z := \tilde{\mu}_1^{(i)}$
      if ($\varphi_j(z) > 0$ for $j = 1(1)i+1$) then $\tilde{\mu}_1^{(i-1)} := z$ else   }
        $y := 0$                                       }
        while $z - y > uy$                             }
          $x := (y+z)/2$                               }  (algorithm 2)
          if ($\varphi_j(x) > 0$ for $j = 1(1)i+1$) then $y := x$ else   }
            $z := x$                                   }
      $\tilde{\mu}_1^{(i+1)} := x.$                    }
```

4. Analysis of termination strategy.

In this section the reliability of the termination criterion

$$\|r_i\| \leq \mu_1^{(i)}\|x_i\|\varepsilon/(1+\varepsilon) \tag{4.1}$$

is discussed instead of (3.1), i.e. $\tilde{\mu}_1^{(i)}$ is replaced by $\mu_1^{(i)}$. This does not make much difference, because an approximation $\tilde{\mu}_1^{(i)}$ to $\mu_1^{(i)}$ is obtained with a preordained (small) tolerance $u > 0$.

Since $\mu_1^{(i)}$ tends to be closer and closer to μ_1 for increasing i, the termination criterion (4.1) might be expected to be more useful than the commonly used termination criteria, based on the residual, i.e. $\|r_i\| \leq \varepsilon$, or on the reduction of the residual, i.e. $\|r_i\|/\|r_0\| \leq \varepsilon$. To discuss (4.1), first note that

$$x - x_0 = \sum_{j=1}^{N} \xi_j u_j, \tag{4.2}$$

where $u_1, \ldots, u_N$ are normalized eigenvectors of A corresponding to eigenvalues $0 < \mu_1 < \ldots < \mu_N$. If the weights ξ_j are small for small indices j, a reasonably accurate approximation $\mu_1^{(i)}$ of μ_1 is only obtained after many iterations of algorithm 3 (see [5], [7]). Note that in this case the criterion (1.3) is too strong.

In fact, we need a "lower bound" for the active eigenvalues μ_j of A relative to the weights ξ_j. In other words, it is necessary, that

$$\mu_1^{(i)} \leq \|r_i\|/\|x - x_i\|. \tag{4.3}$$

If this condition is satisfied, then using theorem 2 and (4.1) guarantees that $\|x - x_i\|/\|x\| \leq \varepsilon$.

Various numerical experiments, e.g. with an isolated smallest active eigenvalue or with a cluster of smallest active eigenvalues of A and with different weights $\xi_1, \ldots, \xi_N$, give confidence that (4.3) is satisfied long before exact termination should occur.

To analyse (4.3) in more detail, note that

$$x - x_i = \phi_i(A)(x - x_0) = \sum_{j=1}^{N} \xi_j \phi_i(\mu_j) u_j \quad \text{for } i = 0(1)N, \tag{4.4}$$

where $\phi_i(0) = 1$ (see [6], (2.9)), so that

$$\|r_i\|^2/\|x - x_i\|^2 = \sum_{j=1}^{N} \gamma_j^{(i)} \mu_j^2 \quad \text{for } i = 0(1)N-1, \quad \text{where} \tag{4.5a}$$

$$\gamma_j^{(i)} = \frac{\xi_j^2 \phi_i^2(\mu_j)}{\sum_{k=1}^{N} \xi_k^2 \phi_j^2(\mu_k)} \quad \text{for } j = 1(1)N. \tag{4.5b}$$

Note that $\sum_{j=1}^{N}\gamma_j^{(i)} = 1$. The quotient $\|r_i\|^2/\|x-x_i\|^2$ is equal to the weighted mean of $\mu_1^2,\ldots,\mu_N^2$ with the weights $\gamma_1^{(i)},\ldots,\gamma_N^{(i)}$.

In theorem 6 it is proved that (4.3) is satisfied, if $\mu_1^{(i)}$ has converged sufficiently, i.e.

$$\mu_1 < \mu_1^{(i)} \leq 2\mu_1\mu_2/(\mu_1+\mu_2). \tag{4.6}$$

For this purpose the following two lemmata are needed:

Lemma 4: *For $i = 1(1)N$ define the polynomial χ_i by*

$$\chi_i(x) = \frac{\mu_1^{(i)}}{\mu_1^{(i)}-x}\,\phi_i(x) = \prod_{j=2}^{i}\frac{\mu_j^{(i)}-x}{\mu_j^{(i)}}. \tag{4.7}$$

Then

$$\mu_1^{(i)} = \sum_{j=1}^{N}\delta_j^{(i)}\mu_j \qquad \textit{for } i = 1(1)N, \quad \textit{where} \tag{4.8a}$$

$$\delta_j^{(i)} = \frac{\xi_j^2\chi_i^2(\mu_j)\mu_j^2}{\sum_{k=1}^{N}\xi_k^2\chi_i^2(\mu_k)\mu_k^2} \qquad \textit{for } j = 1(1)N. \tag{4.8b}$$

Proof. See [7], (5.14).

Note that the weights in [7] correspond to r_0, so that in our case the weights of $x-x_0$ have to be multiplied by the corresponding eigenvalues. ■

Note that $\sum_{j=1}^{N}\delta_j^{(i)} = 1$, and hence $\mu_1^{(i)}$ is equal to the weighted mean of $\mu_1,\ldots,\mu_N$ with the weights $\delta_1^{(i)},\ldots,\delta_N^{(i)}$.

Lemma 5: *If $0 < \mu_1 < \ldots < \mu_N$, $\sum_{k=1}^{j}\delta_k \geq \sum_{k=1}^{j}\gamma_k$ for $j = 1(1)N-1$ and $\sum_{k=1}^{N}\delta_k = \sum_{k=1}^{N}\gamma_k$, then*

$$\sum_{j=1}^{N}\delta_j\mu_j \leq \sum_{j=1}^{N}\gamma_j\mu_j.$$

Proof. For $j = 1(1)N-1$ it holds that

$$\begin{aligned}
&\sum_{k=1}^{j}\gamma_k\mu_k + \left\{\sum_{k=1}^{j}(\delta_k-\gamma_k)\right\}\mu_j + \sum_{k=j+1}^{N}\delta_k\mu_k\\
\leq &\sum_{k=1}^{j}\gamma_k\mu_k + \left\{\sum_{k=1}^{j}(\delta_k-\gamma_k)\right\}\mu_{j+1} + \sum_{k=j+1}^{N}\delta_k\mu_k\\
= &\sum_{k=1}^{j+1}\gamma_k\mu_k + \left\{\sum_{k=1}^{j+1}(\delta_k-\gamma_k)\right\}\mu_{j+1} + \sum_{k=j+2}^{N}\delta_k\mu_k.
\end{aligned}$$

Since $\sum_{k=1}^{N}(\delta_k-\gamma_k) = 0$, it holds that

$$\sum_{k=1}^{N} \delta_k\mu_k = \gamma_1\mu_1+(\delta_1-\gamma_1)\mu_1+\sum_{k=2}^{N} \delta_k\mu_k$$

$$\leq \sum_{k=1}^{N} \gamma_k\mu_k+\left\{\sum_{k=1}^{N}(\delta_k-\gamma_k)\right\}\mu_N = \sum_{k=1}^{N} \lambda_k\mu_k,$$

where the inequality follows by an induction argument. ■

THEOREM 6: *If* $0 < (\mu_1^{(i)}-\mu_1)/\mu_1 \leq (\mu_2-\mu_1^{(i)})))/\mu_2$, *then* $\mu_1^{(i)} \leq \|r_i\|/\|x-x_i\|$.

PROOF. For $j = 1(1)N$ it holds that

$$\left\{\sum_{k=1}^{j} \gamma_k^{(i)}\right\}^{-1} = \frac{\sum_{k=1}^{N} \xi_k^2\phi_i^2(\mu_k)}{\sum_{k=1}^{j} \xi_k^2\phi_i^2(\mu_k)} = \frac{\sum_{k=1}^{N} \xi_k^2\chi_i^2(\mu_k)(\mu_1^{(i)}-\mu_k)^2}{\sum_{k=1}^{j} \xi_k^2\chi_i^2(\mu_k)(\mu_1^{(i)}-\mu_k)^2}$$

$$= 1+\frac{\sum_{k=j+1}^{N} \xi_k^2\chi_i^2(\mu_k)(\mu_1^{(i)}-\mu_k)^2}{\sum_{k=1}^{j} \xi_k^2\chi_i^2(\mu_k)(\mu_1^{(i)}-\mu_k)^2} \geq 1+\frac{\sum_{k=j+1}^{N} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}{\sum_{k=1}^{j} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}\ \frac{(\mu_1^{(i)}-\mu_{j+1})^2/\mu_{j+1}^2}{(\mu_1^{(i)}-\mu_j)^2/\mu_j^2}$$

$$\geq 1+\frac{\sum_{k=j+1}^{N} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}{\sum_{k=1}^{j} \xi_k^2\chi_i^2(\mu_k)\mu_k^2} = \left\{\sum_{k=1}^{j} \delta_k^{(i)}\right\}^{-1}$$

(the assumption has been used in both inequalities), so that $\sum_{k=1}^{j}\delta_k^{(i)} \geq \sum_{k=1}^{j}\gamma_k^{(i)}$. From lemma 5 it follows, that

$$\sum_{j=1}^{N} \delta_j^{(i)}\mu_j \leq \sum_{j=1}^{N} \gamma_j^{(i)}\mu_j. \tag{4.9}$$

According to the Cauchy-Schwarz inequality it holds that

$$\left\{\sum_{j=1}^{N} \gamma_j^{(i)}\mu_j\right\}^2 \leq \sum_{j=1}^{N} \gamma_j^{(i)}\mu_j^2 \sum_{k=1}^{N} \gamma_k^{(i)} = \sum_{j=1}^{N} \gamma_j^{(i)}\mu_j^2.$$

The rest of the proof follows from (4.5a) and (4.8a). ■

Note that the assumption in theorem 6 is equivalent to (4.6).

To illustrate theorem 6 the results of a numerical experiment are presented, where the matrix $A \in \mathbb{R}^{900\times 900}$ is equal to the five point finite difference discretized Laplace operator over a square region with gridspacing 1 and Dirichlet boundary conditions.

The system of linear equations $Ax = b$ is solved according to algorithm 3. We take $b = Ax$, where $x = (1,\ldots,1)^T \in \mathbb{R}^{900}$, and $x_0 = 0$. The results obtained are displayed in figure 1 $(\mu_1^{(i)} \approx 0.021$ if $i \geq 26)$.

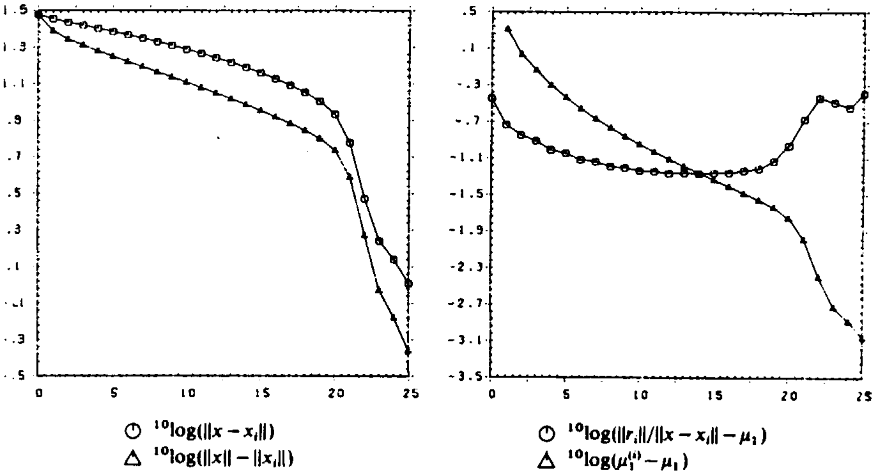


Fig. 1. Experimental results.

From the results shown in figure 1 it is clear that the Euclidean norms of the errors $x-x_i$ form a decreasing sequence (for a proof see [4], theorem 6:3) and the Euclidean norms of the approximations x_i form an increasing sequence (for a proof see the following theorem).

Theorem 7: *The Euclidean norms of the differences $x_i - x_0$ form an increasing sequence, i.e.*

$$0 = \|x_0 - x_0\| < \|x_1 - x_0\| < \ldots < \|x_N - x_0\| = \|x - x_0\|.$$

Proof: For $i = 0(1)N-1$ it holds that $x_i - x_0 = \sum_{j=0}^{i-1} \alpha_j p_j$, where $\alpha_j > 0$. Therefore $(x_i - x_0)^T p_i = \sum_{j=0}^{i-1} \alpha_j p_j^T p_i > 0$, because $p_j^T p_i > 0$ (see [4], theorem 5:3). From this it follows, that

$$\begin{aligned}(x_{i+1}-x_0)^T(x_{i+1}-x_0) &= (x_i - x_0 + \alpha_i p_i)^T(x_i - x_0 + \alpha_i p_i)\\ &= (x_i-x_0)^T(x_i-x_0) + 2\alpha_i(x_i-x_0)^T p_i + \alpha_i^2 p_i^T p_i\\ &> (x_i-x_0)^T(x_i-x_0). \qquad \blacksquare\end{aligned}$$

If $x_0 = 0$, then it follows, that the quotients $\|x-x_i\|/\|x_i\|$ form a decreasing sequence. In this case it follows from theorem 7, that the termination criterion (4.1) can be replaced by

$$\|r_i\| \leq \mu_1^{(i)}\|x_i\|\varepsilon.$$

Note that in some situations it is desirable to compute an approximation x_i to x with maximal machine-precision. The properties mentioned above give the possibility to stop on a criterion based on monotony.

It is well-known, that

$$\mu_1 = 2\left(2-2\cos\frac{\pi}{31}\right) \approx 0.021, \mu_2 = 2\left(2-\cos\frac{\pi}{31}-\cos\frac{3\pi}{31}\right) \approx 0.102,$$

so that $2\mu_1\mu_2/(\mu_1+\mu_2) \approx 0.034$.

According to theorem 6 condition (4.3) holds if $i \geq 21$. From figure 1 it follows that (4.3) holds if $i \geq 15$. The theoretical and the experimental results agree, but the discrepancy is rather large. Note, however, that for $i = 21$ the error reduction in x_i is very moderate, i.e. the termination criterion is reliable long before real convergence of the *cg*-method occurs (see [6]).

To understand the discrepancy, note that theorem 6 holds independently of the values of the weights ξ_j. If $\xi_j \to 0$ for $j = 3(1)N$, then

$$\sum_{j=1}^{N} \delta_j^{(i)}\mu_j \to \sum_{j=1}^{2} \delta_j^{(i)}\mu_j,$$

$$\sum_{j=1}^{N} \gamma_j^{(i)}\mu_j \to \sum_{j=1}^{2} \gamma_j^{(i)}\mu_j.$$

In this limit (4.9) passes into

$$\sum_{j=1}^{2} \delta_j^{(i)}\mu_j \leq \sum_{j=1}^{2} \gamma_j^{(i)}\mu_j.$$

This inequality is equivalent to (4.6). Hence (4.6) is necessary for (4.9) to hold for every initial error $x-x_0$ (defining the ξ_j).

To derive a necessary condition so that (4.3) holds for every initial error $x-x_0$, it must first be noted that according to the Cauchy-Schwarz inequality:

$$\text{(4.10)} \qquad \|r_i\|/\|x-x_i\| \leq \|r_i\|^2/\|x-x_i\|_A^2.$$

Since $\|r_i\|^2/\|x-x_i\|_A^2 = \sum_{j=1}^{N} \varepsilon_j^{(i)}\mu_j$ for $i = 0(1)N-1$, where

$$\varepsilon_j^{(i)} = \frac{\xi_j^2\varphi_i^2(\mu_j)\mu_j}{\sum_{k=1}^{N} \xi_k^2\varphi_i^2(\mu_k)\mu_k} \quad \text{for } j = 1(1)N,$$

the inequality $\mu_1^{(i)} \leq \|r_i\|^2/\|x-x_i\|_A^2$ passes in the limit $\xi_j \to 0$ for $j = 3(1)N$ into

$$\sum_{j=1}^{2} \delta_j^{(i)}\mu_j \leq \sum_{j=1}^{2} \varepsilon_j^{(i)}\mu_j.$$

This inequality is equivalent to

$$0 < \mu_1^{(i)} \leq (\mu_1\mu_2)^{1/2}. \tag{4.11}$$

From this and (4.10) it follows that (4.11) is necessary in order for (4.3) to hold for every initial error $x-x_0$. Thus it is impossible to replace condition (4.6) in theorem 6 by a condition that is stronger than (4.11) (for general weights ξ_j).

For a further discussion of the result in figure 1, we note that $(\mu_1\mu_2)^{1/2} \approx 0.046$. The condition (4.11) corresponds to $i \geq 19$. To understand the remaining discrepancy between the theoretical and experimental results, note that the weights ξ_j for $j = 3(1)N$ realize a weaker condition for $\mu_1^{(i)}$.

In view of the results in [7], we also analyse the reliability of (4.1) in the case of an almost double smallest eigenvalue. In this case $\mu_1^{(i)}$ initially converges to a point between the close eigenvalues, so that condition (4.6) might be unrealistic. A weaker conditon is given in

THEOREM 8: *If* $0 < (\mu_1^{(i)}-\mu_1)/\mu_1 \leq (\mu_3-\mu_1^{(i)})/\mu_3$, *then*

$$\{\mu_1^{(i)}\}^2 \leq (\mu_2^2-\mu_1^2)+\|r_i\|^2/\|x-x_i\|^2. \tag{4.12}$$

PROOF. If $0 < (\mu_1^{(i)}-\mu_1)/\mu_1 \leq (\mu_2-\mu_1^{(i)})/\mu_2$, then (4.12) follows directly from theorem 6. Therefore it is assumed that

$$|\mu_2-\mu_1^{(i)}|/\mu_2 < (\mu_1^{(i)}-\mu_1)/\mu_1 \leq (\mu_3-\mu_1^{(i)})/\mu_3.$$

Note that the first inequality holds if $\mu_2 \leq \mu_1^{(i)}$.

According to the Cauchy-Schwartz inequality it holds that

$$\left\{\sum_{j=1}^{N} \delta_j^{(i)}\mu_j\right\}^2 \leq \sum_{j=1}^{N} \delta_j^{(i)}\mu_j^2 \sum_{k=1}^{N} \delta_k^{(i)} = \sum_{j=1}^{N} \delta_j^{(i)}\mu_j^2.$$

For $j = 3(1)N$ it holds that

$$\left\{\sum_{k=1}^{j} \gamma_k^{(i)}\right\}^{-1} \geq 1+\frac{\sum_{k=j+1}^{N} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}{\sum_{k=1}^{j} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}\,\frac{(\mu_1^{(i)}-\mu_{j+1})^2/\mu_{j+1}^2}{(\mu_1^{(i)}-\mu_j)^2/\mu_j^2} \geq \left\{\sum_{k=1}^{j} \delta_k^{(i)}\right\}^{-1}.$$

Further,

$$\left\{\sum_{k=1}^{2} \gamma_k^{(i)}\right\}^{-1} \geq 1+\frac{\sum\limits_{k=3}^{N} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}{\sum\limits_{k=1}^{2} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}\ \frac{(\mu_1^{(i)}-\mu_3)^2/\mu_3^2}{(\mu_1^{(i)}-\mu_1)^2/\mu_1^2} \geq \left\{\sum_{k=1}^{2} \delta_k^{(i)}\right\}^{-1}$$

and

$$\{\gamma_1^{(i)}\}^{-1} \geq 1+\frac{\sum\limits_{k=3}^{N} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}{\xi_1^2\chi_i^2(\mu_1)\mu_1^2}\ \frac{(\mu_1^{(i)}-\mu_3)^2/\mu_3^2}{(\mu_1^{(i)}-\mu_1)^2/\mu_1^2} \leq F\{\delta_1^{(i)}\}^{-1}, \quad \text{where}$$

$$F = \frac{\sum\limits_{k\neq 2} \xi_k^2\chi_i^2(\mu_k)\mu_k^2}{\sum\limits_{k=1}^{N} \xi_k^2\chi_i^2(\mu_k)\mu_k^2} = 1-\delta_2^{(i)} < 1.$$

Since $\sum_{k=1}^{j} \delta_k^{(i)} \geq \sum_{k=1}^{j} \gamma_k^{(i)}$ for $j = 2(1)N$ and $\delta_1^{(i)} \geq F\gamma_1^{(i)}$, it holds that

$$\begin{aligned}\sum_{k=1}^{N} \delta_k^{(i)}\mu_k^2 &= \gamma_1^{(i)}\mu_1^2+(\delta_1^{(i)}-\gamma_1^{(i)})\mu_1^2+\sum_{k=2}^{N}\delta_k^{(i)}\mu_k^2\\ &\leq \gamma_1^{(i)}\delta_2^{(i)}(\mu_2^2-\mu_1^2)+\sum_{k=1}^{2}\gamma_k^{(i)}\mu_k^2+\left\{\sum_{k=1}^{2}(\delta_k^{(i)}-\gamma_k^{(i)})\right\}\mu_2^2+\sum_{k=3}^{N}\delta_k^{(i)}\mu_k^2\\ &\leq \gamma_1^{(i)}\delta_2^{(i)}(\mu_2^2-\mu_1^2)+\sum_{k=1}^{N}\gamma_k^{(i)}\mu_k^2 \leq (\mu_2^2-\mu_1^2)+\sum_{k=1}^{N}\gamma_k^{(i)}\mu_k^2\end{aligned}$$

(see the proof of lemma 4). The rest of the proof follows from (4.5a) and (4.8a). ■

Note that (4.3) is approximately satisfied if $\mu_1 \approx \mu_2$, the difference being of no importance in practical situations.

To illustrate theorem 8, the results of a numerical experiment are presented, where $A = \mathrm{diag}(\mu_1,\ldots,\mu_{900})$, and

$$\mu_1 = 0.034, \quad \mu_2 = 0.0341, \quad \mu_3 = 0.082,$$

$$\mu_4 = 0.127, \quad \mu_5 = 0.155, \quad \mu_6 = 0.190,$$

$\mu_7, \mu_8, \ldots, \mu_{900}$ uniformly in [0.2, 1.2].

Let $b = 0$ and $x_0 = (\xi_1,\ldots,\xi_{900})^T$, where $\xi_j = \mu_j^{-1}$. This experiment was used in [7]. The diagonal matrix A is inspired by the spectrum of the preconditioned discretized Laplace operator on a 30 × 30 grid. The results obtained are displayed in figure 2 ($\mu_1^{(i)} \approx 0.0340000$ if $i \geq 36$).

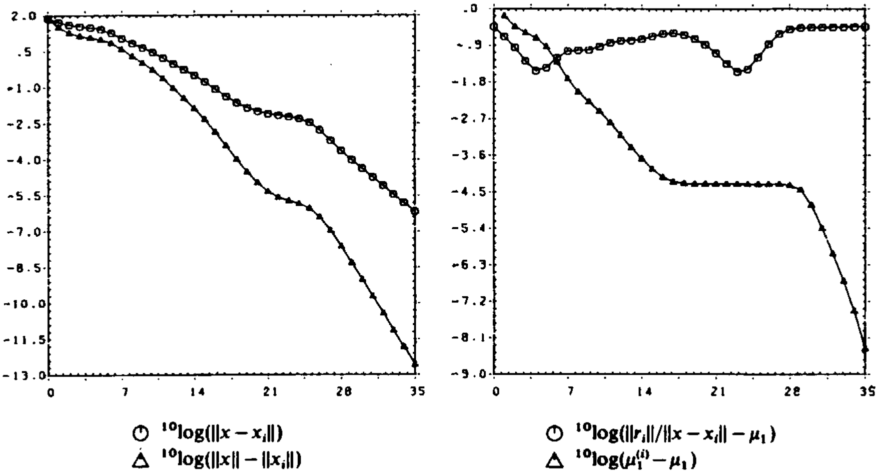


Fig. 2. Experimental results.

Let us now discuss the results in figure 2. We know that

$$2\mu_1\mu_2/(\mu_1+\mu_2) \approx 0.0340499, 2\mu_1\mu_3/(\mu_1+\mu_3) \approx 0.0480689.$$

According to theorem 6 condition (4.3) holds if $i \geq 21$. According to theorem 8 condition (4.12) holds with the small factor $\mu_2^2-\mu_1^2 = 10^{-6}$ if $i \geq 8$. From figure 2 it follows that (4.3) holds if $i \geq 6$.

If dealing with a cluster of eigenvalues $\mu_1, \ldots, \mu_k$ for some $1 < k < N$, a useful condition is given in

THEOREM 9: *If* $0 < (\mu_1^{(i)}-\mu_1)/\mu_1 \leq (\mu_{k+1}-\mu_1^{(i)})/\mu_{k+1}$, *then*

$$\{\mu_1^{(i)}\}^2 \leq (\mu_k^2-\mu_1^2)+\|r_i\|^2/\|x-x_i\|^2.$$

PROOF. Along the same lines as the proof of theorem 8. ■

Hence it can be concluded that the termination criterion (3.1) is very reliable, provided that the conjugate gradient process is not stopped in a too early phase, i.e. ε is not too large, and $\mu_1^{(i)}$ is approximated reasonably, i.e. u is not too large.

Acknowledgements.

I wish to thank Henk van der Vorst and Peter Sonneveld for their criticism and inspiring ideas.

REFERENCES

1. G. H. Golub and C. F. Van Loan, *Matrix Computations*, North Oxford Academic, Oxford (1983).
2. L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York (1981).
3. M. R. Hestenes, *Conjugate Direction Methods in Optimization*, Springer, Berlin (1980).
4. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 409–436.
5. B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, New Jersey (1980).
6. A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numerische Mathematik, 48 (1986), pp. 543–560.
7. A. van der Sluis and H. A. van der Vorst, *The convergence behaviour of Ritz values in the presence of close eigenvalues*, Linear Algebra and its Applications, 88/89 (1987), pp. 651–694.

CHAPTER II

Preconditioned conjugate gradients for solving singular systems

E.F. KAASSCHIETER
TNO, DGV Institute of Applied Geoscience, 2600 AG Delft, The Netherlands

Received 8 February 1988
Revised 18 June 1988

Abstract: In this paper the preconditioned conjugate gradient method is used to solve the system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is a singular symmetric positive semi-definite matrix. The method diverges if $\boldsymbol{b}$ is not exactly in the range $R(\boldsymbol{A})$ of $\boldsymbol{A}$. If the null space $N(\boldsymbol{A})$ of $\boldsymbol{A}$ is explicitly known, then this divergence can be avoided by subtracting from $\boldsymbol{b}$ its orthogonal projection onto $N(\boldsymbol{A})$.

As well as analysing this subtraction, conditions necessary for the existence of a nonsingular incomplete Cholesky decomposition are given. Finally, the theory is applied to the discretized semi-definite Neumann problem.

Keywords: Preconditioned conjugate gradients, symmetric positive semi-definite matrices, incomplete Cholesky decomposition, semi-definite Neumann problem.

1. Introduction

In this paper the system of linear equations

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{1.1}$$

is considered, where $\boldsymbol{A}$ is a symmetric positive semi-definite matrix. Two cases can be distinguished: the case where $\boldsymbol{A}$ is nonsingular and consequently positive definite, and the case where $\boldsymbol{A}$ is singular.

Much is known about the first case (see e.g. [5]). If $\boldsymbol{A}$ is a large and sparse matrix, then iterative methods for the approximate solution of (1.1) are often to be preferred over direct methods, because iterative methods help to reduce both memory requirements and computing time. The conjugate gradient method is a successful iterative method (see [5, section 10.2] and [8]).

The convergence rate of the conjugate gradient method is determined by the spectrum of eigenvalues of the matrix $\boldsymbol{A}$ (see [8]). An acceleration of the convergence rate can often be achieved by replacing the system (1.1) by the preconditioned system

$$\boldsymbol{M}^{-1}\boldsymbol{Ax} = \boldsymbol{M}^{-1}\boldsymbol{b}. \tag{1.2}$$

The symmetric positive definite matrix $\boldsymbol{M}$ must be chosen in such a way that the system $\boldsymbol{Mz} = \boldsymbol{r}$ can be solved with less computational work than the original system (1.1) for every vector $\boldsymbol{r}$ on the right-hand side of the equation, and so that the matrix $\boldsymbol{M}^{-1}\boldsymbol{A}$ has a more 'favourable' spectrum of eigenvalues than $\boldsymbol{A}$.

Numerical experiments indicate that in many situations the construction of the preconditioning matrix M by a suitable incomplete Cholesky decomposition of A is a good choice (see [6,7]). If A is a symmetric M-matrix, then every incomplete Cholesky decomposition exists. However, this condition is not necessary.

If A is singular, then the system (1.1) has a solution if, and only if, b is in the range $R(A)$ of A. In that case the solution is not unique. Nevertheless, a solution can be determined by the preconditioned conjugate gradient method, because only those eigenvalues and eigenvectors of $M^{-1}A$ that are represented in the right-hand side of (1.2) participate in the conjugate gradient process (see e.g. [8, section 2.2]).

However, the method diverges if $b \notin R(A)$, e.g. as a result of perturbation of domain errors. This divergence can usually be avoided by eliminating the singularity of A, i.e. by fixing some entries of the solution x (as many as the dimension of the null space $N(A) = R(A)^{\perp}$ of A), deleting the corresponding rows and columns of A, adjusting the right-hand side and solving the resulting system $\hat{A}\hat{x} = \hat{b}$ by the preconditioned conjugate gradient method.

If $N(A)$ is explicitly known, then there is another way of avoiding the divergence mentioned above. It is then obvious to subtract from b its orthogonal projection onto $N(A)$, thereby yielding the vector b_R, and to solve the adjacent $Ax = b_R$. In many situations this results in a faster convergence rate than when solving the nonsingular system $\hat{A}\hat{x} = \hat{b}$. This approach is discussed in Section 2.

The construction of an incomplete Cholesky decomposition of A may fail. Conditions for the existence of a nonsingular incomplete Cholesky decomposition of a symmetric positive semi-definite matrix are given in Section 3.

Finally, an important application, the discretized semi-definite Neumann problem, is dealt with in Section 4. The results are illustrated by a numerical experiment.

2. The preconditioned conjugate gradient method

Consider the system of linear equations

$$Ax = b. \tag{2.1}$$

where $A \in \mathbb{R}^{n \times n}$ is a singular symmetric positive semi-definite matrix and $b \in \mathbb{R}^n$.

The system (2.1) has a solution if, and only if, $b \in R(A)$ where $R(A) = \{\, y \in \mathbb{R}^n \mid y = Az$ for $z \in \mathbb{R}^n \}$ is the range of A. If the system (2.1) has a solution, then it is not unique. Indeed, Let $x \in \mathbb{R}^n$ be a solution of (2.1), then $\hat{x} = x + y$ is a solution for every $y \in N(A)$, where $N(A) = \{z \in \mathbb{R}^n \mid Az = 0\}$ is the null space of A (note that $N(A) = R(A)^{\perp}$).

Let $M \in \mathbb{R}^{n \times n}$ be a suitable symmetric positive definite preconditioning matrix; then the corresponding preconditioned conjugate gradient method (cg-method) (see e.g. [5, chapter 10]) generates a sequence x_1, $x_2, \ldots,$ starting with a vector $x_0 \in \mathbb{R}^n$, according to

Algorithm 1

```
$r_0 := b - Ax_0$
for $i = 0, 1, \ldots$
  $z_i := M^{-1} r_i$
  if $r_i = 0$ then stop
```

$$\begin{aligned}
&\beta_{i-1} := z_i^{\mathrm{T}} r_i / z_{i-1}^{\mathrm{T}} r_{i-1} \ (\beta_{-1} := 0)\\
&p_i := z_i + \beta_{i-1} p_{i-1} \ (p_0 := z_0)\\
&\alpha_i := z_i^{\mathrm{T}} r_i / p_i^{\mathrm{T}} A p_i\\
&x_{i+1} := x_i + \alpha_i p_i\\
&r_{i+1} := r_i - \alpha_i A p_i.
\end{aligned}$$

Since M is symmetric positive definite, there is a nonsingular matrix $C \in \mathbb{R}^{n\times n}$, such that $M = CC^T$. The preconditioned cg-method is equivalent to the ordinary cg-method for solving the preconditioned system

$$\tilde{A}\tilde{x} = \tilde{b}, \tag{2.2}$$

where $\tilde{A} = C^{-1}AC^{-\mathrm{T}}$, $\tilde{x} = C^{\mathrm{T}}x$ and $\tilde{b} = C^{-1}b$ (choose $\tilde{x}_0 = C^{\mathrm{T}}x_0$). For the analysis of the preconditioned cg-method we will occasionally switch between these two viewpoints.

Corresponding to $\tilde{r}_0 = C^{-1}r_0$ there are uniquely determined eigenvalues $0 = \mu_0 < \mu_1 < \dots < \mu_m$ and normalized eigenvectors $u_0, \dots, u_m$ of $\tilde{A}$, such that $\tilde{r}_0 = \sum_{j=0}^{m} \xi_j u_j$, where $\xi_0 \geqslant 0$ and $\xi_j > 0$ for $j = 1, \dots, m$ (see [8, section 2.2]). Note that $\xi_0 = 0$ if, and only if, $\tilde{r}_0 \in R(\tilde{A})$, i.e. $b \in R(A)$. These eigenvalues and eigenvectors are the active ones; in view of [8, section 2.1] the other eigenvalues and eigenvectors do not participate in the conjugate gradient process.

If $b \notin R(A)$, then (2.1) does not have an exact solution. In practice this situation may arise because of perturbation of domain errors (see [2]). Using the preconditioned cg-method we can then still generate a sequence $x_1, x_2, \dots$. However, from numerical experiments it appears that the Euclidean norm of the residual $\tilde{r}_i$ initially tends to decrease, but at some stage suddenly increases. It seems that the orthogonal projection of the vector $\tilde{x}_i$ onto $R(\tilde{A})$ converges to a certain vector $\tilde{x}$, before it suddenly diverges. Three questions arise:

- In what sense does $x = C^{-\mathrm{T}}\tilde{x}$ represent a solution?
- Can we understand the sudden divergence?
- How can we preclude this divergence?

The last question will be answered in this section; the first two will be discussed in Section 4.

If $b \notin R(A)$, then one often resorts to a least squares solution of (2.1) (which always exists), i.e. a vector x for which $\| b - Ax \|_2$ is minimal (see [5, section 6.1]). Since A is singular, there is an infinite number of least squares solutions. In this whole set of least squares solutions there is a unique vector x whose Euclidean norm is minimal. This is referred to as the minimum norm least squares solution of (2.1). Note that x is a least squares solution of (2.1) if, and only if, x is a solution of the projected system

$$Ax = b_R, \tag{2.3}$$

where b_R is the orthogonal projection of b onto $R(A)$.

If $R(A)$ is explicitly known, then we can prove that a solution x of (2.3) can be determined using the preconditioned cg-method. For the preconditioned starting residual $\tilde{r}_{0,R} = \tilde{b}_R - \tilde{A}\tilde{x}_0$, where $\tilde{b}_R = C^{-1}b_R$, we have $\tilde{r}_{0,R} = \sum_{j=1}^{m} \tilde{\xi}_j u_j$, where $\tilde{\xi}_j > 0$ for $j = 1, \dots, m$ (note that in general $\tilde{\xi}_j \neq \xi_j$, because of the non-orthogonality of the projection of $\tilde{r}_0$ onto $R(\tilde{A})$, resulting in $\tilde{r}_{0,R}$). From this it follows that the cg-method for solving the system

$$\tilde{A}\tilde{x} = \tilde{b}_R \tag{2.4}$$

generates a sequence $\tilde{x}_1, \tilde{x}_2, \dots,$ starting with a vector $\tilde{x}_0 = C^{\mathrm{T}}x_0$. This sequence has the following property

$$\|\tilde{x}-\tilde{x}_i\|_{\tilde{A}}=\min_{y-\tilde{x}_0\in K_i}\|\tilde{x}-y\|_{\tilde{A}}, \tag{2.5}$$

where $K_i=\text{span}\{\tilde{r}_{0,R}, \tilde{A}\tilde{r}_{0,R},\ldots,\tilde{A}^{i-1}\tilde{r}_{0,R}\}$ is the ith Krylov subspace of $\tilde{A}$ with respect to $\tilde{r}_{0,R}$ ($\|z\|_{\tilde{A}}=(z^T\tilde{A}z)^{1/2}$ for all $z\in\mathbb{R}^n$). For the basic relations of the cg-method, which also hold in this case, see e.g. [5, section 10.2] and [8, section 2]. Since $\|\cdot\|_{\tilde{A}}$ is a norm in $R(\tilde{A})$ and $K_i\subset R(\tilde{A})$ for $i=1, 2\ldots$, it follows from (2.5) that $\tilde{x}_i$ converges to a solution $\tilde{x}$ of (2.4), and thus $x_i=C^{-T}\tilde{x}_i$ converges to a solution $x=C^{-T}\tilde{x}$ of (2.3).

This solution is not necessarily the minimum norm solution of (2.3), i.e. the minimum norm least squares solution of (2.1). With the popular choice $x_0=\mathbf{0}$ (and thus $\tilde{x}_0=C^Tx_0=\mathbf{0}$) it follows from (2.5) that $\tilde{x}_i\in R(\tilde{A})$ for $=0, 1,\ldots$. Therefore $\tilde{x}_i$ converges to the minimum norm solution of (2.4) (note that $\tilde{x}$ is the minimum norm solution of (2.4) if, and only if, $\tilde{x}\in R(\tilde{A})$). An approximation to the minimum norm least squares solution of (2.1) can be determined by subtracting from $x_i=C^{-T}\tilde{x}_i$ its orthogonal projection onto $N(A)$.

3. Incomplete Cholesky decompositions

A symmetric positive definite preconditioning matrix $M=CC^T$, where C is a lower triangular matrix, may be determined by an incomplete Cholesky decomposition of the symmetric positive semi-definite matrix A (see [6], [7]). The most general form of an incomplete Cholesky decomposition is indicated in [7, section 1], where it is suggested that a Cholesky decomposition of A be made, during which elimination corrections are partly ignored in C in appropriate places. The ignoration factors will be given by a symmetric matrix $\Theta\in\mathbb{R}^{n\times n}$, where $0\leqslant\theta_{ij}\leqslant 1$ for $i, j=1,\ldots,n$. In this way we obtain

Algorithm 2

```
for $i=1,\ldots,n$
  for $j=1,\ldots,i-1$
      $c_{ij}:=\left(a_{ij}-\theta_{ij}\sum_{k=1}^{j-1}c_{ik}c_{jk}\right)/c_{jj}$
  $c_{ii}:=\left(a_{ii}-\theta_{ii}\sum_{k=1}^{i-1}c_{ik}^2\right)^{1/2}$
```

(we define $0/0=0$). If Algorithm 2 does not fail, i.e. if $a_{ii}-\theta_{ii}\Sigma_{k=1}^{i-1}c_{ik}^2\geqslant 0$ for $i=1,\ldots,n$ and $c_{jj}>0$ if $a_{ij}-\theta_{ij}\Sigma_{k=1}^{j-1}c_{ik}c_{jk}>0$, then we will denote by $C=C(A, \Theta)$ the lower triangular matrix C, constructed by the incomplete Cholesky decomposition of A with respect to the ignoration matrix Θ. Thus the matrix C constructed by the complete Cholesky decomposition of A, which exists if A is symmetric positive semi-definite (see [5, chapter 5]), will be denoted by $C=C(A, \mathbf{1})$ where every entry of $\mathbf{1}\in\mathbb{R}^{n\times n}$ is equal to 1. Note that $C(A, \Theta_1)=C(A, \Theta_2)$ might be possible for $\Theta_1\neq\Theta_2$. Henceforth we will say that $C(A, \Theta)$ exists for a matrix A and an ignoration matrix Θ, when Algorithm 2 is executable.

Note that the executability of Algorithm 2 implies that $c_{ii} \geqslant 0$ for $i = 1, \ldots, n$. However, to solve the system of linear equations $\boldsymbol{M}z = \boldsymbol{r}$ in each iteration of Algorithm 1 according to:

Algorithm 3

```
for $i = 1, \ldots, n$
    $z_i := \left( r_i - \sum_{j=1}^{i-1} c_{ij} z_j \right) / c_{ii}$
for $i = n, \ldots, 1$
    $z_i := \left( r_i - \sum_{j=i+1}^{n} c_{ij} z_j \right) / c_{ii},$
```

it is necessary that $c_{ii} > 0$ for $i = 1, \ldots, n$.

It has been proved that $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta})$ exists for every ignoration matrix $\boldsymbol{\Theta}$, if $\boldsymbol{A}$ is a symmetric $\boldsymbol{M}$-matrix (see [6, theorem 2.4; 7, section 1]). In this case, $c_{ii} > 0$ for $i = 1, \ldots, n$. $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an $\boldsymbol{M}$-matrix, if $a_{ij} \leqslant 0$ for all $i \neq j$, $\boldsymbol{A}$ is nonsingular and $\boldsymbol{A}^{-1} \geqslant 0$. Note that $\boldsymbol{A}$ is a symmetric $\boldsymbol{M}$-matrix if, and only if, $\boldsymbol{A}$ is a Stieltjes matrix, i.e. $a_{ij} \leqslant 0$ for all $i \neq j$ and $\boldsymbol{A}$ is symmetric positive definite (see [9, p.85]).

Before deriving a necessary condition for the existence of an incomplete Cholesky decomposition of a singular symmetric positive semi-definite matrix $\boldsymbol{A}$, a definition need to be given.

Definition 3.1. A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a singular Stieltjes matrix if $a_{ij} \leqslant 0$ for all $i \neq j$ and $\boldsymbol{A}$ is singular and symmetric positive semi-definite.

Theorem 3.2. *If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an irreducible singular Stieltjes matrix, then $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta})$ exists for every ignoration matrix $\boldsymbol{\Theta}$. In this case $c_{ii} > 0$ for $i = 1, \ldots, n$ if, and only if, $\boldsymbol{C} \neq \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{1})$.*

Proof. Let $\boldsymbol{\Theta} \in \mathbb{R}^{n \times n}$ be a certain ignoration matrix and consider the incomplete Cholesky decomposition of an irreducible singular Stieltjes matrix $\boldsymbol{A}$ with respect to $\boldsymbol{\Theta}$. Since the leading principal submatrix that is obtained from $\boldsymbol{A}$ by omitting the last row and column is a nonsingular Stieltjes matrix (see [4, section 5]) the first $n-1$ loops of Algorithm 2 are executable and $c_{ii} > 0$ for $i = 1, \ldots, n-1$.

Assume that Algorithm 2 is not executable, i.e. $a_{nn} - \theta_{nn} \Sigma_{k=1}^{n-1} c_{nk}^2 < 0$, and let $\boldsymbol{A}^{(\epsilon)} = \boldsymbol{A} + \epsilon e_n e_n^{\mathrm{T}}$, where $\epsilon > 0$ and $e_n = (0, \ldots, 0, 1)^{\mathrm{T}}$ is the nth unity vector, then $\boldsymbol{A}^{(\epsilon)}$ is a nonsingular Stieltjes matrix (see [4, (5, 11)]) and thus $\boldsymbol{C}(\boldsymbol{A}^{(\epsilon)}, \boldsymbol{\Theta})$ exists. If $\epsilon > 0$ is small enough we have

$$a_{nn}^{(\epsilon)} - \theta_{nn} \sum_{k=1}^{n-1} c_{nk}^2 = a_{nn} + \epsilon - \theta_{nn} \sum_{k=1}^{n-1} c_{nk}^2 < 0.$$

This gives a contradiction, thus Algorithm 2 is executable, i.e. $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta})$ exists.

($\Rightarrow$) Suppose that $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{1})$, then $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ (see [5, chapter 5]) and thus $\Pi_{i=1}^{n} c_{ii}^2 = (\det \boldsymbol{C})^2 = \det \boldsymbol{A} = 0$, i.e. $c_{nn} = 0$.

($\Leftarrow$) Suppose that $\boldsymbol{C} \neq \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{1})$.

Assume that

$$\max\left\{ i \mid c_{ii} > \left(a_{ii} - \sum_{k=1}^{i-1} c_{ik}^2 \right)^{1/2} \right\} \geqslant \max\left\{ i \mid c_{ij} > \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk} \right) / c_{jj} \text{ for some } j \right\}. \tag{3.1}$$

Define in this case

$$i_0 = \max\left\{ i \mid c_{ii} > \left(a_{ii} - \sum_{k=1}^{i-1} c_{ik}^2 \right)^{1/2} \right\},$$

i.e. i_0 agrees with the last partial ignoration in Algorithm 2.

If $i_0 = n$, then $c_{nn} > (a_{nn} - \Sigma_{k=1}^{n-1} c_{nk}^2)^{1/2} \geqslant 0$. Thus, assume that $i_0 < n$ and define $\boldsymbol{C}' = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta}')$, where

$$\theta'_{ij} = \begin{cases} 1 & \text{if } i = j = i_0, \\ \theta_{ij} & \text{otherwise.} \end{cases}$$

Since $\boldsymbol{A}$ is irreducible, there is a row of integers $\{i_s\}_{s=0}^r$, such that $i_r = n$ and $a_{i_{s-1} i_s} < 0$ for $s = 1, \dots, r$ (see [9, p.20]). A subrow $\{i'_\sigma\}_{\sigma=0}^{\rho}$ of the row $\{i_s\}_{s=0}^r$ exists, such that $i_0 = i'_0 < i'_1 < \cdots < i'_\rho = n$. By complete induction it follows that

$$\begin{aligned} c_{i'_\sigma i'_{\sigma-1}} &= \left(a_{i'_\sigma i'_{\sigma-1}} - \sum_{k=1}^{i'_{\sigma-1}-1} c_{i'_\sigma k} c_{i'_{\sigma-1} k} \right) \Big/ c_{i'_{\sigma-1} i'_{\sigma-1}} \\ &< \left(a_{i'_\sigma i'_{\sigma-1}} - \sum_{k=1}^{i'_{\sigma-1}-1} c'_{i'_\sigma k} c'_{i'_{\sigma-1} k} \right) \Big/ c'_{i'_{\sigma-1} i'_{\sigma-1}} = c'_{i'_\sigma i'_{\sigma-1}} \leqslant 0, \\ c_{i'_\sigma i'_\sigma} &= \left(a_{i'_\sigma i'_\sigma} - \sum_{k=1}^{i'_\sigma - 1} c_{i'_\sigma k}^2 \right)^{1/2} > \left(a_{i'_\sigma i'_\sigma} - \sum_{k=1}^{i'_\sigma - 1} (c'_{i'_\sigma k})^2 \right)^{1/2} = c'_{i'_\sigma i'_\sigma} \geqslant 0 \end{aligned}$$

for $\sigma = 1, \dots, \rho$.

Thus, in particular $c_{nn} > c'_{nn} \geqslant 0$.

Assume that (3.1) does not hold. Define in this case

$$i_0 = \max\left\{ i \mid c_{ij} > \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk} \right) \Big/ c_{jj} \text{ for some } j \right\},$$

i.e., i_0 agrees with the last loop in Algorithm 2, in which an elimination correction is partly ignored. Define $\boldsymbol{C}' = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta}')$, where

$$\theta'_{ij} = \begin{cases} 1 & \text{if } j < i = i_0, \\ \theta_{ij} & \text{otherwise.} \end{cases}$$

Now we have

$$c_{i_0 i_0} = \left(a_{i_0 i_0} - \sum_{k=1}^{i_0 - 1} c_{i_0 k}^2 \right)^{1/2} > \left(a_{i_0 i_0} - \sum_{k=1}^{i_0 - 1} (c'_{i_0 k})^2 \right)^{1/2} = c'_{i_0 i_0} \geqslant 0.$$

The rest of the proof is analogous. □

If a symmetric matrix $\boldsymbol{A}$ is reducible, then a permutation matrix $\boldsymbol{P}$ exists, such that

$$\tilde{\boldsymbol{A}} = \boldsymbol{P}^T \boldsymbol{A} \boldsymbol{P} = \begin{pmatrix} \tilde{\boldsymbol{A}}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{\boldsymbol{A}}_N \end{pmatrix}, \tag{3.2}$$

where every submatrix $\tilde{A}_i \in \mathbf{R}^{p_i \times p_i}$ is irreducible or equal to the 1×1 null matrix, $0 < p_i < n$ and $\sum_{i=1}^{N} p_i = n$. We say that (3.2) is the normal form of A (see [9, p.46]).

The normal form (3.2) is unique up to permutations in and of submatrices $\tilde{A}_i$. In the folllowing we choose P such that the rows and columns of every submatrix $\tilde{A}_i$ correspond to successive rows and columns of A. The normal form (3.2) is then unique up to permutations of submatrices $\tilde{A}_i$.

Define $\tilde{\Theta} = P^T \Theta P$. It follows from Algorithm 2, that $C = C(A, \Theta)$ exists if, and only if, $\tilde{C} = C(\tilde{A}, \tilde{\Theta})$ exists. In this case we have

$$\tilde{C} = P^T C P = \begin{pmatrix} \tilde{C}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{C}_N \end{pmatrix}, \tag{3.3}$$

where $\tilde{C}_i = C(\tilde{A}_i, \tilde{\Theta}_i)$ and $\tilde{\Theta}_i$ is the principal submatrix of $\tilde{\Theta}$ corresponding to $\tilde{A}_i$.

At this stage we can prove:

Theorem 3.3. *If $A \in \mathbf{R}^{n \times n}$ is a reducible singular Stieltjes matrix, then $C = C(A, \Theta)$ exists for every ignoration matrix Θ. Let $\tilde{A} = P^T A P$ be the normal form* (3.2) *of A, where the rows and columns of every submatrix $\tilde{A}_i$ correspond to successive rows and columns of A. In this case $c_{ii} > 0$ for $i = 1, \ldots, n$ if, and only if, $\tilde{C}_i \neq C(\tilde{A}_i, \mathbf{1}_i)$, where $\tilde{C}_i$ and $\mathbf{1}_i$ are the principal submatrices of $\tilde{C}$ and $\mathbf{1}$ corresponding to $\tilde{A}_i$.*

Proof. Follows directly from Theorem 3.2 and (3.3). □

4. The semi-definite Neumann problem

An important practical example of a system (2.1) is obtained after the discretization of the semi-definite Neumann boundary value problem

$$-\nabla \cdot (A \nabla u) = f \quad \text{in } \Omega, \qquad -n \cdot (A \nabla u) = g \text{ on } \partial\Omega, \tag{4.1}$$

where $\Omega \subset \mathbf{R}^d$ is an open, bounded and connected domain with a piecewise smooth boundary $\partial\Omega$. Further, let $A \in L_\infty(\Omega, \mathbf{R}^{d \times d})$, where $A(x)$ is symmetric positive definite for almost every $x \in \Omega$, and $f \in L_2(\Omega)$, $g \in L_2(\partial\Omega)$ satisfying the compatibility condition $\int_\Omega f \, dx = \int_{\partial\Omega} g \, ds$ (see [3, section 1.2]).

The discretization of (4.1) by a suitable finite difference or finite element method, leads to a system (2.1), where A is a singular Stieltjes matrix (for details see [1], [9]). If the discretization grid is connected (see [9, p. 20]), then A is irreducible. Note that $N(A) = \text{span}\{e\}$, where $e = (1, \ldots, 1)^T$, because the solution u of (4.1) is unique up to a constant factor. As a result of perturbation of domain errors (Ω is approximated by a polygon $\tilde{\Omega}$) the system (2.1) may not have a solution, i.e. $b \notin R(A)$ (see [2], where b is projected onto $R(A)$ to overcome this problem).

As an illustration we take the Laplace equation on $\Omega = (0,1)^2$ with Neumann boundary conditions:

$$-\Delta u = 0 \quad \text{in } \Omega, \qquad -\partial u / \partial n = g \text{ on } \partial\Omega, \tag{4.2}$$

with g such that $u(x) = x_1 + x_2 - 1$ for $x = (x_1, x_2)^T \in \Omega$ (it then follows that $\int_{\partial\Omega} g \, ds = 0$). We

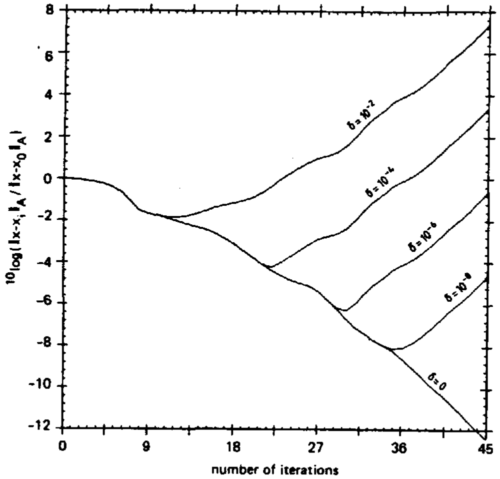


Fig. 1. Experimental results for different perturbations (with $x_{i+1} := x_i + \alpha_i p_i$).

choose a five-point finite difference discretization with step sizes of 1/29. This results in an irreducible singular Stieltjes matrix $A \in \mathbb{R}^{900\times 900}$ (see [9, section 6.3]). The resulting system (2.1) is solved by the preconditioned cg-method with the preconditioning matrix $M = CC^{\mathrm{T}}$, where the lower triangular matrix C is constructed by the incomplete Cholesky decomposition of A with respect to the ignoration matrix $\Theta \in \mathbb{R}^{n\times n}$, where

$$\theta_{ij} = \begin{cases} 1 & \text{if } a_{ij} \neq 0, \\ 0 & \text{if } a_{ij} = 0, \end{cases} \tag{4.3}$$

(see Section 3). This is the so-called ICCG(1,1) preconditioning (see [7, section 2.1.2]). From Theorem 3.2 it follows that $C = C(A, \Theta)$ exists and $c_{ii} > 0$ for $i = 1, \dots, n$. In Algorithm 2 we choose the starting vector $x_0 = \mathbf{0}$.

To simulate a perturbation of the right-hand side we choose the vector $b_R = Ax$ as an unperturbed right-hand side, where $x \in R(A)$ corresponds to the solution of (4.2) ($b_R \in R(A)$). Next to this system we consider the perturbed systems $Ax = b$, where $b = b_R + \gamma e$ and $\gamma = \|b_R\|_2\delta/\sqrt{n(1-\delta^2)}$ for $0 < \delta < 1$. Note that $b_R = b - (b^{\mathrm{T}}e/n)e$ is the orthogonal projection of b onto $R(A)$ (see Section 2). A good measure for the perturbation of a system $Ax = b$ is the angle θ between b and $R(A)$. We find

$$\sin\theta = \|b - b_R\|_2/\|b\|_2 = \gamma\sqrt{n}/\|b\|_2 = \delta. \tag{4.4}$$

The preconditioned cg-method for solving the unperturbed system $Ax = b_R$, i.e. $\delta = 0$, converges monotonically (see Fig. 1). The preconditioned cg-method for solving a perturbed system $Ax = b$, i.e. $0 < \delta < 1$, initially seems to converge monotonically to the minimum norm solution of the unperturbed system $Ax = b_R$, but then suddenly starts to diverge (see Fig. 1 for $\delta = 10^{-2}$, 10^{-4}, 10^{-6}, 10^{-8}). The smaller $\delta > 0$, the longer it takes before the preconditioned cg-method starts to diverge. Two questions remain:

- In what sense does the preconditioned cg-method for solving a perturbed system $Ax = b$ initially converge to a solution of the unperturbed system $Ax = b_R$?
- Can we understand the sudden divergence of the preconditioned cg-method for solving the perturbed system $Ax = b$?

In order to answer the first question, note that the results in Fig. 1 are not influenced by the component of $\tilde{x}_i$ orthogonal to $R(\tilde{A})$. Thus the preconditioned cg-method for solving a perturbed system $Ax = b$ and the unperturbed system $Ax = b_R$ would generate the same results, if the constants α_i and β_i were equal in both cases. However, since $r_i \neq A(x - x_i)$ in the perturbed case, these constants are not equal in both cases. If $\delta > 0$ is small, then initially roughly the same constants α_i and β_i are computed in both cases and thus roughly the same results are generated, i.e. the orthogonal projections onto $R(\tilde{A})$ of the approximations $\tilde{x}_i$ generated by the cg-method for solving a perturbed system $\tilde{A}\tilde{x} = \tilde{b}$ converges initially to a solution of the unperturbed system $\tilde{A}\tilde{x} = \tilde{b}_R$.

In order to answer the second question note that the constants α_i in Algorithm 2 are chosen in accordance with the property

$$\| x - x_{i+1} \|_A = \| x - x_i - \alpha_i p_i \|_A = \min_{\alpha \in \mathbf{R}} \| x - x_i - \alpha p_i \|_A, \tag{4.5}$$

at least if $b \in R(A)$ (see [5, section 10.3]). If $b \notin R(A)$, then (4.5) is not true and can be replaced by

$$\| x - x_i - \hat{\alpha}_i p_i \|_A = \min_{\alpha \in \mathbf{R}} \| x - x_i - \alpha p_i \|_A, \tag{4.6}$$

where $\hat{\alpha}_i = p_i^{\mathrm{T}}A(x - x_i)/p_i^{\mathrm{T}}Ap_i$. Since $z_i^{\mathrm{T}}r_i = p_i^{\mathrm{T}}r_i \neq p_i^{\mathrm{T}}A(x - x_i)$ in the perturbed case, we have $\alpha_i \neq \hat{\alpha}_i$. If $\delta > 0$ is small, then initially $0 < \alpha_i < 2\hat{\alpha}_i$ and thus it follows from (4.6) that $\| x - x_{i+1} \|_A < \| x - x_i \|_A$, i.e. the sequence $\| x - x_1 \|_A, \| x - x_2 \|_A, \ldots$ converges, though not optimally. If the cg-process is perturbed too much, then $\alpha_i < 0$ or $\alpha_i > 2\hat{\alpha}_i$ and $\| x - x_i \|_A$ starts diverging.

If the computation of x_{i+1} in Algorithm 2 is replaced by $x_{i+1} := x_i + \hat{\alpha}_i p_i$ (note that $\tilde{A}(\tilde{x} - \tilde{x}_i)$ is the orthogonal projection of $\tilde{r}_i$ onto $R(\tilde{A})$, thus $\hat{\alpha}_i$ can be computed without knowing the solution x of (2.3)), then the sequence $\| x - x_1 \|_A, \| x - x_2 \|_A, \ldots$ converges (see Fig. 2). The sudden divergence in the perturbed case is replaced by a stagnation of $\| x - x_i \|_A$. This stagnation can be explained by realizing that (2.5) is not true, if $b \notin R(A)$. This is not

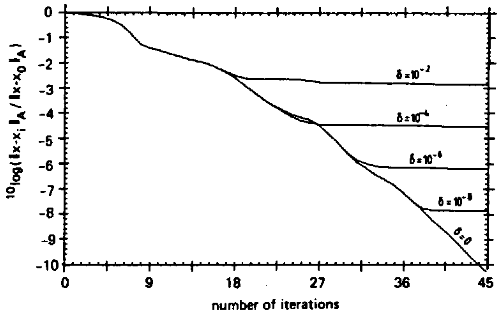


Fig. 2. Experimental results for different perturbations (with $x_{i+1} := x_i + \hat{\alpha}_i p_i$).

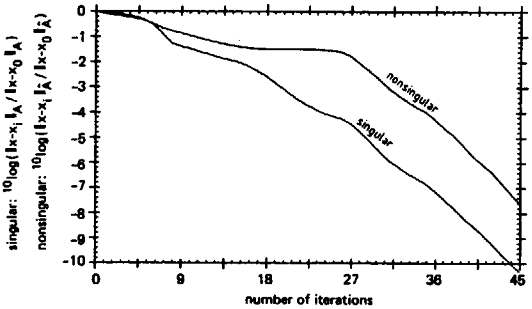


Fig. 3. Experimental results for the singular and nonsingular case.

caused by a loss of orthogonality (because from Algorithm 1 it follows that $z_i^{\mathrm{T}}r_j = 0$ and $p_i^{\mathrm{T}}Ap_j = 0$ if $i \neq j$) but is the result of $r_i \neq A(x - x_i)$.

To get rid of the stagnation of $\| x - x_i \|_A$ it suffices to project b on $R(A)$, resulting in the vector $b_R = b - (b^{\mathrm{T}}e/n)e$, and to solve the adjacent system $Ax = b_R$, resulting in a least squares solution of the perturbed system $Ax = b$ (see Section 2). Note that the convergence of the preconditioned cg-method for solving the projected system can be disturbed by rounding errors, if the matrix A is ill conditioned. In this case it may be advisable to project $\tilde{x}_i$ and $\tilde{r}_i$ on $R(\tilde{A})$ repeatedly, which is not a very expenasive process by itself.

In conclusion, note that the classic approach for eliminating the singularity of the matrix A is to fix an entry in the solution x, to delete the corresponding row and column of A, to adjust the right-hand side and to solve the resulting system $\hat{A}\hat{x} = \hat{b}$. Though the matrix $\hat{A}$ is nonsingular, the convergence rate of the precondition cg-method appears to be slower than in the nonsingular case (see Fig. 3 for the results of the experiment, where $x(900)$, which corresponds to the value $u(1,1)$ of the solution u of (4.2), is fixed). This experiment motivated the use of the preconditioned cg-method for the original singular system itself, as is described in this paper.

Acknowledgement

I wish to thank Henk van der Vorst for his criticisms and inspiring ideas.

References

[1] O. Axelsson and V.A. Barker, *Finite Element Solution of Boundary Value Problems* (Academic Press, New York, 1984).

[2] J.W. Barrett and C.M. Elliott, A practical finite element approximation of a semi-definite Neumann problem on a curved domain, *Numer. Math.* **51** (1987) 23–36.

[3] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems* (North-Holland, Amsterdam, 1978).

[4] M. Fiedler and V. Pták, On matrices with non-positive off-diagonal elements and positive principal minors, *Czechoslovakian Math. J.* **12** (1962) 382–400.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computations* (North Oxford Academic, Oxford, 1983).

[6] J.A. Meijerink and H.A. Van der Vorst, An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, *Math. Comput.* **31** (1977) 148–162.

[7] J.A. Meijerink and H.A. Van der Vorst, Guidelines for the usage of incomplete decompositions in solving sets of linear systems as they occur in practical problems, *J. of Comput. Phys.* **44** (1981) 134–155.

[8] A. Van der Sluis and H.A. Van der Vorst, The rate of convergence of conjugate gradients, *Numer. Math.* **48** (1986) 543–560.

[9] R.S. Varga, *Matrix Iterative Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1962).

CHAPTER III

A GENERAL FINITE ELEMENT PRECONDITIONING FOR THE CONJUGATE GRADIENT METHOD

E.F. KAASSCHIETER

TNO Institute of Applied Geoscience, P.O. Box 285,
2600 AG Delft, The Netherlands

Abstract.

Discretizing a symmetric elliptic boundary value problem by a finite element method results in a system of linear equations with a symmetric positive definite coefficient matrix. This system can be solved iteratively by a preconditioned conjugate gradient method. In this paper a preconditioning matrix is proposed that can be constructed for all finite element methods if a mild condition for the node numbering is fulfilled. Such a numbering can be constructed using a variant of the reverse Cuthill-McKee algorithm.

AMS(MOS) subject classifications: 65 F 10, 65 F 50, 65 N 30.

Keywords: Elliptic boundary value problems, finite element methods, preconditioned conjugate gradients, reverse Cuthill-McKee algorithm.

1. Introduction.

Discretizing a symmetric elliptic boundary value problem by a finite element method results in a system of linear equations with a symmetric positive definite coefficient matrix. This system can be solved iteratively by a preconditioned conjugate gradient method. Nowadays many methods for constructing efficient preconditioning matrices are known (for surveys see [1] and [5]).

However, finite difference methods have been the main inspiration for the preconditioning methods invented up to now (for interesting exceptions see [9], [13] and [19]). As a result, existence theorems have been inspired by coefficient matrices which arise from discretizing elliptic boundary value problems by finite difference techniques. For example, the preconditioning matrices constructed by incomplete Cholesky decompositions or modified incomplete Cholesky decompositions exist if the coefficient matrix is a symmetric M-matrix (see [17: theorem 2.4]) or a symmetric weakly diagonally dominant M-matrix (see [8: theorem 3.1]), respectively. This condition is fulfilled if central second-order finite differences are used. For finite element methods this condition results in serious restrictions to the shape of the element subdomains, e.g. if linear triangles are used, all vertex angles have to be nonobtuse. Using quadratic triangles results in a coefficient matrix that is never an M-matrix. Nevertheless, numerical experiments suggest that incomplete Cholesky decompositions and modified incomplete Cholesky decompositions precondition quite well if the finite element mesh is not too irregular.

This paper introduces a preconditioning matrix that can be constructed for all finite element methods if a mild condition for the global node numbering is fulfilled. More explicitly, there may be no maximal global node numbers. A global node has a maximal number if all of its neighbours have a lower number and this node and its neighbours are not on the Dirichlet-type portion of the boundary. Such a numbering can be constructed using the reverse Cuthill-McKee algorithm (see [16]) starting with a node on the Dirichlet-type portion of the boundary.

The main idea of this preconditioning method is to write the contribution of each finite element to the global matrix as a product of a lower triangular, a diagonal and an upper triangular matrix. These various matrices are then assembled to form a global lower triangular, a global diagonal and a global upper triangular matrix. The preconditioning matrix is defined as the product of these global matrices. The idea can be generalized by assembling the sum of the factors of contributions of several finite elements to the global matrix.

In section 2 the finite element method is briefly reviewed, concentrating on the construction of finite element spaces by partitioning the domain, interpolating locally in each subdomain and assembling the local basis functions.

The preconditioned conjugate gradient method, the assembly of the coefficient matrix as the sum of contributions of each finite element and some well-known preconditioning matrices are discussed in section 3. Sufficient conditions for the existence of these preconditioning matrices are mentioned.

In section 4 the finite element preconditioning matrix is defined. A necessary and sufficient condition for the existence of such a preconditioning matrix is given. A variant of the reverse Cuthill-McKee algorithm to fulfil this condition is discussed. It is proved that the smallest eigenvalue of the preconditioned matrix is equal to 1. This property can be used to verify the correctness of a computer implementation of the finite element preconditioning, using the fact that the smallest Ritz value of the preconditioned matrix converges monotonically to its smallest active eigenvalue during the conjugate gradient process (see, e.g., [21: section 2.3]). A generalization of the finite element preconditioning is also discussed.

In section 5 the Neumann boundary value problem, where the solution is only unique up to a constant value, is considered. The consequences of this nonuniqueness for the conjugate gradient method and the finite element preconditioning are discussed.

In section 6 the performance of the preconditioned conjugate gradient method, using the finite element preconditioning, is illustrated by several numerical experiments.

Final conclusions are presented in section 7.

2. The finite element method.

Although it is assumed that the reader is familiar with the finite element method, it may be helpful to review some of the basic concepts from a general point of view (for details see [2], [4] and [20]).

Consider the linear variational problem: Find $u \in V$, such that

$$(2.1) \qquad a(u,v) = f(v) \quad \text{for all } v \in V,$$

where V is a real Hilbert space, the continuous bilinear form $a(\cdot,\cdot)$: $V \times V \rightarrow \mathbb{R}$ is symmetric and V-elliptic (see [4: (1.1.3)]) and f: $V \rightarrow \mathbb{R}$ is a continuous linear form.

Then the Ritz-Galerkin method for approximating the solution of (2.1) consists in defining similar problems in finite-dimensional subspaces of V. More specifically, with any finite-dimensional subspace V_h of V, we associate the discrete problem: Find $u_h \in V_h$, such that

$$(2.2) \qquad a(u_h,v_h) = f(v_h) \quad \text{for all } v_h \in V_h.$$

The variational problems (2.1) and (2.2) have a unique solution u and u_h, respectively (see [4: theorem 1.1.1]).

Let us henceforth assume that (2.1) corresponds to an elliptic boundary value problem posed over a connected, bounded, open domain Ω in $\mathbb{R}^N$ with a Lipschitzian boundary $\partial\Omega$. For typical examples of such problems see [4: section 1.2].

The finite element method, in its simplest form, is a specific process of constructing finite element spaces V_h. The construction is characterized by three basic aspects:

2.1. Partitioning of $\overline{\Omega}$

We construct a partitioning of $\overline{\Omega}$ by subdividing $\overline{\Omega}$ in a finite number E of subdomains $\overline{\Omega}_e$, $e = 1,...,E$, called finite elements, such that

(i) $\overline{\Omega} = \bigcup_{e=1}^{E} \overline{\Omega}_e$;

(ii) every subdomain $\overline{\Omega}_e$ is closed and consists of a nonempty interior Ω_e and a Lipschitzian boundary $\partial\Omega_e$;

(iii) $\Omega_e \cap \Omega_f = \emptyset$ if $e \neq f$;

(iv) any face of every subdomain $\overline{\Omega}_e$ is either a face of another subdomain, or a subset of the boundary $\partial\Omega$.

2.2. Local interpolation

For each $\overline{\Omega}_e$, $e = 1,...,E$, we introduce finite-dimensional spaces $\mathcal{P}_e$ spanned by linearly independent local basis functions $\{\phi_i^e\}_{i=1}^{n_e}$. Locally, we approximate the restriction $u^e = u|_{\overline{\Omega}_e}$ of $u \in V$ by linear combinations of the form

$$u_h^e(x) = \sum_{i=1}^{n_e} u_i^e \phi_i^e(x) \quad \text{for } x \in \overline{\Omega}_e, \tag{2.3}$$

where the coefficients u_i^e are usually taken to be the values of u_h^e and the values of various partial derivatives of u_h^e at a preassigned collection of local nodes $\{x_i^e\}_{i=1}^{n_e}$ within $\overline{\Omega}_e$. In general we demand that, for a certain k

$$\mathcal{P}_k(\overline{\Omega}_e) \subset \mathcal{P}_e, \tag{2.4}$$

where $\mathcal{P}_k(\overline{\Omega}_e)$ is the space of polynomials of degree $\leq k$ over $\overline{\Omega}_e$.

2.3. Assembly

The collection of subdomains $\overline{\Omega}_e$ is assembled by connecting any adjacent subdomains along their mutual boundaries; by also matching corresponding local basis functions and taking into account essential boundary conditions, a system of n linearly independent global basis functions $\{\phi_i\}_{i=1}^n$ is obtained. Globally, $u \in V$ is approximated by linear combinations of the form

$$(2.5) \qquad u_h(x) = \sum_{i=1}^{n} u_i \phi_i(x) \quad \text{for } x \in \overline{\Omega},$$

where the coefficients u_i are taken to be the values of u_h and the values of various partial derivatives of u_h at the assembled collection of global nodes $\{x_i\}_{i=1}^n$ within $\overline{\Omega}$. Being linearly independent, the collection $\{\phi_i\}_{i=1}^n$ provides a basis for the finite element subspace $V_h \subset V$. Clearly, the solution $u_h = \sum_{i=1}^{n} u_i \phi_i \in V_h$ of problem (2.2) is such that the coefficients u_i are the solution of the system of linear equations

$$(2.6) \qquad \sum_{j=1}^{n} a(\phi_i,\phi_j)\, u_j = f(\phi_i) \quad \text{for } i = 1,\dots,n.$$

2.4. Examples of elliptic boundary value problems

The examples (see also [4: (1.2.23)]) correspond to the following data:

$$(2.7) \qquad \begin{cases} V = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega_0\}, \\ a(u,v) = \int_\Omega (A\nabla u) \cdot \nabla v dx, \\ f(v) \;\; = \int_\Omega fv dx - \int_{\partial\Omega_1} gv ds, \end{cases}$$

where $\partial\Omega_0 = \partial\Omega \backslash \partial\Omega_1$ is a measurable subset of the boundary $\partial\Omega$, and the following assumptions to the functions A, f and g:

(i) $A \in L^\infty(\Omega,\mathbb{R}^{N\times N})$, $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega_1)$;

(ii) A is symmetric and uniformly positive definite, i.e. $(A\xi)\cdot\xi \geq \alpha \,\|\xi\|_2^2$ almost everywhere in Ω for all $\xi \in \mathbb{R}^N$ and some $\alpha > 0$.

If the measure of $\partial\Omega_0 \subset \partial\Omega$ is positive, then the bilinear form $a(\cdot,\cdot)$ is V-elliptic (see [4: theorem 1.2.1]), and thus a unique solution $u \in V$ exists for the variational problem (2.1). Using Green's formula, it can be concluded that u is the formal solution of the boundary value problem

$$\text{(2.8)} \qquad \begin{cases} -\nabla\cdot(A\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega_0, \\ -\boldsymbol{n}\cdot(A\nabla u) = g & \text{on } \partial\Omega_1. \end{cases}$$

2.5. Examples of finite element spaces

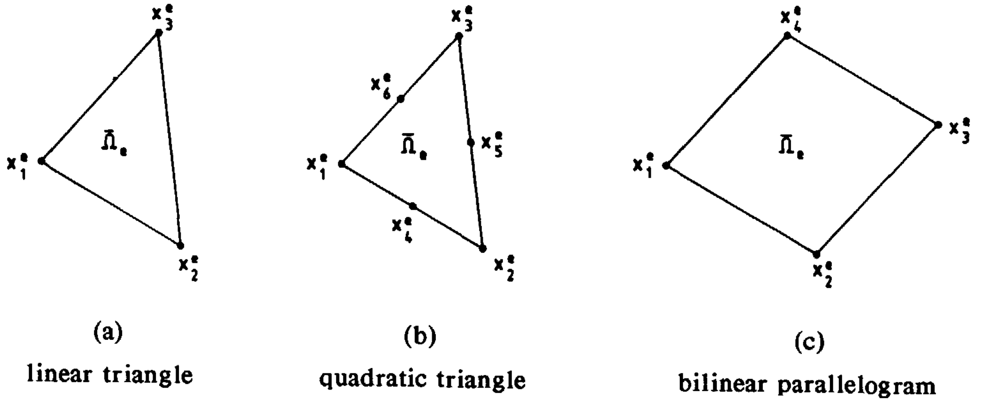
Consider the examples mentioned in section 2.4. It is assumed that $N = 2$, $\overline{\Omega} \subset \mathbb{R}^2$ is a polygon and $\partial\Omega_0$ is the union of a number of sides of subdomains $\overline{\Omega}_e$. Furthermore, the basis functions $\phi_i^e \in \mathcal{P}_e$ are uniquely defined by

$$\text{(2.9)} \qquad \phi_i^e(x_j^e) = \delta_{ij} \quad \textit{for } i,j = 1,\ldots,n_e.$$

Linear triangles: Let $\overline{\Omega}_e$, $e = 1,\ldots,E$, be triangles, $n_e = 3$ and $\{x_i^e\}_{i=1}^3$ the vertices of $\overline{\Omega}_e$ (see figure 1a). Demand that $\mathcal{P}_e = \mathcal{P}_1(\overline{\Omega}_e)$ (see [20: section 2.2.2]).

Quadratic triangles: Let $\overline{\Omega}_e$, $e = 1,\ldots,E$, be triangles, $n_e = 6$ and $\{x_i^e\}_{i=1}^6$ the vertices and the mid-side points of $\overline{\Omega}_e$ (see figure 1b). Demand that $\mathcal{P}_e = \mathcal{P}_2(\overline{\Omega}_e)$ (see [20: section 2.2.3]).

Bilinear parallelograms: Let $\overline{\Omega}_e$, $e = 1,\ldots,E$, be parallelograms, $n_e = 4$ and $\{x_i^e\}_{i=1}^4$ the vertices of $\overline{\Omega}_e$ (see figure 1c). Demand that u_h^e is a bilinear function of the natural coordinates in $\overline{\Omega}_e$. Clearly, $\mathcal{P}_1(\overline{\Omega}_e) \subset \mathcal{P}_e$ (see [20: section 2.2.4]).



(a) linear triangle

(b) quadratic triangle

(c) bilinear parallelogram

Figure 1. Examples of finite elements.

3. The preconditioned conjugate gradient method.

In [12] the conjugate gradient method (cg-method) is introduced to solve a system of linear equations

$$(3.1) \qquad Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, and $b \in \mathbb{R}^n$. For the basic relations in the cg-method see [7: section 10.2], [11: section IV-2] and [12: section 5].

Unfortunately the cg-method converges rather slowly for ill-conditioned matrices. An important way around this difficulty is to precondition A (see [2: section 1.4] and [7: section 10.3]).

This refers to finding a nonsingular matrix C, such that $\tilde{A} = C^{-1}AC^{-T}$ has a more favourable distribution of its eigenvalues than the original matrix A (for details on the rate of convergence of the cg-method see [3] and [21]). We can then apply the cg-method (with improved convergence properties) to the transformed system

$$(3.2) \qquad \tilde{A}\tilde{x} = \tilde{b},$$

where $\tilde{x} = C^T x$ and $\tilde{b} = C^{-1}b$. After transforming the iterates we obtain the preconditioned cg-method with respect to the preconditioning matrix $M = CC^T$:

```
*Algorithm 1:*
$r_0 := b - Ax_0$
for $i = 0,1,...$
    $z_i := M^{-1}r_i$
    if $r_i = 0$ then stop
    $\beta_{i-1} := z_i^T r_i / z_{i-1}^T r_{i-1}$   $(\beta_{-1} := 0)$
    $p_i := z_i + \beta_{i-1}p_{i-1}$   $(p_0 := z_0)$
    $\alpha_i := z_i^T r_i / p_i^T A p_i$
    $x_{i+1} := x_i + \alpha_i p_i$
    $r_{i+1} := r_i - \alpha_i A p_i$ .
```

3.1. Assembly of the matrix

In the following it is assumed that the preconditioned cg-method is used to solve the system of linear equations (2.6), i.e. the system (3.1), where $A = [a(\phi_i,\phi_j)]_{i,j=1,...,n}$ and $b = [f(\phi_i)]_{i=1,...,n}$, is solved. Note that the symmetry and the positive definiteness of the matrix A follow directly from the symmetry and the V-ellipticity of the bilinear form $a(\cdot,\cdot)$.

A key feature of the finite element method is the fact that A (and also b) can be assembled as the sum of contributions from each element (see, e.g., [2: section 5.2] and [20: section 3.1]). Locally, we construct the element matrices

$A_e = [a_e(\phi_i^e, \phi_j^e)]_{i,j=1,\dots,n_e}$, and then assemble A as the sum

$$(3.3)\qquad A = \sum_{e=1}^{E} N_e A_e N_e^T,$$

where the Boolean connectivity matrices $N_e \in \mathbb{R}^{n \times n_e}$ are defined by

$$(3.4)\qquad (N_e)_{i,j} = \begin{cases} 1 \text{ if } \phi_i|_{\overline{\Omega}_e} = \phi_j^e, \\ 0 \text{ if } \phi_i|_{\overline{\Omega}_e} \neq \phi_j^e . \end{cases}$$

Note that the number of unity entries of N_e is less than n_e if the solution of the boundary value problem is fixed by essential boundary conditions in some nodes within $\overline{\Omega}_e$.

3.2. Examples of preconditioning matrices

In general, a good preconditioning matrix has the following properties:

(i) $\tilde{A}$ has a more favourable distribution of its eigenvalues than A;

(ii) the factors of M can be determined quickly and do not require excessive storage in relation to A;

(iii) the system $Mz_i = r_i$ can be solved much more efficiently than $Ax = b$.

A variety of choices for the preconditioning matrix M has been discussed in the literature (see, e.g., [2: section 1.4] and [10: section 7.4]). Popular methods for computing M are to use an incomplete Cholesky decomposition (see [17] and [18]) or a modified incomplete Cholesky decomposition (see [2: section 1.4] and [8]). In the following examples we shall use the form $M = (D + L)\, D^{-1}\, (D + L^T)$, where D is a diagonal matrix and L is a strictly lower triangular matrix:

Symmetric Gauss-Seidel preconditioning (SGS): A Cholesky decomposition of A is made, during which all elimination corrections are ignored. Consequently, $D = \text{diag}(A)$ and L is equal to the strictly lower triangular part of A. For an efficient implementation see [6].

Diagonal incomplete Cholesky decomposition (DIC): A Cholesky decomposition of A is made, during which the elimination corrections are ignored in all nondiagonal places. Consequently, L is equal to the strictly lower triangular part of A and D is defined by $\text{diag}(M) = \text{diag}(A)$. For an efficient implementation see [6].

Incomplete Cholesky decomposition (IC): A Cholesky decomposition of A is made, during which the elimination corrections are ignored in appropriate nondiagonal places, which are given by the set $P \subset P_n = \{(i,j) \mid i \neq j,\ 1 \leq i,j \leq n\}$ having the property that $(i,j) \in P$ implies $(j,i) \in P$. Consequently, D and L are defined by $\ell_{i,j} = a_{i,j}$ if $(i,j) \in P$ and $m_{i,j} = a_{i,j}$ if $(i,j) \notin P$. Note that the choice $P = P_n$ results in the diagonal incomplete Cholesky decomposition.

Modified incomplete Cholesky decomposition (MIC): A Cholesky decomposition of A is made, during which elimination corrections are moved to the diagonal in appropriate places, which are given by the set $P \subset P_n$ having the property that $(i,j) \in P$ implies $(j,i) \in P$. Consequently, D and L are defined by $\ell_{i,j} = a_{i,j}$ if $(i,j) \in P$, $m_{i,j} = a_{i,j}$ if $(i,j) \in P$ and $i \neq j$, and $\sum_{j=1}^{n} m_{i,j} = \sum_{j=1}^{n} a_{i,j}$ for $i = 1,\ldots,n$.

3.3. Existence of preconditioning matrices

Since the matrix A is symmetric and positive definite, the diagonal entries of $D = \text{diag}(A)$ are positive and therefore the Gauss-Seidel preconditioning matrix exists and is positive definite.

It has been proved (see [17: theorem 2.4]) that a positive definite preconditioning matrix can be constructed by the incomplete Cholesky decomposition for each set $P \subset P_n$ having the property that $(i,j) \in P$ implies $(j,i) \in P$ if A is a symmetric M-matrix. The same statement holds for the modified incomplete

Cholesky decomposition if A is a symmetric weakly diagonally dominant M-matrix (see [8: theorem 3.1]). Note that a symmetric and positive definite matrix A is a symmetric M-matrix if, and only if, the nondiagonal entries of A are nonpositive (see [22: theorem 1]). It follows from (3.3) that the latter is true if the nondiagonal entries of A_e are nonpositive for $e = 1,\ldots,E$. For the examples in section 2.5 the following can be proved:

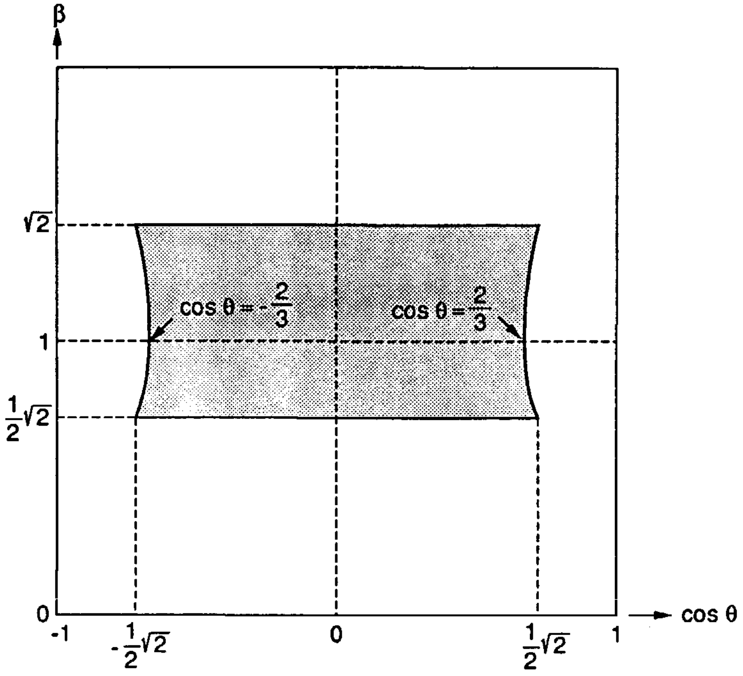


Figure 2. Nonpositivity region for nondiagonal entries of A_e (bilinear parallelograms).

Linear triangles: The nondiagonal entries of A_e are nonpositive if, and only if, $\theta \leq \pi/2$, where θ is any vertex angle in $\overline{\Omega}_e$ (see [2: page 201]).

Quadratic triangles: Some entries of A_e are positive.

Bilinear parallelograms: The nondiagonal entries of A_e are nonpositive if, and only if, $1/\sqrt{2} \leq \beta \leq \sqrt{2}$ and $\beta^2 - 3\,\beta \cos\theta + 1 > 0$, where β is the ratio of the two edge lengths of $\overline{\Omega}_e$ and θ is any vertex angle in $\overline{\Omega}_e$ (see figure 2).

Note, however, that the condition for A to be an M-matrix is not necessary. In fact, the incomplete Cholesky decomposition and the modified incomplete Cholesky decomposition of A for a nontrivial set $P \subset P_n$ exist in many situations, where A is not an M-matrix.

4. A finite element preconditioning.

The main idea of the finite element preconditioning method is to write the contribution of each subdomain $\overline{\Omega}_e$ to the global matrix A as the product of a lower triangular, a diagonal and an upper triangular matrix. These various matrices are then assembled to form a global lower triangular, a global diagonal and a global upper triangular matrix. The preconditioning matrix is defined as the product of these global matrices (an analogous idea has been presented in [9]).

For the exact definition of the finite element preconditioning matrix we start from the sum (3.3). Deleting all zero rows and columns of the matrix $N_e A_e N_e^T$ results in the matrix $\hat{A}_e \in \mathbb{R}^{\hat{n}_e \times \hat{n}_e}$, where $\hat{n}_e \leq n_e$. There are uniquely defined Boolean connectivity matrices $P_e \in \mathbb{R}^{\hat{n}_e x n_e}$ and $\hat{N}_e \in \mathbb{R}^{n x \hat{n}_e}$, such that $\hat{A}_e = P_e A_e P_e^T$ and $N_e = \hat{N}_e P_e$. Thus the sum (3.3) is equivalent to $A = \sum_{e=1}^{E} \hat{N}_e \hat{A}_e \hat{N}_e^T$.

Note that the number of unity entries of N_e, P_e and $\hat{N}_e$ is equal to $\hat{n}_e$, where $\hat{n}_e < n_e$ if the solution of the boundary value problem is fixed by essential boundary conditions in some nodes within $\overline{\Omega}_e$.

LEMMA 4.1: *If $\hat{A}_e$ is symmetric and positive semidefinite, then $\hat{A}_e$ can be written in the form*

$$\hat{A}_e = (D_e + L_e)\, D_e^+ \,(D_e + L_e^T), \tag{4.1}$$

where $D_e \geq 0$ is a diagonal matrix, L_e a strictly lower triangular matrix and $D_e^+ \geq 0$ is the generalized inverse of D_e, i.e. D_e^+ is a diagonal matrix, where

$$[D_e^+]_{i,i} = \begin{cases} 1/d^e_{i,i} & \text{if } d^e_{i,i} > 0, \\ 0 & \text{if } d^e_{i,i} = 0. \end{cases} \tag{4.2}$$

PROOF: In the proof the subscript e is omitted.

If $\hat{A}$ is positive definite, then, using the complete Cholesky decomposition, $\hat{A}$ can be written in the form $\hat{A} = (D + L)\, D^{-1}\, (D + L^T)$, where $D \geq 0$ is a diagonal matrix with positive diagonal entries and L is a strictly lower triangular matrix.

If $\hat{A}$ is singular, then it can be proved by contradiction that $\hat{A}$ can be written in the form $\hat{A} = (D + L)\, D^+\, (D + L^T)$, where $D \geq 0$ is a diagonal matrix and L a strictly lower triangular matrix, using the fact that $\hat{A}_\epsilon = A + \epsilon I$, where $\epsilon > 0$, is symmetric and positive definite. Thus the complete Cholesky decomposition of $\hat{A}_\epsilon$ exists. □

DEFINITION 4.2: If the diagonal entries of the matrix $D = \sum_{e=1}^{E} \hat{N}_e D_e \hat{N}_e^T$ are positive, then the finite element preconditioning (FEP) matrix M is defined by

$$M = (D + L)\, D^{-1}\, (D + L^T), \tag{4.3}$$

where $L = \sum_{e=1}^{E} \hat{N}_e L_e \hat{N}_e^T$.

Note that in [9] the matrices $\hat{A}_e$ are written in the form $\hat{A}_e = C_e C_e^T$. If the diagonal entries of the matrix $C = \sum_{e=1}^{E} \hat{N}_e C_e \hat{N}_e^T$ are positive, then the element matrix factorization (EMF) preconditioning matrix M is defined by CC^T. Thus, the finite element preconditioning and the element matrix factorization do not result in the same preconditioning matrix.

In section 4.1 a variant of the reverse Cuthill-McKee algorithm to fulfil the condition in definition 4.2 will be discussed.

4.1. Existence of the finite element preconditioning

For the examples in section 2.5 the element matrices $A_e = [a_e(\phi_i^e,\phi_j^e)]_{i,j=1,...,n_e}$ are constructed, where

$$(4.4) \qquad a_e(u_e,v_e) = \int_{\Omega_e} (A\nabla u_e) \cdot \nabla v_e dx$$

for $u_e, v_e \in V_e = \{v \in H^1(\Omega_e) \mid v = 0 \text{ on } \partial\Omega_e \cap \partial\Omega_0\}$.

If $\partial\Omega_e \cap \partial\Omega_0 \neq \emptyset$, i.e. the solution of (2.8) is fixed by essential boundary conditions in some nodes within $\overline{\Omega}_e$, then $\hat{n}_e < n_e$ and $\hat{A}_e$ is a symmetric and positive definite matrix, which follows directly from the symmetry and the V-ellipticity of the bilinear form $a_e(\cdot,\cdot)$. From lemma 4.1 it follows that $\hat{A}_e$ can be written in the form $\hat{A}_e = (D_e + L_e) D_e^{-1} (D_e + L_e^T)$, where the diagonal entries are positive.

If $\partial\Omega_e \cap \partial\Omega_0 = \emptyset$, then $\hat{n}_e = n_e$ and A_e is a symmetric and positive semidefinite matrix with null space $N(\hat{A}_e) = \text{span } \{1\}$, where $1 = (1,...,1)^T$. From lemma 4.1 it follows that $\hat{A}_e$ can be written in the form (4.1). Since the leading principle submatrix that is obtained from $\hat{A}_e$ by omitting the last row and column is symmetric and positive definite and $\prod_{i=1}^{n_e} d_{i,i}^e = \det \hat{A}_e = 0$, it follows that $d_{i,i}^e > 0$ for $i = 1,...,n_e-1$ and $d_{n_e,n_e}^e = 0$.

A necessary and sufficient condition for the fulfilment of the condition of definition 4.2 can be derived for these examples.

DEFINITION 4.3 : A global node number i is maximal, if the following holds for all subdomains $\overline{\Omega}_e$, where $x_i \in \overline{\Omega}_e$:

(i) if $x_j \in \overline{\Omega}_e$ and $x_j \neq x_i$, then $j < i$;
(ii) $\hat{n}_e = n_e$, i.e. the solution of (2.8) is not fixed by essential boundary conditions in some nodes within $\overline{\Omega}_e$.

THEOREM 4.4 : *The diagonal entries of the matrix* $D = \sum_{e=1}^{E} \hat{N}_e D_e \hat{N}_e^T$ *are positive if, and only if, there are no maximal global node numbers.*

PROOF : Let $1 \leq i \leq n$.

($\Rightarrow$): If $d_{i,i} = 0$, then $d^e_{n_e,n_e} = 0$ for all subdomains $\overline{\Omega}_e$, where $x_i \in \overline{\Omega}_e$. Thus $\hat{n}_e = n_e$ and $j < i$ for all $x_j \in \overline{\Omega}_e$, where $x_j \neq x_i$, i.e. i is a maximal global node number.

($\Leftarrow$): If i is a maximal global node number, then $d^e_{n_e,n_e} = 0$ for all subdomains $\overline{\Omega}_e$, where $x_i \in \overline{\Omega}_e$. Thus $d_{i,i} = \sum_{\{e | x_i \in \overline{\Omega}_e\}} d^e_{n_e,n_e} = 0$. □

Note that the second condition in theorem 4.4 can easily be checked before assembling the matrix A, because the global node numbers of all subdomains $\overline{\Omega}_e$ are known after the generation of the finite element mesh, i.e. after $\overline{\Omega}$ is subdivided into subdomains $\overline{\Omega}_e$ and all global nodes are numbered.

A global node numbering without maximal global node numbers can be constructed using a variant of the reverse Cuthill-McKee algorithm (see, e.g., [2: page 276], [16] and [20: section 3.2.3]).

Let $\{x_i\}_{i=1}^{m}$, where $m > n$, be the collection of all nodes. Introduce the undirected graph $G = (X,E)$ associated with the partitioning of $\overline{\Omega}$. Here, $X = \{x_i\}_{i=1}^{m}$ is the set of nodes and $\{x_i,x_j\} \in E$, if $x_i,x_j \in \overline{\Omega}_e$ for a subdomain $\overline{\Omega}_e$. Let $G' = (X,E')$, where $E' \subset E$, be a connected subgraph of G. A global node numbering without maximal global node numbers can be constructed as follows:

First step: Choose a node in X in which the solution of (2.8) is fixed by an essential boundary condition, e.g. with minimal degree corresponding to the subgraph G'. This node is the starting node; it receives the number 1. The first level is defined as the collection consisting of this node.

Iteration step: For the nodes of the last numbered level with increasing new numbers, one determines successively their as yet unnumbered adjacent nodes (according to the subgraph G'). The latter are numbered sequentially, e.g. with increasing degree. The next level is defined as the collection of the nodes that were numbered in this iteration.

Last step: After all nodes in X have been numbered, the numbering obtained is reversed. It is compressed to a numbering for the global nodes $\{x_i\}_{i=1}^{n}$ by deleting all nodes in which the solution of (2.8) is fixed by essential boundary conditions, and shifting the numbers accordingly.

4.2. The smallest eigenvalue of $\tilde{A}$

THEOREM 4.5: *Let $C = (D + L)\, D^{-1/2}$, where D and L are defined in definition 4.2, and $\tilde{A} = C^{-1} A C^{-T}$. Then the smallest eigenvalue of $\tilde{A}$ is equal to 1.*

PROOF: We will first prove that the matrix $R = A - M$ is positive semidefinite.

Define $\overline{A}_e = \hat{N}_e \hat{A}_e \hat{N}_e^T$, $\overline{L}_e = \hat{N}_e L_e \hat{N}_e^T$ and $\overline{D}_e = \hat{N}_e D_e \hat{N}_e^T$ for $e = 1,\dots,E$, then $\overline{A}_e = (\overline{D}_e + \overline{L}_e)\, \overline{D}_e^+ \,(\overline{D}_e + \overline{L}_e^T)$ and $A = \sum_{e=1}^{E} \overline{A}_e$. Furthermore, let $\ell_{e;i}$ be the i^{th} column of $\overline{D}_e + \overline{L}_e$ and $\mathrm{d}_{e;i}$ the i^{th} diagonal entry of $\overline{D}_e$. Then we can define for $i = 1,\dots,n$:

$$A^{(i)} = \sum_{e=1}^{E} \ell_{e;i}\, \mathrm{d}_{e;i}^{+}\, \ell_{e;i}^{T},$$

$$M^{(i)} = \Big(\sum_{e=1}^{E} \ell_{e;i}\Big) \Big(\sum_{e=1}^{E} \mathrm{d}_{e;i}\Big)^{-1} \Big(\sum_{e=1}^{E} \ell_{e;i}\Big)^{T},$$

$$R^{(i)} = A^{(i)} - M^{(i)},$$

so that $A = \sum_{i=1}^{n} A^{(i)}$, $M = \sum_{i=1}^{n} M^{(i)}$ and $R = \sum_{i=1}^{n} R^{(i)}$.

For all $x \in \mathbb{R}^n$ we have

$$x^T R^{(i)} x = \sum_{e=1}^{E} \lambda_{e;i}^2 \, d_{e;i}^{+} - \Big(\sum_{e=1}^{E} \lambda_{e;i}\Big)^2 \Big(\sum_{e=1}^{E} d_{e;i}\Big)^{-1} \geq 0,$$

where $\lambda_{e;i} = x^T \ell_{e;i}$ for $e = 1,\dots,E$ (use the Cauchy-Schwarz inequality for the vectors $[\lambda_{e;i}(d_{e;i}^{+})^{1/2}]_{i=1,\dots,n}$ and $[d_{e;i}^{1/2}]_{i=1,\dots,n}$ and remind that $\ell_{e;i} = 0$ if $d_{e;i} = 0$). Therefore $x^T R x = \sum_{i=1}^{n} x^T R^{(i)} x \geq 0$ for all $x \in \mathbb{R}^n$.

Using the positive definiteness of M (this is assumed in definition 4.2) we find that

$$x^T A x / x^T M x = 1 + x^T R x / x^T M x \geq 1$$

for all $x \in \mathbb{R}^n$ and thus $x^T \tilde{A} x \geq 1$ for all $x \in \mathbb{R}^n$.

From the equality of the first column of A and M, i.e. $Ae_1 = Me_1$ for $e_1 = (1,0,\dots,0)^T$, it follows that $C^T e_1$ is an eigenvector of $\tilde{A}$ with eigenvalue 1. □

4.3. Generalizations of the finite element preconditioning

In order to generalize the finite element preconditioning, the finite element method is represented in a slightly generalized form. Again, the construction of finite element spaces V_h is characterized by three basic aspects:

Partitioning of $\overline{\Omega}$*:* Construct a partitioning of $\overline{\Omega}$ in a finite number E of subdomains $\overline{\Omega}_e$, $e = 1,\dots,E$, such that the conditions (i)-(iii) of section 2.1 are fulfilled.

Local interpolation : For each $\overline{\Omega}_e$, $e = 1,...,E$, introduce finite-dimensional spaces $\mathcal{P}_e$ spanned by linearly independent local basis functions $\{\phi_i^e\}_{i=1}^{n_e}$. Locally, we approximate the restriction $u^e = u|_{\overline{\Omega}_e}$ of $u \in V$ by linear combinations of the form

$$(4.5) \qquad u_h^e(x) = \sum_{i=1}^{n_e} u_i^e \phi_i^e(x) \quad \text{for } x \in \overline{\Omega}_e,$$

where the coefficients u_i^e are usually taken to be values of u_h^e and the values of various partial derivatives of u_h^e at a preassigned collection of local nodes $\{x_i^e\}_{i=1}^{n_e}$ within $\overline{\Omega}_e$. We construct a partitioning of $\overline{\Omega}_e$ in a finite number F_e of subsubdomains $\overline{\Omega}_{e,f}$, $f = 1,...,F_e$, called finite elements, such that

(i) $\overline{\Omega}_e = \sum_{f=1}^{F_e} \overline{\Omega}_{e,f}$;

(ii) every subsubdomain $\overline{\Omega}_{e,f}$ is closed and consists of a nonempty interior $\Omega_{e,f}$ and a Lipschitzian boundary $\partial\Omega_{e,f}$;

(iii) $\Omega_{e,f} \cap \Omega_{e,g} = \emptyset$ if $f \neq g$;

(iv) any face of every subsubdomain $\overline{\Omega}_{e,f}$ is either a face of another subsubdomain, or a subset of the boundary $\partial\Omega$.

Let $\mathcal{P}_{e,f} = \{u^{e,f} = u|_{\overline{\Omega}_{e,f}} \mid u \in \mathcal{P}_e\}$, $f = 1,...,F_e$. In general, we demand that, for a certain k,

$$(4.6) \qquad \mathcal{P}_k(\overline{\Omega}_{e,f}) \subset \mathcal{P}_{e,f} \,.$$

Assembly : See section 2.3.

It is again assumed that the preconditioned cg-method is used to solve the system of linear equations (2.6), i.e. the system (3.1) is solved, where $A = [a(\phi_i,\phi_j)]_{i,j=1,...,n}$ and $b = [f(\phi_i)]_{i=1,...,n}$.

A can be assembled as the sum of contributions $A_e = [a_e(\phi^e_i,\phi^e_j)]_{i,j=1,...,n_e}$ from each subdomain Ω_e.

For the definition of the generalized finite element preconditioning matrix we start from the assembly of the matrix and then follow the procedure given in section 4 literally.

Two special cases of the generalized finite element preconditioning must be mentioned. If $F_e = 1$, $e = 1,...,E$, then we obtain the nongeneralized finite element preconditioning (see the beginning of section 4). If $E = 1$, then $M = A$, i.e. a complete Cholesky decomposition of A is made. Between these two extremes there are many possibilities for selecting an appropriate preconditioning that balances the expected improvement of the convergence properties of the cg-method and the increase of computational work per iteration.

Note that the element matrix factorization (see [9]) can be generalized along the same lines.

5. The Neumann boundary value problem.

Consider the examples in section 2.4. If the measure of $\partial\Omega_0 \subset \partial\Omega$ is zero, then we have to define

$$(5.1) \qquad V = H^1(\Omega) \,/\, \mathcal{P}_0(\Omega),$$

where $\mathcal{P}_0(\Omega)$ is the space of constant functions over Ω and accordingly assume that f and g satisfy the compatability condition, i.e.

$$(5.2) \qquad \int_\Omega f dx = \int_{\partial\Omega} g ds,$$

to ensure that the bilinear form $a(\cdot,\cdot)$ is V-elliptic.

Again, a unique solution $u \in V$ exists for (2.1) and now u is the formal solution of the boundary value problem

$$(5.3)\qquad \begin{cases} -\nabla \cdot (A\nabla u) = f \text{ in } \Omega, \\ \\ -\boldsymbol{n} \cdot (A\nabla u) = g \text{ on } \partial\Omega. \end{cases}$$

5.1. The preconditioned conjugate gradient method

In [15] the preconditioned cg-method is used to solve a system of linear equations

$$(5.4)\qquad Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric, singular and positive semidefinite, and b is in the range $R(A)$ of A (for examples of preconditioning matrices see section 3.2). Note that $A(x+y) = b$ for all y in the null space $N(A)$ of A (clearly, $N(A) = R(A)^{\perp}$).

In the following it is assumed that the preconditioned cg-method is used to solve the system of linear equations (2.6) corresponding to the Neumann boundary value problem, i.e. the system (3.1) is solved, where $A = [a(\phi_i,\phi_j)]_{i,j=1,...,n}$ and $b = [f(\phi_i)]_{i=1,...,n}$. Note that the symmetry and the positive semidefiniteness of the matrix A follow directly from the symmetry and the V-ellipticity of the bilinear form $a(\cdot,\cdot)$. It follows directly that $N(A) = \text{span}\ \{\mathbf{1}\}$, where $\mathbf{1} = (1,...,1)^T$, and from (5.2) that $b \in R(A)$.

5.2. Existence of preconditioning matrices

Since the matrix A is symmetric and positive semidefinite, and $N(A) = \text{span}\ \{\mathbf{1}\}$, the diagonal entries of $D = \text{diag}\ (A)$ are positive and therefore the symmetric Gauss-Seidel preconditioning matrix exists and is positive definite.

It has been proved (see [15: theorem 3.2]) that a positive definite preconditioning matrix can be constructed by the incomplete Cholesky decomposition for each nontrivial set $P \subset P_n$ that has the property that $(i,j) \in P$ implies $(j,i) \in P$ if A is a singular Stieltjes matrix. Note that a symmetric, singular and positive semidefinite matrix is a singular Stieltjes matrix if, and only if, the nondiagonal entries of A are nonpositive (see [15: definition 3.1]).

It has also been proved (see [8: theorem 3.1]) that a positive semidefinite preconditioning matrix can be constructed by the modified incomplete Cholesky decomposition for each set $P \subset P_n$ that has the property that $(i,j) \in P$ implies $(j,i) \in P$ if A is a singular weakly diagonally dominant Stieltjes matrix. In this case we have to use the form $M = (D + L)\, D^+\, (D + L^T)$ for the resulting preconditioning matrix and in algorithm 1 $z_i := M^+ r_i$ has to be computed instead of $z_i := M^{-1} r_i$.

5.3. The finite element preconditioning

For the definition of the finite element preconditioning see section 4, where definition 4.2 is replaced by

DEFINITION 5.1: If the first $n-1$ diagonal entries of the matrix $D = \sum_{e=1}^{E} \hat{N}_e D_e \hat{N}_e^T$ are positive, then the finite element preconditioning matrix M is defined by

$$M = (D + L)\, D^+\, (D + L^T), \tag{5.5}$$

where $L = \sum_{e=1}^{E} \hat{N}_e L_e \hat{N}_e^T$.

Note that in algorithm 1 $z_i := M^+ r_i$ has to be computed instead of $z_i := M^{-1} r_i$. With respect to the Neumann boundary problem, a necessary and sufficient condition for the fulfilment of the condition of definition 5.1 can be derived for the examples in section 2.5.

THEOREM 5.2: *The first $n-1$ diagonal entries of the matrix $D = \sum_{e=1}^{E} \hat{N}_e D_e \hat{N}_e^T$ are positive if, and only if, there is only one maximal global node number.*

PROOF: See the proof of theorem 4.4. Note that $\hat{n}_e = n_e$, $e = 1,\dots,E$. □

A global node numbering with only one maximal global node number can be constructed using a variant of the Cuthill-McKee algorithm (see section 4.1), where the starting node can be chosen arbitrarily.

THEOREM 5.3: *Let $C = (D + L)(D^+)^{1/2}$, where D and L are defined in definition 5.1, and $\tilde{A} = C^+A(C^+)^T$. Then the smallest positive eigenvalue of $\tilde{A}$ is equal to* 1.

PROOF: The first part of this proof is the same as the proof of theorem 4.5. For all $x \in \mathbb{R}^n$ we have

$$x^T R^{(i)} x = \sum_{e=1}^{E} \lambda_{e;i}^2\, d_{e;i}^+ - \Big(\sum_{e=1}^{E} \lambda_{e;i}\Big)^2 \Big(\sum_{e=1}^{E} d_{e;i}\Big)^+ \geq 0,$$

where $\lambda_{e;i} = x^T \ell_{e;i}$ for $e = 1,\dots,E$. Therefore $x^T Rx = \sum_{i=1}^{n} x^T R^{(i)} x \geq 0$ for all $x \in \mathbb{R}^n$. Using the positive semidefiniteness of M and $N(M) = \text{span }\{1\}$ (this follows from definition 5.1) we find that

$$x^T Ax\, / x^T Mx = 1 + x^T Rx / x^T Mx \geq 1$$

for all $x \in (\text{span }\{1\})^\perp$, and thus $x^T \tilde{A} x \geq 1$ for all $x \in R(\tilde{A}) = (\text{span }\{1\})^\perp$.

From the equality of the first column of A and M, i.e. $Ae_1 = Me_1$ for $e_1 = (1,0,\dots,0)^T$, it follows that $C^T e_1$ is an eigenvector of $\tilde{A}$ with eigenvalue 1. □

Note that the finite element preconditioning can be generalized for the Neumann boundary value problem analogously to section 4.3.

6. Numerical experiments.

In the preceding sections the definition and the existence of the finite element preconditioning were discussed. In this section the convergence properties of the preconditioned cg-method using the finite element preconditioning, the element matrix factorization (see [9]) and some well-known preconditioning methods (see section 3.2) will be evaluated, using some numerical experiments.

We are especially interested in the situations, where the coefficient matrix is not an M-matrix and thus the existence of incomplete Cholesky decompositions and modified incomplete Cholesky decompositions is not guaranteed (see section 3.3). We are also interested in the expected convergence properties of the preconditioned cg-method using generalized finite element preconditionings.

All computations were made in double precision on an Alliant FX/40.

6.1. The Dirichlet model problem

Let $\Omega = \{(x,y) \mid 0 < x - y \cot \theta < 1,\ 0 < y/\sin \theta < 1\}$ be the unit rhombus with the vertex angle $0 < \theta < \pi$. Consider the Dirichlet boundary value problem

$$(6.1)\qquad \begin{cases} -\Delta u = f \text{ in } \Omega, \\ u = 0 \quad \text{on } \partial\Omega, \end{cases}$$

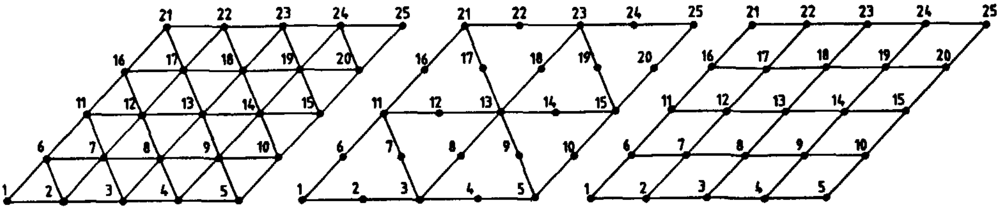
where f is chosen, such that (6.1) has the solution $u(x,y) = x'(1 - x')\, y'(1 - y') \cdot e^{x'y'}$, $(x,y) \in \Omega$, where x' and y' are the natural coordinates of the rhombus, i.e.

$$(6.2)\qquad x' = x - y \cot \theta,\ y' = y/\sin \theta.$$

Construct a regular partitioning of $\overline{\Omega}$ by subdividing $\overline{\Omega}$ into

(i) $2 \cdot 30^2$ linear triangles;
(ii) $2 \cdot 15^2$ quadratic triangles;
(iii) 30^2 bilinear parallelograms

(see section 2.5), as in figure 3.



(a)
linear triangles (b)
quadratic triangles (c)
bilinear parallelograms

Figure 3. Partitioning of numerical examples.

The global nodes are numbered lexicographically from bottom to top and from left to right (see figure 3). The matrix $A \in \mathbb{R}^{841 \times 841}$ is assembled as the sum of the element matrices (see section 3.1).

In order to avoid discretization errors, the preconditioned cg-method is applied to the system $Ax = b$, where the solution x corresponds to the solution of the boundary value problem (6.1). We start with the vector $x_0 = 0$ and terminate if $\|x - x_i\|_A / \|x\|_A \leq 10^{-8}$.

For the preconditioning matrix M we opt for the following:

(i) No preconditioning, i.e. $M = I$ (-);

(ii) Diagonal scaling, i.e. $M = \text{diag}\,(A)$ (diag);

(iii) Symmetric Gauss-Seidel preconditioning (SGS);

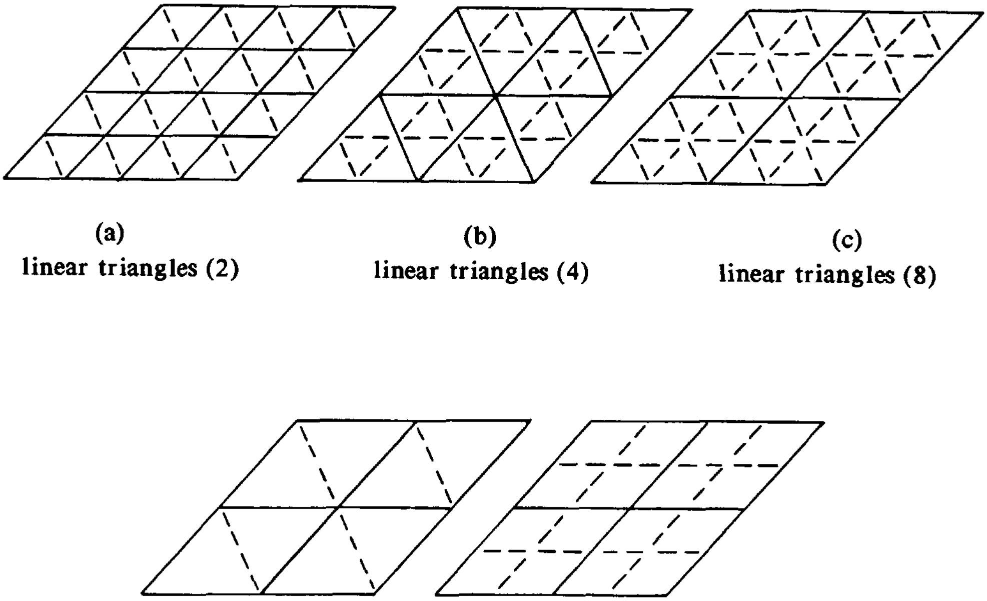
(iv) Diagonal incomplete Cholesky decomposition (DIC);

(v) Incomplete Cholesky decomposition (IC);

(vi) Modified incomplete Cholesky decomposition (MIC);

(vii) Element matrix factorization (EMF);

(viii) Finite element preconditioning (FEP).



(a) linear triangles (2)

(b) linear triangles (4)

(c) linear triangles (8)

(d) quadratic triangles (2)

(e) bilinear parallelograms (4)

Figure 4. Generalized partitioning of numerical examples.

Note that in [9: (2.4)] a preconditioning parameter $\xi \geq 0$ is used in order to modify the element matrix factorization. In these experiments the influence of ξ is not studied, i.e. $\xi = 0$ is chosen. The finite element preconditioning can be modified analogously, possibly resulting in a faster convergence for the optimal parameter ξ.

The triangular factors of M have the same sparsity pattern as the corresponding triangular parts of the matrix A for the preconditionings (iii)-(viii). This implies that during the incomplete Cholesky decomposition elimination corrections are ignored and during the modified incomplete Cholesky decomposition elimination corrections are moved to the diagonal in the nondiagonal places that correspond to the nonzero entries of A (for all vertex angles θ).

In order to generalize the finite element preconditioning and the element matrix factorization for the examples, the factors of the preconditioning matrix M are computed by summing the contributions of several finite elements, as in figure 4.

Table I. Number of nonzero entries of the matrix A.

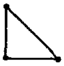
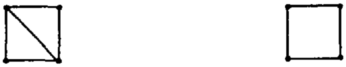
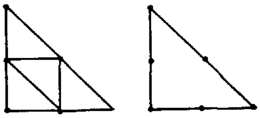
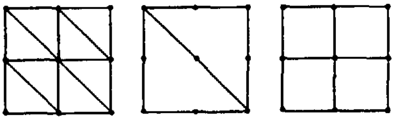
	$7N^2 + 6N + 1$
	$9N^2 + 6N + 1$
	$\frac{23}{2} N^2 + 8N + 1$
	$16N^2 + 8N + 1$

Table II. Number of iterations for the Dirichlet model problem († indicates that the preconditioning matrix does not exist).

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
-, diag	88	71	81	106	135
SGS	34	33	31	42	53
DIC	28	30	27	31	32
IC(1)	23	23	20	14	10
MIC(1)	18	19	17	12	8
EMF(1)	34	33	31	30	35
FEP(1)	30	29	27	26	27
IC(2)	23	23	20	14	10
MIC(2)	18	19	17	12	8
EMF(2)	34	33	30	29	34
FEP(2)	30	29	27	25	25
IC(4)	20	20	18	13	9
MIC(4)	15	17	15	12	8
EMF(4)	37	37	36	35	42
FEP(4)	26	25	25	25	27
IC(8)	20	20	17	13	9
MIC(8)	15	16	15	12	8
EMF(8)	36	35	33	31	33
FEP(8)	25	24	23	23	23

(a) linear triangles

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
-	101	80	95	121	165
diag	105	83	96	122	167
SGS	38	36	35	47	63
DIC	31	32	28	30	†
IC(1)	22	22	19	13	20
MIC(1)	16	17	16	13	11
EMF(1)	41	39	38	36	43
FEP(1)	27	26	26	26	28
IC(2)	22	22	19	13	13
MIC(2)	16	17	16	13	10
EMF(2)	40	38	35	35	40
FEP(2)	27	25	24	24	27

(b) quadratic triangles

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
-, diag	92	67	58	63	80
SGS	37	31	30	30	37
DIC	33	28	28	28	32
IC(1)	26	23	22	18	11
MIC(1)	20	20	18	16	10
EMF(1)	36	33	31	28	30
FEP(1)	31	28	28	27	27
IC(4)	22	20	19	16	10
MIC(4)	16	16	16	14	10
EMF(4)	38	35	33	31	30
FEP(4)	26	24	23	23	23

(c) bilinear parallellograms

-	no preconditioning
diag	diagonal scaling
SGS	symmetric Gauss-Seidel
DIC	diagonal incomplete Cholesky
IC	incomplete Cholesky
MIC	modified incomplete Cholesky
EMF	element matrix factorization
FEP	finite element preconditioning

Note that the amount of computational work per iteration of algorithm 1 is equal to the computational work for two inner products ($2n$ flops), three vector updates ($3n$ flops) and two matrix-vector products, where the number of floating point operations is equal to the number of possible nonzero entries of the matrix A (see table I). In our experiments $N = 28$ and $n = (N+1)^2$. The amount of computational work per iteration can be reduced for the symmetric Gauss-Seidel preconditioning and the diagonal incomplete Cholesky decomposition by using an idea, which is described in [6].

In order to compare the generalized finite element preconditionings we compute the preconditioning matrix by the incomplete Cholesky decomposition and the modified incomplete Cholesky decomposition with respect to the set $P \subset P_n$, which corresponds to the nonzero entries of the factors of the finite element preconditioning matrix (for all angles θ). Thus for each generalized finite element preconditioning and element matrix factorization and their corresponding incomplete Cholesky decomposition and modified incomplete Cholesky decomposition the same amount of computational work is required per iteration.

The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table II.

6.2. The Neumann model problem

Let $\Omega = \{(x,y) \mid 0 < x - y \cot \theta < 1, 0 < y/\sin \theta < 1\}$ be the unit rhombus with vertex angle $0 < \theta < \pi$. Consider the Neumann boundary value problem

$$(6.3) \qquad \begin{cases} -\Delta u = f & \text{in } \Omega, \\ -\dfrac{\partial u}{\partial n} = g & \text{on } \partial\Omega, \end{cases}$$

where f and g are chosen, such that (6.3) has the same solution as the Dirichlet boundary value problem (6.1).

Table III. Number of iterations for the Neumann boundary value problem († indicates that the preconditioning matrix does not exist).

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
-	166	111	123	158	222
diag	155	104	118	146	210
SGS	69	54	44	71	104
DIC	55	49	38	55	†
IC(1)	45	38	30	23	†
MIC(1)	38	35	30	22	17
EMF(1)	67	52	46	49	62
FEP(1)	41	36	33	35	52
IC(2)	45	38	30	23	†
MIC(2)	38	35	30	22	17
EMF(2)	61	47	43	48	69
FEP(2)	40	35	32	32	42
IC(4)	38	32	27	21	†
MIC(4)	31	29	27	20	16
EMF(4)	59	47	45	49	69
FEP(4)	35	33	31	32	39
IC(8)	38	31	26	20	16
MIC(8)	31	29	26	21	16
EMF(8)	53	44	39	39	50
FEP(8)	34	30	28	26	30

(a) linear triangles

-	no preconditioning
diag	diagonal scaling
SGS	symmetric Gauss-Seidel
DIC	diagonal incomplete Cholesky
IC	incomplete Cholesky
MIC	modified incomplete Cholesky
EMF	element matrix factorization
FEP	finite element preconditioning

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
-	191	125	142	186	280
diag	184	121	137	165	258
SGS	77	60	49	80	126
DIC	63	53	40	†	†
IC(1)	43	36	28	22	†
MIC(1)	35	33	29	26	23
EMF(1)	61	52	49	55	75
FEP(1)	38	36	35	36	46
IC(2)	42	35	28	22	†
MIC(2)	36	32	29	24	19
EMF(2)	60	46	43	45	60
FEP(2)	36	33	30	30	36

(b) quadratic triangles

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
-	173	105	88	95	137
diag	158	98	83	84	127
SGS	74	51	48	51	75
DIC	64	46	45	48	65
IC(1)	49	38	34	29	21
MIC(1)	40	36	35	31	22
EMF(1)	58	50	49	45	52
FEP(1)	41	36	36	34	40
IC(4)	42	32	29	25	19
MIC(4)	31	30	28	26	19
EMF(4)	53	44	40	41	46
FEP(4)	34	31	29	29	31

(c) bilinear parallellograms

Assume the same partitionings and preconditionings as in section 6.1. Note that for the Neumann boundary value problem $N = 30$.

The preconditioned cg-method is again applied to the system $Ax = b$, where x corresponds to the solution of the boundary value problem (6.3). Start with the vector $x_0 = 0$ and terminate if $\|x - x_i\|_A \,/\, \|x\|_A \le 10^{-8}$.

The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table III.

6.3. An inhomogeneous model problem

Let $\Omega = \{(x,y) \mid 0 < x - y \cot\theta < 1,\ 0 < y/\sin\theta < 1\}$ be the unit rhombus with vertex angle $0 < \theta < \pi$. Consider the boundary value problem

$$(6.4)\qquad \begin{cases} -10^3\,\Delta u = 1 \text{ in } \Omega_1, & -10^{-3}\,\Delta u = 0 \text{ in } \Omega_2, \quad -\Delta u = 0 \text{ in } \Omega_3, \\ u = 0 \text{ on } \partial\Omega_0, & -\dfrac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega_1. \end{cases}$$

Choose

$$(6.5)\qquad \begin{cases} \Omega_1 = \{(x,y) \mid 0 < x' < 1/3,\ 0 < y' < 1/3\}, \\ \Omega_2 = \{(x,y) \mid 0 < x' < 2/3,\ 0 < y' < 2/3\}\backslash\overline{\Omega}_1, \\ \Omega_3 = \Omega\backslash\overline{\Omega}_2, \\ \partial\Omega_0 = \{(x,y) \mid x' = 1,\ 0 \le y' \le 1\} \cup \{(x,y) \mid 0 \le x' \le 1,\ y' = 1\}, \\ \partial\Omega_1 = \partial\Omega\backslash\partial\Omega_0, \end{cases}$$

where x' and y' are the natural coordinates of the rhombus (see figure 5).

Table IV. Number of iterations for the inhomogeneous model problem († indicates that the preconditioning matrix does not exist).

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
diag	143	102	123	152	198
SGS	70	56	48	71	90
DIC	54	50	41	57	†
IC(1)	44	39	31	23	†
MIC(1)	33	28	25	19	19
EMF(1)	75	47	43	45	67
FEP(1)	40	29	29	29	38
IC(2)	44	39	31	23	†
MIC(2)	32	27	26	19	19
EMF(2)	60	41	41	44	68
FEP(2)	40	29	28	26	32
IC(4)	37	34	28	22	†
MIC(4)	29	24	23	18	17
EMF(4)	72	49	47	48	67
FEP(4)	36	27	27	27	31
IC(8)	37	34	28	21	17
MIC(8)	28	24	22	18	16
EMF(8)	63	44	42	41	52
FEP(8)	36	26	24	23	27

(a) linear triangles

diag	diagonal scaling
SGS	symmetric Gauss-Seidel
DIC	diagonal incomplete Cholesky
IC	incomplete Cholesky
MIC	modified incomplete Cholesky
EMF	element matrix factorization
FEP	finite element preconditioning

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
diag	170	119	143	177	244
SGS	78	62	53	79	108
DIC	64	55	44	†	†
IC(1)	41	38	30	24	†
MIC(1)	33	26	26	21	21
EMF(1)	79	55	52	54	81
FEP(1)	39	28	28	29	37
IC(2)	41	37	30	23	†
MIC(2)	34	27	26	21	19
EMF(2)	71	48	46	47	70
FEP(2)	38	27	27	26	34

(b) quadratic triangles

θ	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$
diag	146	97	87	88	119
SGS	73	53	48	52	66
DIC	62	48	46	48	58
IC(1)	48	39	35	30	20
MIC(1)	35	27	26	24	20
EMF(1)	59	43	44	42	54
FEP(1)	41	29	28	27	30
IC(4)	41	34	30	26	18
MIC(4)	27	24	23	20	17
EMF(4)	63	44	42	41	44
FEP(4)	36	26	24	24	26

(c) bilinear parallelograms

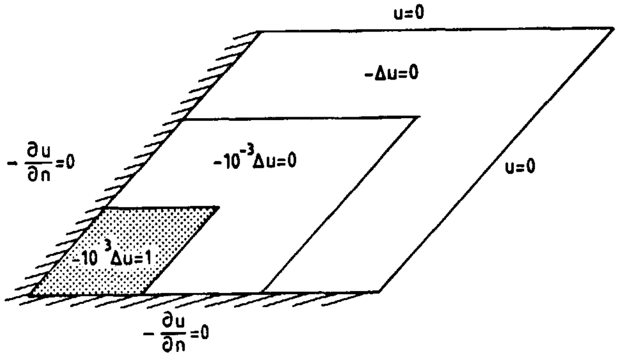

u=0
-Δu=0
$-\frac{\partial u}{\partial n}=0$
$-10^{-3}\Delta u=0$
u=0
$-10^{3}\Delta u=1$
$-\frac{\partial u}{\partial n}=0$

Figure 5. An inhomogeneous model problem.

The partitionings and preconditionings are the same as in section 6.1. Note that $N = 29$ for the inhomogeneous model problem under consideration.

The preconditioned cg-method is used to solve the system of linear equations (2.6), i.e. the system $Ax = b$ is solved, where $A = [a(\phi_i,\phi_j)]_{i,j=1,...,n}$ and $b = [f(\phi_i)]_{i=1,...,n}$ (see section 2.4).

We start with the vector $x_0 = 0$ and choose a termination criterion, such that the condition $\|x - x_i\|_A / \|x\|_A \leq 10^{-8}$ is fulfilled (see [14] for this termination criterion). The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table IV.

6.4. Inferences

From the results shown in tables II-IV it is clear that the incomplete Cholesky decomposition exists in many situations, where A is not an M-matrix. The modified incomplete Cholesky decomposition even exists for all executed experiments.

Furthermore, the modified incomplete Cholesky decomposition gives the best results. The finite element preconditioning and the incomplete Cholesky decomposition are rather competitive, but the Dirichlet model problem is more favourable for the incomplete Cholesky decomposition. Both preconditioning methods result in less iterations than the diagonal incomplete Cholesky decomposition. The element matrix factorization results in more iterations than the finite element preconditioning and seems to be very sensitive for inhomogeneities.

In these experiments the convergence properties improve by generalizing the considered preconditioning methods, but this is more than offset by the increased amount of computational work per iteration. Thus, it is not profitable to generalize.

7. Conclusions.

From the numerical experiments it is clear that the modified incomplete Cholesky decomposition can exist and can result in a very effective preconditioning matrix for the conjugate gradient method in situations, where the coefficient matrix is not an M-matrix. However, the existence of modified incomplete Cholesky decompositions is then not guaranteed. In these situations the existence of the finite element preconditioning is guaranteed and it results in an effective preconditioning matrix. Moreover, when the existence of both preconditioning matrices is guaranteed the finite element preconditioning only is slightly less effective.

It can be concluded that the finite element preconditioning is a robust and effective preconditioning method for solving second-order symmetric elliptic boundary value problems by the finite element method and the conjugate gradient method.

The finite element preconditioning is very promising for higher order symmetric elliptic boundary value problems, i.e. plate problems (see [4: section 6]) and shell problems (see [4: section 8]) and, in combination with generalized conjugate gradient methods, for nonsymmetric elliptic boundary value problems. Conditions for the existence of the preconditioning matrix can be derived easily for these problems.

Acknowledgements.

I wish to thank Henk van der Vorst for his criticism and inspiring ideas, and Guus Segal for stimulating discussions and assistance with the programming of the methods used.

REFERENCES

[1] Axelsson, O., *A survey of preconditioned iterative methods for linear systems of algebraic equations*, BIT 25 (1985), pp. 166-187.

[2] Axelsson, O. and V.A. Barker, *Finite Element Solution of Boundary Value Problems*, Academic Press, New York (1984).

[3] Axelsson, O. and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numerische Mathematik 48 (1986), pp. 499-523.

[4] Ciarlet, P.G., *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam (1978).

[5] Concus, P., G.H. Golub and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM Journal of Scientific and Statistical Computation 6 (1985), pp. 220-252.

[6] Eisenstat, S.C., *Efficient implementation of a class of preconditioned conjugate gradient methods*, SIAM Journal of Scientific and Statistical Computation 2 (1981), pp. 1-4.

[7] Golub, G.H. and C.F. van Loan, *Matrix Computations*, North Oxford Academic, Oxford (1983).

[8] Gustafsson, I., *Modified incomplete Cholesky (MIC) methods*, in: D.J. Evans (ed.), *Preconditioning Methods, Theory and Applications*, Gordon and Breach, New York (1983), pp. 265-293.

[9] Gustafsson, I. and G. Lindskog, *A preconditioning technique based on element matrix factorizations*, Computer Methods in Applied Mechanics and Engineering 55 (1986), pp. 201-220.

[10] Hageman, L.A. and D.M. Young, *Applied Iterative Methods*, Academic Press, New York (1981).

[11] Hestenes, M.R., *Conjugate Direction Methods in Optimization*, Springer, Berlin (1980).

[12] Hestenes, M.R. and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards 49 (1952), pp. 409-436.

[13] Hughes, T.J.R., I. Levit and J. Winget, *An element-by-element solution algorithm for problems of structural and solid mechanics*, Computer Methods in Applied Mechanics and Engineering 36 (1983), pp. 241-254.

[14] Kaasschieter, E.F., *A practical termination criterion for the conjugate gradient method*, BIT 28 (1988), pp. 308-322.

[15] Kaasschieter, E.F., *Preconditioned conjugate gradients for solving singular systems*, Journal of Computational and Applied Mathematics 24 (1988), pp. 265-275.

[16] Liu, W.H. and A.H. Sherman, *Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordening algorithms for sparse matrices*, SIAM Journal of Numerical Analysis 13 (1976), pp. 198-213.

[17] Meijerink, J.A. and H.A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Mathematics of Computation 31 (1977), pp. 148-162.

[18] Meijerink, J.A. and H.A. van der Vorst, *Guidelines for the usage of incomplete decompositions in solving sets of linear systems as they occur in practical problems*, Journal of Computational Physics 44 (1981), pp. 131-155.

[19] Nour-Omid, B. and B.N. Parlett, *Element preconditioning using splitting techniques*, SIAM Journal of Scientific and Statistical Computation 6 (1985), pp. 761-770.

[20] Schwarz, H.R., *The Finite Element Method*, Academic Press, New York (1988).

[21] van der Sluis, A. and H.A. van der Vorst, *The rate of convergence of conjugate gradients*, Numerische Mathematik 48 (1986), pp. 543-560.

[22] Varga, R.S., *On recurring theorems on diagonal dominance*, Linear Algebra and Its Applications 13 (1976), pp. 1-9.

CHAPTER IV

MIXED-HYBRID FINITE ELEMENTS AND STREAMLINE COMPUTATION FOR THE POTENTIAL FLOW PROBLEM

E.F. Kaasschieter

TNO Institute of Applied Geoscience, P.O. Box 285,
2600 AG Delft, The Netherlands

A.J.M. Huijben*

Royal Netherlands Naval College, P.O. Box 10000,
1780 CA Den Helder, The Netherlands

Abstract

An important class of problems in mathematical physics involves equations of the form $-\nabla\cdot(A\nabla\phi) = f$. In a variety of problems it is desirable to obtain an accurate approximation of the flow quantity $u = -A\nabla\phi$. Such an accurate approximation can be determined by the mixed finite element method. In this paper the lowest order mixed method is discussed in detail.

The mixed finite element method results in a large system of linear equations with an indefinite coefficient matrix. This drawback can be circumvented by the hybridization technique, which leads to a symmetric positive definite system. This system can be solved efficiently by the preconditioned conjugate gradient method.

After approximating u by the lowest order mixed finite element method, streamlines and residence times can be determined easily and accurately by computations at the element level.

AMS (MOS) subject classifications: 35 J 25, 65 F 10, 65 F 50, 65 N 30.

Keywords: Potential flow problem, mixed-hybrid finite elements, streamline computation, preconditioned conjugate gradients.

* Present address: Philips, CFT Automation, P.O. Box 218, 5600 MD Eindhoven, The Netherlands.

1. Introduction.

An important class of problems in mathematical physics involves equations of the form

(1.1a) $\boldsymbol{u} = -A\nabla\phi$,

(1.1b) $\nabla\cdot\boldsymbol{u} = f$,

where A is a symmetric and uniformly positive definite second rank tensor. The equations (1.1) are fundamental in the theory of heat conduction, electrostatics and groundwater hydraulics.

For instance, in modelling the flow of an incompressible fluid in a saturated porous medium, the piezometric head (potential) ϕ and the specific discharge (Darcy velocity) $\boldsymbol{u}$ are related by Darcy's law (equation (1.1a)), where A is the tensor of hydraulic conductivity (permeability), and $\boldsymbol{u}$ has to fulfil the continuity equation (equation (1.1b)). The function f is used to represent sources and sinks (see, e.g., [4: chapter 5]).

An accurate approximation of the specific discharge is crucial in the numerical solution of a variety of groundwater flow problems. In approximating $\boldsymbol{u}$ from (1.1) by standard finite difference or finite element techniques, first an approximation of ϕ is determined as a set of cell averages, nodal values or piecewise smooth functions. This approximation of ϕ is then numerically differentiated and multiplied by an often rough tensor A to obtain an approximation of $\boldsymbol{u}$. In many cases an inaccurate specific discharge results from this approach, i.e. the approximation thus obtained does not fulfil the continuity equation (1.1b) sufficiently well.

The mixed-hybrid finite element method

In a physical context it is desirable to obtain an approximation of $\boldsymbol{u}$, that fulfils (1.1b) as well as possible with respect to the finite difference grid or finite element mesh. Such an approximation can be determined by the mixed finite element method, which Raviart and Thomas proposed for two-dimensional problems and Nedelec proposed for three-dimensional problems (see [29], [32], [26], [27], [30: chapter IV]). In this paper the mixed finite element method will be discussed for general two- and three-dimensional problems. Only the lowest order mixed method will be considered, firstly, because higher order methods result in some conceptual complications and, secondly, because the lowest order method is comparatively easy and straightforward to use for practical problems (see [11], [8: chapter V], [18]).

The mixed finite element method results in a large system of linear equations. The choice of a numerical method to solve this system is restricted by the fact that its coefficient matrix is indefinite. This drawback can be circumvented by an implementation technique called hybridization, which leads to a symmetric positive definite system of linear equations (see [1]). Since this system is sparse, it can be solved efficiently by the preconditioned conjugate gradient method (see, e.g., [12: chapter 10]).

Unfortunately, the available literature on the mixed-hybrid finite element method is very theoretical, and this hampers its application. Several important aspects are discussed implicitly or are not discussed explicitly for general inhomogeneous problems of the form (1.1). Therefore, it is difficult to apply this method to practical problems. Sections 2 to 8 inclusive aim to give a straightforward and integrated presentation of the mixed-hybrid finite element method. Hopefully, this will facilitate its application.

Computation of streamlines and residence times

In groundwater hydraulics, streamlines and residence times induced by the specific discharge $\boldsymbol{u}$ are important secondary quantities. Given the velocity $\boldsymbol{w} = \boldsymbol{u}/p$, where p is the porosity, a streamline is defined as a curve that is everywhere tangential to the velocity. The time required by a water particle to flow along a streamline from one point to another is called the residence time between these two points. In modelling two-dimensional groundwater flow problems streamlines can be obtained easily as contour lines of the stream function (see [4: section 5-8]). However, branch cuts have to be introduced if there are sources or sinks in the flow domain. Furthermore, one cannot obtain residence times directly from the stream function. Moreover, when modelling three-dimensional groundwater flow the concept of a stream function is complicated.

Fortunately, streamlines and residence times can be determined easily and accurately for two- and three-dimensional problems with or without sources and sinks, using the lowest order mixed finite element method. After the computation of the approximation u_h of u the approximate velocity $w_h = u_h/p$ can be determined. Streamlines and residence times can be determined from w_h by computations at the element level. This approach allows the streamlines and residence times to be determined exactly with respect to the approximate velocity w_h.

Outline of the paper

The remainder of this paper is organized as follows. In section 2 we define notations and give some preliminary results. A mixed variational formulation of (1.1) is stated in section 3. Section 4 deals with the change of variables, in preparation for the definition of the local basis functions in section 5. The lowest order mixed finite element method is treated in section 6 and its hybridization in section 7. The assembly of the resulting system of linear equations and its solution are discussed in section 8. In section 9 the elementwise computation of streamlines and residence times is exposed. In section 10 the applicability and advantages of the mixed finite element method and the efficient solution of the resulting system of linear equations are illustrated by several numerical experiments. Final conclusions are drawn in section 11.

2. Notations and preliminaries.

In this section some functional spaces and some preliminary results are introduced (for details, see [32: chapter II]).

Throughout this paper, Ω shall denote an open simply connected domain in $\mathbb{R}^d$ ($d = 2$ or $d = 3$) with a Lipschitz-continuous boundary $\partial\Omega$ (see [9: page 12]). Let $\partial\Omega_D$ and $\partial\Omega_N$ be measurable portions of $\partial\Omega$, such that $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ and $\overline{\partial\Omega_D} \cup \overline{\partial\Omega_N} = \partial\Omega$ (in the next section Dirichlet and Neumann boundary conditions will be defined on $\partial\Omega_D$ and $\partial\Omega_N$, respectively).

The Lebesgue space $L^2(\Omega)$ contains the square-integrable scalar functions on Ω, i.e.

$$(2.1a) \qquad L^2(\Omega) = \left\{\phi\colon \Omega \to \mathbb{R} \;\middle|\; \int_\Omega \phi^2 \, dx < \infty\right\}.$$

$L^2(\Omega)$ is a Hilbert space with respect to the norm

$$(2.1b) \qquad \|\phi\|_{0,\Omega} = \left[\int_\Omega \phi^2 \, dx\right]^{1/2}, \; \phi \in L^2(\Omega).$$

The Lebesgue space $\mathbf{L}^2(\Omega)$ is analogously defined as

$$(2.2a) \qquad \mathbf{L}^2(\Omega) = \left\{\mathbf{u}\colon \Omega \to \mathbb{R}^d \;\middle|\; \int_\Omega \|\mathbf{u}\|_2^2 \, dx < \infty\right\}.$$

$\mathbf{L}^2(\Omega)$ is a Hilbert space with respect to the norm

$$(2.2b) \qquad \|\mathbf{u}\|_{0,\Omega} = \left[\int_\Omega \|\mathbf{u}\|_2^2 \, dx\right]^{1/2}, \; \mathbf{u} \in \mathbf{L}^2(\Omega).$$

In order to state variational formulations of the equations (1.1) we have to consider functions that are differentiable in the weak sense (see, e.g., [15: definition 6.2.3]). Spaces containing these kinds of functions are the Sobolev spaces. Note that throughout this chapter differentiation is understood in the weak sense.

The Sobolev space $H^1(\Omega)$ contains the square-integrable scalar functions whose gradients are also square-integrable, i.e.

(2.3a) $$H^1(\Omega) = \{\phi \in L^2(\Omega) \mid \nabla\phi \in L^2(\Omega)\}.$$

$H^1(\Omega)$ is a Hilbert space with respect to the norm

(2.3b) $$\|\phi\|_{1,\Omega} = [\|\phi\|^2_{0,\Omega} + \|\nabla\phi\|^2_{0,\Omega}]^{1/2}, \phi \in H^1(\Omega).$$

If $\phi \in H^1(\Omega)$, then the trace $\gamma_D\phi = \phi|_{\partial\Omega}$ is well defined, and we denote

(2.4a) $$H^{1/2}(\partial\Omega) = \{\gamma_D\phi \mid \phi \in H^1(\Omega)\}$$

with the norm

(2.4b) $$\|\lambda\|_{1/2,\partial\Omega} = \inf_{\phi\in H^1(\Omega)} \{\|\phi\|_{1,\Omega} \mid \lambda = \gamma_D\phi\}, \lambda \in H^{1/2}(\partial\Omega).$$

The linear subspaces $H^1_D(\Omega)$ and $H^{1/2}_D(\partial\Omega)$ are defined as

(2.5a) $$H^1_D(\Omega) = \{\phi \in H^1(\Omega) \mid \phi = 0 \text{ on } \partial\Omega_D\},$$

(2.5b) $$H^{1/2}_D(\partial\Omega) = \{\lambda \in H^{1/2}(\partial\Omega) \mid \lambda = 0 \text{ on } \partial\Omega_D\}.$$

Let $g_D \in H^{1/2}(\partial\Omega)$, then the linear variety $H^1_*(\Omega)$ is defined as

(2.6) $$H^1_*(\Omega) = \{\phi \in H^1(\Omega) \mid \phi = g_D \text{ on } \partial\Omega_D\}.$$

The Sobolev space $H(\text{div};\Omega)$ contains the square-integrable vectorial functions whose divergences are also square-integrable, i.e.

(2.7a) $$H(\text{div};\Omega) = \{\boldsymbol{u} \in L^2(\Omega) \mid \nabla\cdot\boldsymbol{u} \in L^2(\Omega)\}.$$

$H(\mathrm{div};\Omega)$ is a Hilbert space with respect to the norm

$$\text{(2.7b)} \qquad \|\boldsymbol{u}\|_{\mathrm{div},\Omega} = \left[\|\boldsymbol{u}\|^2_{0,\Omega} + \|\nabla\cdot\boldsymbol{u}\|^2_{0,\Omega}\right]^{1/2}, \ \boldsymbol{u} \in H(\mathrm{div};\Omega).$$

If $\boldsymbol{u} \in H(\mathrm{div};\Omega)$, then the trace $\gamma_N \boldsymbol{u} = \boldsymbol{n}\cdot\boldsymbol{u}\big|_{\partial\Omega}$, where $\boldsymbol{n}$ is the outward normal to $\partial\Omega$, is well defined, and we denote

$$\text{(2.8a)} \qquad H^{-1/2}(\partial\Omega) = \{\gamma_N \boldsymbol{u} \,|\, \boldsymbol{u} \in H(\mathrm{div};\Omega)\}$$

with the norm

$$\text{(2.8b)} \qquad \|\mu\|_{-1/2,\partial\Omega} = \inf_{\boldsymbol{u}\in H(\mathrm{div};\Omega)} \{\|\boldsymbol{u}\|_{\mathrm{div},\Omega} \,|\, \mu = \gamma_N \boldsymbol{u}\}, \ \mu \in H^{1/2}(\partial\Omega).$$

The linear subspaces $H_N(\mathrm{div};\Omega)$ and $H_N^{-1/2}(\partial\Omega)$ are defined as

$$\text{(2.9a)} \qquad H_N(\mathrm{div};\Omega) = \{\boldsymbol{u} \in H(\mathrm{div};\Omega) \,|\, \boldsymbol{n}\cdot\boldsymbol{u} = 0 \text{ on } \partial\Omega_N\},$$

$$\text{(2.9b)} \qquad H_N^{-1/2}(\partial\Omega) = \{\mu \in H^{-1/2}(\partial\Omega) \,|\, \mu = 0 \text{ on } \partial\Omega_N\}.$$

Let $g_N \in H^{-1/2}(\partial\Omega)$, then the linear variety $H_*(\mathrm{div};\Omega)$ is defined as

$$\text{(2.10)} \qquad H_*(\mathrm{div};\Omega) = \{\boldsymbol{u} \in H(\mathrm{div};\Omega) \,|\, \boldsymbol{n}\cdot\boldsymbol{u} = g_N \text{ on } \partial\Omega_N\}.$$

NOTE 2.1: The Sobolev spaces defined above are related by the following duality properties (for a proof, see [32: proposition II-2.1]):

$$\text{(2.10a)} \qquad \|\lambda\|_{1/2,\partial\Omega} = \sup_{\boldsymbol{u}\in H(\mathrm{div};\Omega)\setminus\{0\}} \frac{\int_{\partial\Omega} \lambda\, \boldsymbol{n}\cdot\boldsymbol{u}\, ds}{\|\boldsymbol{u}\|_{\mathrm{div},\Omega}} \quad \forall\lambda \in H^{1/2}(\partial\Omega),$$

$$\text{(2.10b)} \qquad \|\mu\|_{-1/2,\partial\Omega} = \sup_{\phi\in H^1(\Omega)\setminus\{0\}} \frac{\int_{\partial\Omega} \mu\phi\, ds}{\|\phi\|_{1,\Omega}} \quad \forall\mu \in H^{-1/2}(\partial\Omega).$$

From these duality properties it follows immediately that $H^{1/2}(\partial\Omega)$ and $H^{-1/2}(\partial\Omega)$ are dual spaces, i.e.

$$\text{(2.11a)} \qquad \|\lambda\|_{1/2,\partial\Omega} = \sup_{\mu\in H^{-1/2}(\partial\Omega)\backslash\{0\}} \frac{\int_{\partial\Omega} \lambda\mu \, ds}{\|\mu\|_{-1/2,\partial\Omega}} \quad \forall\lambda \in H^{1/2}(\partial\Omega),$$

$$\text{(2.11b)} \qquad \|\mu\|_{-1/2,\partial\Omega} = \sup_{\lambda\in H^{1/2}(\partial\Omega)\backslash\{0\}} \frac{\int_{\partial\Omega} \mu\lambda \, ds}{\|\lambda\|_{1/2,\partial\Omega}} \quad \forall\mu \in H^{-1/2}(\partial\Omega). \square$$

3. A mixed variational formulation.

We shall concentrate on the following elliptic boundary value problem (see also [9: (1.2.28)]):

$$\text{(3.1)} \qquad \boxed{\begin{array}{l} -\nabla\cdot(A\nabla\phi) = f \text{ in } \Omega, \\ \phi = g_D \text{ on } \partial\Omega_D, \ -\boldsymbol{n}\cdot(A\nabla\phi) = g_N \text{ on } \partial\Omega_N, \end{array}}$$

where $f \in L^2(\Omega)$, $g_D \in H^{1/2}(\partial\Omega)$, $g_N \in H^{-1/2}(\partial\Omega)$ and $A \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ is symmetric and uniformly positive definite, i.e. $(A\xi)\cdot\xi \geq \alpha \|\xi\|_2^2$ almost everywhere in Ω for all $\xi \in \mathbb{R}^d$ for a constant $\alpha > 0$, that does not depend on $x \in \Omega$. Again, the vector $\boldsymbol{n}$ is the outward normal to $\partial\Omega$.

If the measure of $\partial\Omega_D \subset \partial\Omega$ is positive, then a function $\phi \in H^1_*(\Omega)$ (see (2.6)) is the unique weak solution of (3.1) if

$$\text{(3.2)} \qquad \int_\Omega (A\nabla\phi)\cdot\nabla\psi \, dx = \int_\Omega f\psi \, dx - \int_{\partial\Omega} g_N \, \psi \, ds \quad \forall\psi \in H^1_D(\Omega)$$

(see, e.g., [9: (1.2.26)]). Since $f \in L^2(\Omega)$ it follows from (3.2), using Green's formula, that

$$(3.3)\qquad -\int_\Omega \nabla\cdot(A\nabla\phi)\,\psi\,dx = \int_\Omega f\psi\,dx \quad \forall\psi\in L^2(\Omega)$$

(note that $H^1_D(\Omega)$ is a dense subspace of $L^2(\Omega)$). Using Green's formula again, the mixed variational formulation of (3.1) can be stated (see [32: IX-(1.4)]):

Find $(\boldsymbol{u},\phi)\in H_*(\mathrm{div};\Omega)\times L^2(\Omega)$, such that

$$(3.4\mathrm{a})\qquad \int_\Omega (C\boldsymbol{u})\cdot\boldsymbol{v}\,dx - \int_\Omega \phi\,\nabla\cdot\boldsymbol{v}\,dx = -\int_{\partial\Omega} g_D\,\boldsymbol{n}\cdot\boldsymbol{v}\,ds \quad \forall\boldsymbol{v}\in H_N(\mathrm{div};\Omega),$$

$$(3.4\mathrm{b})\qquad -\int_\Omega \nabla\cdot\boldsymbol{u}\,\psi\,dx = -\int_\Omega f\psi\,dx \quad \forall\psi\in L^2(\Omega),$$

where $C = A^{-1}$ is the compliance tensor (for the definition of $H_*(\mathrm{div};\Omega)$, see (2.10)).

The problem (3.4) has a unique solution $(\boldsymbol{u},\phi)$. Moreover, $\phi\in H^1_*(\Omega)$ is the solution of (3.2) and $\boldsymbol{u} = -A\nabla\phi$ in Ω (see [32: theorem IX-1.1]).

NOTE 3.1: The reason for choosing this variational formulation of problem (3.1) is the necessity to have $\boldsymbol{u}$ as well as $\nabla\cdot\boldsymbol{u}$ in the basic formulation. In this case approximations of $\boldsymbol{u}$ and $\nabla\cdot\boldsymbol{u}$ can be obtained directly from the corresponding mixed finite element method (see section 6). Only then can the continuity equation (1.1b) be fulfilled as well as possible with respect to the finite element mesh. □

NOTE 3.2: An essential condition when proving that problem (3.4) has a unique solution is the inf-sup condition (also called the Babuška-Brezzi condition)

$$(3.5)\qquad \inf_{\phi\in L^2(\Omega)\setminus\{0\}}\ \sup_{\boldsymbol{v}\in H_N(\mathrm{div};\Omega)\setminus\{0\}} \frac{\int_\Omega \phi\,\nabla\cdot\boldsymbol{v}\,dx}{\|\phi\|_{0,\Omega}\,\|\boldsymbol{v}\|_{\mathrm{div},\Omega}} \ge \beta,$$

where the constant $\beta > 0$ only depends on Ω (for a proof of this condition, see [32: theorem IX-1.1]). The existence and uniqueness of the solution of (3.4) then follows, using [5: theorem 1.1]. □

NOTE 3.3: If $f = 0$ almost everywhere in Ω, then it follows from (3.4b) that $\nabla\cdot\boldsymbol{u} = 0$ in Ω. In this case the dual variational formulation of (3.1) can be stated. Define the space

$$(3.6) \qquad X(\Omega) = \{\boldsymbol{u} \in H(\mathrm{div};\Omega) \mid \nabla\cdot\boldsymbol{u} = 0 \text{ in } \Omega\}.$$

$X(\Omega)$ is a Hilbert space with respect to the norm $\|\cdot\|_{0,\Omega}$. The linear subspace $X_N(\Omega)$ and the linear variety $X_*(\Omega)$ are defined as

$$(3.7a) \qquad X_N(\Omega) = X(\Omega) \cap H_N(\mathrm{div};\Omega),$$
$$(3.7b) \qquad X_*(\Omega) \ = X(\Omega) \cap H_*(\mathrm{div};\Omega).$$

From (3.4a) it follows that $\boldsymbol{u} \in X_*(\Omega)$ is the unique solution of the following problem:

Find $\boldsymbol{u} \in X_*(\Omega)$, such that

$$(3.8) \qquad \int_\Omega (C\boldsymbol{u})\cdot\boldsymbol{v}\, dx = -\int_{\partial\Omega} g_D\, \boldsymbol{n}\cdot\boldsymbol{v}\, ds \quad \forall \boldsymbol{v} \in X_N(\Omega). \ \square$$

NOTE 3.4: If $d = 2$ and $f = 0$ almost everywhere in Ω, then $\boldsymbol{u}$ can be obtained from the stream function Ψ as $\boldsymbol{u} = -\nabla \times \Psi = -\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \nabla\Psi$ in Ω. In order to prove this assertion, assume that $\boldsymbol{u} \in X_*(\Omega)$ is the unique solution of problem (3.8). Let s be the anti-clockwise tangent to $\partial\Omega$, then $\boldsymbol{n} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} s$ almost everywhere on $\partial\Omega$. Let $\overline{\partial\Omega_D} = \bigcup_{k=1}^{K} \partial\Omega_k$, where $\partial\Omega_k$ are closed simply connected portions of $\overline{\partial\Omega_D}$, which are mutually disjoint. Choose $G_D \in H^{-1/2}(\partial\Omega)$, such that $G_D = \dfrac{\partial g_D}{\partial s}$ on $\partial\Omega_D$, and $G_N \in H^{1/2}(\partial\Omega)$, such that $g_N = \dfrac{\partial G_N}{\partial s}$ on $\partial\Omega_N$ and $G(b_k) - G(a_k) = \int_{\partial\Omega_k} \boldsymbol{n}\cdot\boldsymbol{u}\, ds$, $k = 1,\ldots,K$, where a_k and b_k are respectively the begin and end points of $\partial\Omega_k$ (in the anti-clockwise notion). Note that G_N is unique up to a constant, because $\int_{\partial\Omega} \boldsymbol{n}\cdot\boldsymbol{u}\, ds = 0$. Consider the following boundary value problem:

$$\text{(3.9)} \quad \boxed{\begin{array}{l} -\nabla\cdot(C'\nabla\Psi) = 0 \text{ in } \Omega, \\ \Psi = G_N \text{ on } \partial\Omega_N,\ -\boldsymbol{n}\cdot(C'\nabla\Psi) = G_D \text{ on } \partial\Omega_D, \end{array}}$$

where $C' = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} C \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Define the linear subspace $Y_N(\Omega)$ and the linear variety $Y_*(\Omega)$ as

$$\text{(3.10a)} \quad Y_N(\Omega) = \{\Psi \mid \Psi \in H^1(\Omega),\ \Psi = 0 \text{ on } \partial\Omega_N\},$$
$$\text{(3.10b)} \quad Y_*(\Omega) = \{\Psi \mid \Psi \in H^1(\Omega),\ \Psi = G_N \text{ on } \partial\Omega_N\}.$$

If the measure of $\partial\Omega_N \subset \partial\Omega$ is positive, then a function $\Psi \in Y_*(\Omega)$ is the unique weak solution of (3.9) if

$$\text{(3.11)} \quad \int_\Omega (C'\nabla\Psi)\cdot\nabla\Phi\, dx = -\int_{\partial\Omega} G_D\, \Phi\, ds \quad \forall\Phi \in Y_N(\Omega).$$

Define $\tilde{\boldsymbol{u}} = -\nabla \times \Psi$ in Ω, then it follows that $\tilde{\boldsymbol{u}} \in X_*(\Omega)$ fulfils (3.8). Therefore, $\tilde{\boldsymbol{u}} = \boldsymbol{u}$ is the unique solution of (3.8). □

4. Change of variables.

In the next section we will define the basis functions for the approximations of $\boldsymbol{u}$ and ϕ restricted to a finite element. As usual these basis functions can be defined explicitly on a reference element. For an arbitrary element they follow from an affine transformation. Therefore we have to describe the transformations of scalar and vectorial functions from one domain to another. The well-known transformation rules corresponding to the covariant and contravariant vector formalism (see, e.g., [24: section 1.5]) will be used.

For the sake of simplicity, in the arguments of various functional spaces we shall not distinguish between $\hat{S}$ (and S) and its corresponding interior.

Let $\hat{S}$ be a closed simply connected domain in $\mathbb{R}^d$ ($d = 2$ or $d = 3$) with a Lipschitz-continuous boundary $\partial\hat{S}$. Further, let $F \in C^1(\hat{S}, \mathbb{R}^d)$ be an isomorphism onto $S = F(\hat{S})$. The point $x \in S$ is the image of $\hat{x} \in \hat{S}$ obtained by the mapping F, i.e.

(4.1) $$x = F(\hat{x}),\ \hat{x} \in \hat{S}.$$

The functional matrix $B \in C(\hat{S}, \mathbb{R}^{dxd})$ of F is given by

(4.2) $$B(\hat{x}) = [\frac{\partial}{\partial \hat{x}_j} F_i(\hat{x})]_{i,j=1,\dots,d},\ \hat{x} \in \hat{S},$$

and the functional determinant or Jacobian $J \in C(\hat{S})$ of F by

(4.3) $$J(\hat{x}) = \det B(\hat{x}),\ \hat{x} \in \hat{S}.$$

Suppose that

(4.4) $$J(\hat{x}) > 0 \quad \forall \hat{x} \in \hat{S}.$$

With any scalar function $\hat{\phi} \in L^2(\hat{S})$ the function $\phi \in L^2(S)$ is associated, where

(4.5) $$\phi(x) = \hat{\phi}(\hat{x}),\ \hat{x} \in \hat{S}.$$

It is well-known that

(4.6) $$\int_S \phi\, dx = \int_{\hat{S}} \hat{\phi}\, J\, d\hat{x} \quad \forall \hat{\phi} \in L^2(\hat{S}).$$

If $\hat{\phi} \in H^1(\hat{S})$, then $\phi \in H^1(S)$ and

$$(4.7) \qquad \nabla\phi(x) = B(\hat{x})^{-T}\, \nabla\hat{\phi}(\hat{x}) \quad \forall\hat{x} \in \hat{S}.$$

With any vectorial function $\hat{u} \in L^2(\hat{S})$ the function $u \in L^2(S)$ is associated, where

$$(4.8) \qquad u(x) = B(\hat{x})\, \hat{u}(\hat{x})/J(\hat{x}),\ \hat{x} \in \hat{S}.$$

With these choices the following equalities hold (for a proof, see [32: propositions II-5.2 and II-5.4]):

$$(4.9a) \qquad \int_S u\cdot\nabla\, \phi\, dx = \int_{\hat{S}} \hat{u}\cdot\nabla\, \hat{\phi}\, d\hat{x} \quad \forall\hat{u} \in L^2(\hat{S})\ \forall\hat{\phi} \in H^1(\hat{S}),$$

$$(4.9b) \qquad \int_S \phi\, \nabla\cdot u\, dx = \int_{\hat{S}} \hat{\phi}\, \nabla\cdot\hat{u}\, d\hat{x} \quad \forall\hat{\phi} \in L^2(S)\ \forall\hat{u} \in H(\text{div};\hat{S}),$$

$$(4.9c) \qquad \int_{\partial S} \phi\, n_S\cdot u\, ds = \int_{\partial\hat{S}} \hat{\phi}\, n_{\hat{S}}\cdot\hat{u}\, d\hat{s} \quad \forall\hat{\phi} \in H^1(\hat{S})\ \forall\hat{u} \in H(\text{div};\hat{S}),$$

where n_S and $n_{\hat{S}}$ are the outward normals to ∂S and $\partial\hat{S}$, respectively.

From (4.9b) it follows immediately that, if $\hat{u} \in H(\text{div};\hat{S})$, then $u \in H(\text{div};S)$ and

$$(4.10) \qquad \nabla\cdot u(x) = \nabla\cdot\hat{u}(\hat{x})/J(\hat{x}) \quad \forall\hat{x} \in \hat{S}.$$

5. Local basis functions.

Let the reference element $\hat{S}$ be the convex hull of L suitably chosen points $\hat{x}_\ell$, $\ell = 1,...,L$, i.e. $\hat{S} = \left\{\hat{x} = \sum_{\ell=1}^{L} \zeta_\ell \hat{x}_\ell \,|\, 0 \le \zeta_\ell \le 1,\ \sum_{\ell=1}^{L} \zeta_\ell = 1\right\}$. Let $B \in \mathbb{R}^{d\times d}$, such that $J = \det B > 0$, and $b \in \mathbb{R}^d$. The mapping $F \in C^1(\hat{S},\mathbb{R}^{d\times d})$ is defined as

$$(5.1) \qquad F(\hat{x}) = B\hat{x} + b,\ \hat{x} \in \hat{S}.$$

Thus, F is an isomorphism onto an element $S \in F(\hat{S})$. Again, the point $\hat{x} \in S$ is the image of $\hat{x} \in \hat{S}$ by the mapping F, i.e.

$$(5.2) \qquad x = Bx + b,\ \hat{x} \in \hat{S}.$$

Now, B is the functional matrix of F and $J = \det B$ is its Jacobian.

The simplest choice for a finite element formulation of the variational problem (3.4) is to approximate u by a piecewise linear function, such that its normal component on each edge ($d = 2$) or face ($d = 3$) of the finite element mesh is constant. Therefore, let $\hat{e}_i$, $i = 1,\ldots,I$, be the edges ($d = 2$) or faces ($d = 3$) of $\hat{S}$, and $RT^0(\hat{S})$ be the I-dimensional space of linear vectorial functions $\hat{u}$ on $\hat{S}$, such that $n_{\hat{S}}\cdot\hat{u}$ is constant on $\hat{e}_i$, $i = 1,\ldots,I$. Its basis functions are $\hat{v}_i$, $i = 1,\ldots,I$, such that

$$(5.3) \qquad \int_{\hat{e}_j} n_{\hat{S}}\cdot\hat{v}_i \, d\hat{s} = \delta_{ij},\ i,j = 1,\ldots,I.$$

With any vectorial basis function $\hat{v}_i$, the basis function v_i is associated, where

$$(5.4) \qquad v_i(x) = B\hat{v}_i(\hat{x})/J,\ \hat{x} \in \hat{S}$$

(see (4.8)). Let $RT^0(S)$ be the space spanned by the basis functions v_i, $i = 1,\ldots,I$. Clearly, $RT^0(S)$ is the I-dimensional space of linear vectorial functions u on S, such that $n_S\cdot u$ is constant on e_i, $i = 1,\ldots,I$ (use (4.9c)), and also

$$(5.5) \qquad RT^0(S) = \{u \mid u(x) = B\hat{u}(\hat{x})/J \ \ \forall\hat{x} \in \hat{S},\ \hat{u} \in RT^0(\hat{S})\}.$$

Further, ϕ is approximated by a piecewise constant function. Therefore, let $M^0(S)$ be the one-dimensional space of constant scalar functions on S. Its basis function is ψ, where $\psi(x) = 1$, $x \in S$.

The formal representation, and therefore also the numerical computation of the basis functions of $RT^0(S)$, simplifies if natural coordinates are used (see, e.g., [31: section 2.3.1]). Let $x_\ell = B\hat{x}_\ell + b$, $\ell = 1,\dots,L$, then $S = \left\{x = \sum_{\ell=1}^{L} \zeta_\ell x_\ell \,|\, 0 \le \zeta_\ell \le 1, \sum_{\ell=1}^{L} \zeta_\ell = 1\right\}$. For each $x \in S$ the natural coordinates $\zeta_\ell(x)$, $\ell = 1,\dots,L$, are such that

$$(5.6) \qquad x = \sum_{\ell=1}^{L} \zeta_\ell(x)\, x_\ell,\ 0 \le \zeta_\ell(x) \le 1,\ \sum_{\ell=1}^{L} \zeta_\ell(x) = 1.$$

Note that the functions ζ_ℓ are the local basis functions on S corresponding to the conforming finite element method for the variational problem (3.2) (see [9: section 2.2]).

An overview of some possible choices for the reference element $\hat{S}$ and the basis functions of $RT^0(\hat{S})$ and $RT^0(S)$ is given below.

Triangle ($d = 2$, $L = 3$, $I = 3$): See figure 1a.

$\hat{x}_\ell$: $\binom{0}{0}$, $\binom{1}{0}$, $\binom{0}{1}$.

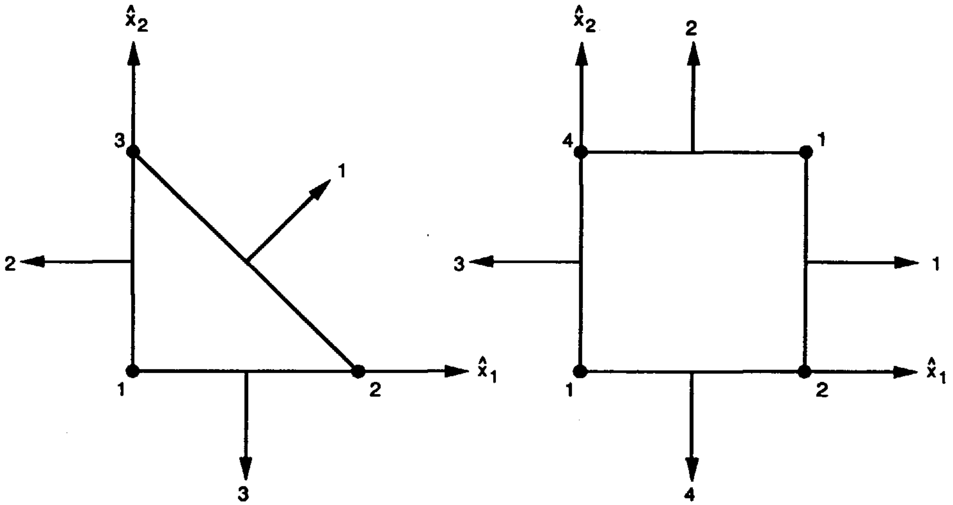
$$\hat{v}_i(\hat{x}): \quad \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}, \begin{bmatrix} \hat{x}_1 - 1 \\ \hat{x}_2 \end{bmatrix}, \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 - 1 \end{bmatrix}.$$

$B = [x_2 - x_1 \,|\, x_3 - x_1]$, $b = x_1$.
$v_1(x) = \{(x_2 - x_1)\,\zeta_2(x) + (x_3 - x_1)\,\zeta_3(x)\}/J = (x - x_1)/J$,
$v_2(x) = \{(x_1 - x_2)\,\zeta_1(x) + (x_3 - x_2)\,\zeta_3(x)\}/J = (x - x_2)/J$,
$v_3(x) = \{(x_1 - x_3)\,\zeta_1(x) + (x_2 - x_3)\,\zeta_2(x)\}/J = (x - x_3)/J$.

Parallelogram ($d = 2$, $L = 4$, $I = 4$): See figure 1b.

$\hat{x}_\ell$: $\binom{0}{0}$, $\binom{1}{0}$, $\binom{1}{1}$, $\binom{0}{1}$.

$$\hat{v}_i(\hat{x}): \quad \begin{bmatrix} \hat{x}_1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \hat{x}_2 \end{bmatrix}, \begin{bmatrix} \hat{x}_1 - 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \hat{x}_2 - 1 \end{bmatrix}.$$



(a)
triangle

(b)
parallelogram

Figure 1. Two-dimensional mixed reference elements.

$B = [x_2 - x_1 \mid x_4 - x_1]$, $b = x_1$.

$v_1(x) = (x_2 - x_1)\,[\zeta_2(x) + \zeta_3(x)]/J$,

$v_2(x) = (x_4 - x_1)\,[\zeta_3(x) + \zeta_4(x)]/J$,

$v_3(x) = (x_1 - x_2)\,[\zeta_1(x) + \zeta_4(x)]/J$,

$v_4(x) = (x_1 - x_4)\,[\zeta_1(x) + \zeta_2(x)]/J$.

Tetrahedron (d = 3, L = 4, I = 4): See figure 2a.

$$\hat{x}_\ell: \begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}.$$

$$\hat{v}_i(\hat{x})\text{: } 2\begin{bmatrix}\hat{x}_1\\ \hat{x}_2\\ \hat{x}_3\end{bmatrix},\ 2\begin{bmatrix}\hat{x}_1-1\\ \hat{x}_2\\ \hat{x}_3\end{bmatrix},\ 2\begin{bmatrix}\hat{x}_1\\ \hat{x}_2-1\\ \hat{x}_3\end{bmatrix},\ 2\begin{bmatrix}\hat{x}_1\\ \hat{x}_2\\ \hat{x}_3-1\end{bmatrix}.$$

$B = [x_2 - x_1 \,|\, x_3 - x_1 \,|\, x_4 - x_1],\ b = x_1.$

$v_1(x) = 2\,\{(x_2 - x_1)\,\zeta_2(x) + (x_3 - x_1)\,\zeta_3(x) + (x_4 - x_1)\,\zeta_4(x)\}/J = 2\,(x - x_1)/J,$

$v_2(x) = 2\,\{(x_1 - x_2)\,\zeta_1(x) + (x_3 - x_2)\,\zeta_3(x) + (x_4 - x_2)\,\zeta_4(x)\}/J = 2\,(x - x_2)/J,$

$v_3(x) = 2\,\{(x_1 - x_3)\,\zeta_1(x) + (x_2 - x_3)\,\zeta_2(x) + (x_4 - x_3)\,\zeta_4(x)\}/J = 2\,(x - x_3)/J,$

$v_4(x) = 2\,\{(x_1 - x_4)\,\zeta_1(x) + (x_2 - x_4)\,\zeta_2(x) + (x_3 - x_4)\,\zeta_3(x)\}/J = 2\,(x - x_4)/J.$

Prism ($d = 3$, $L = 6$, $I = 5$): See figure 2b.

$$\hat{x}_\ell\text{: } \begin{bmatrix}0\\0\\0\end{bmatrix},\ \begin{bmatrix}1\\0\\0\end{bmatrix},\ \begin{bmatrix}0\\1\\0\end{bmatrix},\ \begin{bmatrix}0\\0\\1\end{bmatrix},\ \begin{bmatrix}1\\0\\1\end{bmatrix},\ \begin{bmatrix}0\\1\\1\end{bmatrix}.$$

$$\hat{v}_i(\hat{x})\text{: } \begin{bmatrix}\hat{x}_1\\ \hat{x}_2\\ 0\end{bmatrix},\ \begin{bmatrix}\hat{x}_1-1\\ \hat{x}_2\\ 0\end{bmatrix},\ \begin{bmatrix}\hat{x}_1\\ \hat{x}_2-1\\ 0\end{bmatrix},\ 2\begin{bmatrix}0\\ 0\\ \hat{x}_3\end{bmatrix},\ 2\begin{bmatrix}0\\ 0\\ \hat{x}_3-1\end{bmatrix}.$$

$B = [x_2 - x_1 \,|\, x_3 - x_1 \,|\, x_4 - x_1],\ b = x_1.$

$v_1(x) = \{(x_2 - x_1)\,[\zeta_2(x) + \zeta_5(x)] + (x_3 - x_1)\,[\zeta_3(x) + \zeta_6(x)]\}/J,$

$v_2(x) = \{(x_1 - x_2)\,[\zeta_1(x) + \zeta_4(x)] + (x_3 - x_2)\,[\zeta_3(x) + \zeta_6(x)]\}/J,$

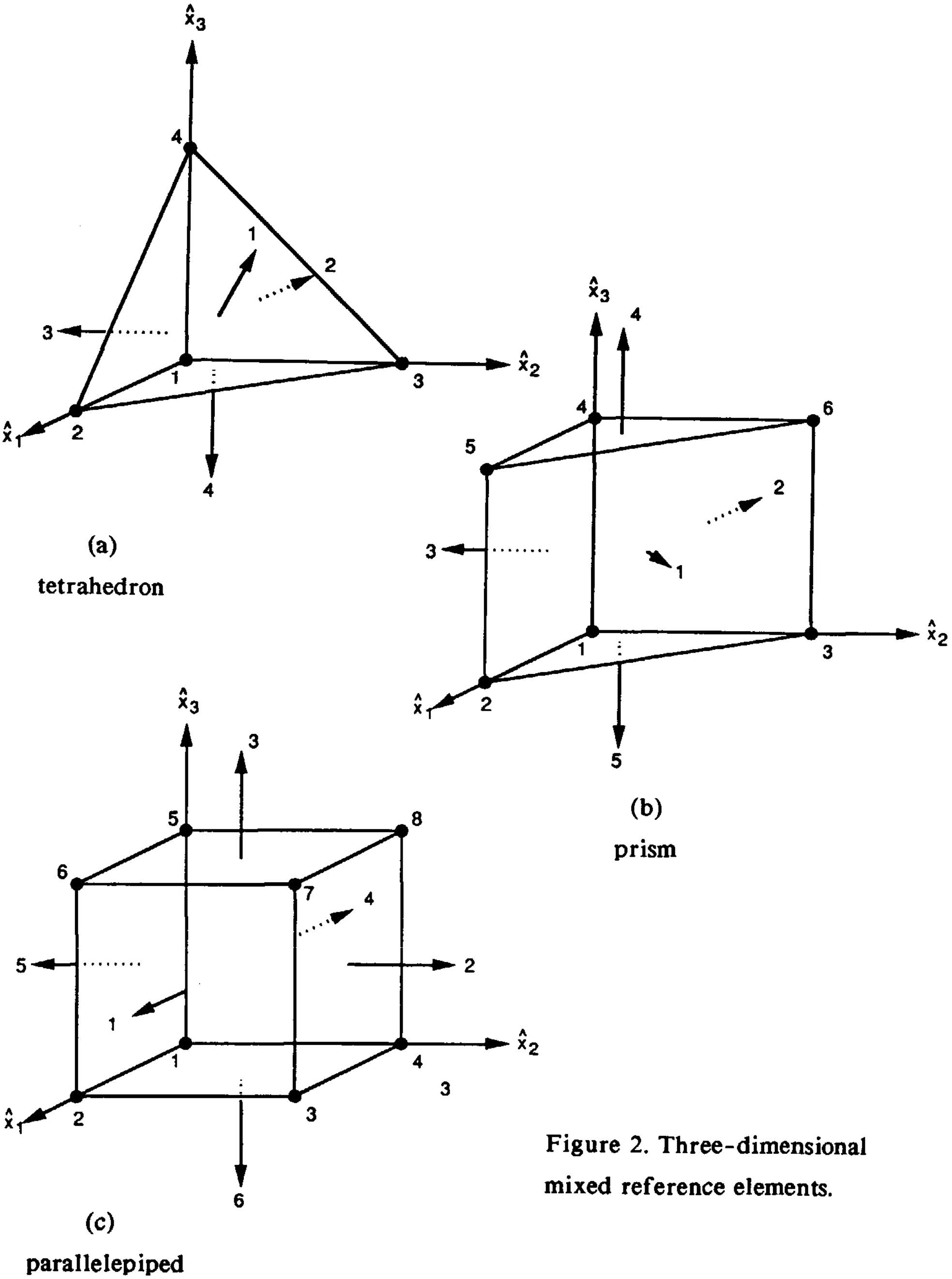
$v_3(x) = \{(x_1 - x_3)\,[\zeta_1(x) + \zeta_4(x)] + (x_2 - x_3)\,[\zeta_2(x) + \zeta_5(x)]\}/J,$

$v_4(x) = 2\,(x_4 - x_1)\,[\zeta_4(x) + \zeta_5(x) + \zeta_6(x)]/J,$

$v_5(x) = 2\,(x_1 - x_4)\,[\zeta_1(x) + \zeta_2(x) + \zeta_3(x)]/J.$

Parallelepiped ($d = 3$, $L = 8$, $I = 6$): See figure 2c.

$$\hat{x}_\ell\text{: } \begin{bmatrix}0\\0\\0\end{bmatrix},\ \begin{bmatrix}1\\0\\0\end{bmatrix},\ \begin{bmatrix}1\\1\\0\end{bmatrix},\ \begin{bmatrix}0\\1\\0\end{bmatrix},\ \begin{bmatrix}0\\0\\1\end{bmatrix},\ \begin{bmatrix}1\\0\\1\end{bmatrix},\ \begin{bmatrix}1\\1\\1\end{bmatrix},\ \begin{bmatrix}0\\1\\1\end{bmatrix}.$$



(a)
tetrahedron

(b)
prism

(c)
parallelepiped

Figure 2. Three-dimensional mixed reference elements.

$$\hat{\boldsymbol{v}}_i(\hat{x}):\ \begin{bmatrix}\hat{x}_1\\ 0\\ 0\end{bmatrix},\ \begin{bmatrix}0\\ \hat{x}_2\\ 0\end{bmatrix},\ \begin{bmatrix}0\\ 0\\ \hat{x}_3\end{bmatrix},\ \begin{bmatrix}\hat{x}_1-1\\ 0\\ 0\end{bmatrix},\ \begin{bmatrix}0\\ \hat{x}_2-1\\ 0\end{bmatrix},\ \begin{bmatrix}0\\ 0\\ \hat{x}_3-1\end{bmatrix}.$$

$B = [x_2 - x_1 \,|\, x_4 - x_1 \,|\, x_5 - x_1],\ b = x_1.$
$\boldsymbol{v}_1(x) = (x_2 - x_1)\,[\zeta_2(x) + \zeta_3(x) + \zeta_6(x) + \zeta_7(x)]/J,$
$\boldsymbol{v}_2(x) = (x_4 - x_1)\,[\zeta_3(x) + \zeta_4(x) + \zeta_7(x) + \zeta_8(x)]/J,$
$\boldsymbol{v}_3(x) = (x_5 - x_1)\,[\zeta_5(x) + \zeta_6(x) + \zeta_7(x) + \zeta_8(x)]/J,$
$\boldsymbol{v}_4(x) = (x_1 - x_2)\,[\zeta_1(x) + \zeta_4(x) + \zeta_5(x) + \zeta_8(x)]/J,$
$\boldsymbol{v}_5(x) = (x_1 - x_4)\,[\zeta_1(x) + \zeta_2(x) + \zeta_5(x) + \zeta_6(x)]/J,$
$\boldsymbol{v}_6(x) = (x_1 - x_5)\,[\zeta_1(x) + \zeta_2(x) + \zeta_3(x) + \zeta_4(x)]/J.$

6. A mixed finite element method.

We now introduce the lowest order Raviart-Thomas discretization of (3.4). Assume henceforth that Ω is a polygon ($d = 2$) or a polyhedron ($d = 3$). A triangulation of $\overline{\Omega}$ (see [9: chapter 2]) is constructed by subdividing $\overline{\Omega}$ in a collection S_h of closed, simply connected subdomains $S \in \mathrm{S}_h$, called finite elements, such that every subdomain $S \in \mathrm{S}_h$ is a triangle or a parallelogram ($d = 2$), or a tetrahedron, a prism or a parallelepiped ($d = 3$). Define $h = \max\limits_{S \in \mathrm{S}_h} \mathrm{diam}\,(S)$.

In order to state a finite element formulation of problem (3.4) it is necessary to define finite-dimensional subspaces of $H(\mathrm{div};\Omega)$ and $L^2(\Omega)$. These spaces are called Raviart-Thomas spaces and multiplier spaces, respectively.

Let E_h be the collection of edges ($d = 2$) or faces ($d = 3$) of subdomains $S \in \mathrm{S}_h$ and define

$$(6.1)\qquad \mathrm{E}_h^{\partial} = \{e \in \mathrm{E}_h \,|\, e \subset \partial\Omega\}.$$

We assume that $\overline{\partial\Omega_D}$ is the union of some $e \in \mathrm{E}_h^{\partial}$. Now, let $g_{N,h} \in H^{-1/2}(\partial\Omega)$

be a piecewise constant approximation of g_N, such that

$$\text{(6.2)} \qquad \int_e (g_{N,h} - g_N)\, ds = 0 \quad \forall e \in \mathrm{E}_h^{\partial}.$$

Define the Raviart-Thomas spaces

$$\text{(6.3)} \qquad RT^0_{-1}(\mathrm{S}_h) = \{u \in L^2(\Omega) \,|\, u|_S \in RT^0(S) \quad \forall S \in \mathrm{S}_h\},$$

$$\text{(6.4)} \qquad \begin{aligned} RT^0_0(\mathrm{S}_h) &= \{u \in RT^0_{-1}(\mathrm{S}_h) \,|\, \text{the normal component of } u \text{ is continuous} \\ &\qquad \text{across the interelement boundaries}\} \\ &= RT^0_{-1}(\mathrm{S}_h) \cap H(\mathrm{div};\Omega), \end{aligned}$$

$$\text{(6.5)} \qquad \begin{aligned} RT^0_{0,N}(\mathrm{S}_h) &= \{u \in RT^0_0(\mathrm{S}_h) \,|\, n\cdot u = 0 \text{ on } \partial\Omega_N\} \\ &= RT^0_{-1}(\mathrm{S}_h) \cap H_N(\mathrm{div};\Omega), \end{aligned}$$

$$\text{(6.6)} \qquad \begin{aligned} RT^0_{0,*}(\mathrm{S}_h) &= \{u \in RT^0_0(\mathrm{S}_h) \,|\, n\cdot u = g_{N,h} \text{ on } \partial\Omega_N\} \\ &= RT^0_{-1}(\mathrm{S}_h) \cap H_*(\mathrm{div};\Omega). \end{aligned}$$

Further, the multiplier space $M^0_{-1}(\mathrm{S}_h)$ is defined as

$$\text{(6.7)} \qquad M^0_{-1}(\mathrm{S}_h) = \{\phi \in L^2(\Omega) \,|\, \phi|_S \in M^0(S) \; \forall S \in \mathrm{S}_h\}.$$

The lowest order Raviart-Thomas mixed method for problem (3.4) now reads as follows:

Find $(u_h, \phi_h) \in RT^0_{0,*}(\mathrm{S}_h) \times M^0_{-1}(\mathrm{S}_h)$, such that

$$\text{(6.8a)} \qquad \int_\Omega (Cu_h)\cdot v_h \, dx - \int_\Omega \phi_h \, \nabla\cdot v_h \, dx = - \int_{\partial\Omega} g_D \; n\cdot v_h \, ds$$

$$\forall v_h \in RT^0_{0,N}(\mathrm{S}_h),$$

$$\text{(6.8b)} \qquad - \int_\Omega \nabla\cdot u_h \; \psi_h \, dx = - \int_\Omega f \, \psi_h \, dx \quad \forall \psi_h \in M^0_{-1}(\mathrm{S}_h).$$

The problem (6.8) has a unique solution (see [32: theorem IX-2.1]). Moreover, if $(\boldsymbol{u},\phi) \in H_*(\mathrm{div};\Omega) \times L^2(\Omega)$ is the unique solution of (3.4), and $(\boldsymbol{u}_h,\phi_h) \in RT^0_{0,*}(\mathrm{S}_h) \times M^0_{-1}(\mathrm{S}_h)$ is the unique solution of (6.8), then

$$(6.9) \qquad \| \boldsymbol{u} - \boldsymbol{u}_h \|_{\mathrm{div},\Omega} + \| \phi - \phi_h \|_{0,\Omega}$$
$$\leq C \Big\{ \inf_{\boldsymbol{v}_h \in RT^0_{0,*}(\mathrm{S}_h)} \| \boldsymbol{u} - \boldsymbol{v}_h \|_{\mathrm{div},\Omega} + \inf_{\psi_h \in M^0_{-1}(\mathrm{S}_h)} \| \phi - \psi_h \|_{0,\Omega} \Big\},$$

where the constant $C > 0$ does not depend on h.

NOTE 6.1: An essential condition when proving that problem (6.8) has a unique solution and when proving (6.9) is the discrete inf-sup condition (also called the discrete Babuška- Brezzi condition)

$$(6.10) \qquad \inf_{\phi_h \in M^0_{-1}(\mathrm{S}_h)\setminus\{0\}} \ \sup_{\boldsymbol{v}_h \in RT^0_{0,N}(\mathrm{S}_h)\setminus\{0\}} \frac{\int_\Omega \phi_h \, \nabla\cdot\boldsymbol{v}_h \, dx}{\| \phi_h \|_{0,\Omega} \, \| \boldsymbol{v}_h \|_{\mathrm{div},\Omega}} \geq \beta,$$

where the constant $\beta > 0$ only depends on Ω (for a proof of this condition, see [32: lemma IX-3.3]). The existence and uniqueness of the solution of (6.8) and the inequality (6.9) then follow using [5: corollary 2.1]. □

NOTE 6.2: Let the Sobolev space $H^1(\Omega)$ (see (2.3)) be equipped with the semi-norm

$$(6.11) \qquad | \phi |_{1,\Omega} = \| \nabla\phi \|_{0,\Omega}, \ \phi \in H^1(\Omega),$$

and the Sobolev space $\boldsymbol{H}^1(\Omega) = \{\boldsymbol{u} = (u_1,\dots,u_d)^T \in \boldsymbol{L}^2(\Omega) \,|\, u_i \in H^1(\Omega),\ i = 1,\dots,d\}$ with the semi-norm

$$(6.12) \qquad | \boldsymbol{u} |_{1,\Omega} = \Big[\sum_{i=1}^{d} | u_i |^2_{1,\Omega} \Big]^{1/2}, \ \boldsymbol{u} \in \boldsymbol{H}^1(\Omega).$$

Let $(u,\phi) \in H_*(\mathrm{div};\Omega) \times L^2(\Omega)$ be the unique solution of (3.4), and $(u_h,\phi_h) \in RT^0_{0,*}(S_h) \times M^0_{-1}(S_h)$ be the unique solution of (6.8). If $u \in H^1(\Omega)$, $\nabla\cdot u \in H^1(\Omega)$ and $\phi \in H^1(\Omega)$, then

$$(6.13) \qquad \| u - u_h \|_{\mathrm{div},\Omega} + \| \phi - \phi_h \|_{0,\Omega}$$

$$\leq Ch\, \{ |u|_{1,\Omega} + |\nabla\cdot u|_{1,\Omega} + |\phi|_{1,\Omega} \},$$

where the constant $C > 0$ does not depend on h (see [32: notes IX-3.2 and IX-4.1]). □

NOTE 6.3: If $f = 0$ in Ω, then it follows from (6.8b) that $\nabla\cdot u_h = 0$ in Ω. In this case u_h can be computed without computing ϕ_h (for the corresponding dual variational problem, see note 3.3). Define the finite-dimensional space

$$(6.14) \qquad X_h(S_h) = \{ u_h \in RT^0_0(S_h) \mid \nabla\cdot u_h = 0 \text{ in } \Omega \}.$$

The linear subspace $X_{N,h}(S_h)$ and the linear variety $X_{*,h}(S_h)$ are defined as

$$(6.15a) \qquad X_{N,h}(S_h) = X_h(S_h) \cap RT^0_{0,N}(S_h),$$

$$(6.15b) \qquad X_{*,h}(S_h) = X_h(S_h) \cap RT^0_{0,*}(S_h).$$

From (6.8a) it follows that $u_h \in X_{0,h}(S_h)$ is the unique solution of the following problem:

Find $u_h \in X_{0,h}(S_h)$, such that

$$(6.16) \qquad \int_\Omega (Cu_h)\cdot v_h \, dx = - \int_{\partial\Omega} g_D \, n\cdot v_h \, ds \quad \forall v_h \in X_{N,h}(S_h)$$

(for more details see [19: section 3.3], [16]). □

NOTE 6.4: If $d = 2$ and $f = 0$ almost everywhere in Ω, then an approximation Ψ_h of the unique solution $\Psi \in Y_*(\Omega)$ of problem (3.11) can be determined using the lowest order conforming finite element method (see [9: chapter 2]). Of course, we demand that $\Psi_h = G_{N,h}$ on $\partial\Omega_N$, where $G_{N,h} \in H^{1/2}(\partial\Omega)$ is a piecewise linear approximation of G_N, such that

$$\text{(6.17)} \qquad \int_e (G_{N,h} - G_N)\, ds = 0 \quad \forall e \in \mathrm{E}_h^{\partial}.$$

Define $\tilde{u}_h = -\nabla \times \Psi_h$ in Ω, then $n{\cdot}\tilde{u}_h = g_{N,h}$ on $\partial\Omega_N$. Moreover, since Ψ_h is continuous across the interelement boundaries, it follows immediately that the normal component of $\tilde{u}_h$ is continuous across the interelement boundaries. In addition, $\nabla{\cdot}\tilde{u}_h = 0$ in S for all $\mathrm{S} \in \mathrm{S}_h$ and thus $\nabla{\cdot}\tilde{u}_h = 0$ in Ω. Therefore, $\tilde{u}_h = u_h$ is the unique solution of (6.16). □

The stage is now set to introduce an equivalent system of linear equations for the variational problem (6.8). Let $\tilde{e}_i$, $i = 1,\dots,\tilde{I}$, be the numbered edges ($d = 2$) or faces ($d = 3$) of $\{e \in \mathrm{E}_h \,|\, e \not\subset \overline{\partial\Omega_N}\}$ and let S_j, $j = 1,\dots,J$, be the numbered subdomains of S_h.

The finite-dimensional space $RT^0_{0,N}(\mathrm{S}_h)$ is spanned by the linearly independent vectorial basis functions $\tilde{v}_i$, $i = 1,\dots,\tilde{I}$, such that

$$\text{(6.18)} \qquad \int_{\tilde{e}_j} \tilde{n}_j{\cdot}\tilde{v}_i\, ds = \delta_{ij},\ i,j = 1,\dots,\tilde{I},$$

where the vector $\tilde{n}_j$ is the normal to $\tilde{e}_j$ pointing from S_k to S_ℓ, $k > \ell$, if $\tilde{e}_j = S_k \cap S_\ell \not\subset \overline{\partial\Omega_N}$, and outwards if $\tilde{e}_j \subset \overline{\partial\Omega_D}$ (see also (5.3)). Thus, a function $u \in RT^0_{0,N}(\mathrm{S}_h)$ has one degree of freedom per edge ($d = 2$) or face ($d = 3$) $\tilde{e}_i$, $i = 1,\dots,\tilde{I}$, which is equal to $\int_{\tilde{e}_i} \tilde{n}_i{\cdot}u\, ds$, i.e. the flux across $\tilde{e}_i$ in the direction of $\tilde{n}_i$.

The finite-dimensional space $M^0_{-1}(\mathrm{S}_h)$ is spanned by the linearly independent scalar basis functions ψ_j, $j = 1,\dots,J$, such that

$$\psi_i(x) = \delta_{ij},\ x \in S_j,\ i,j = 1,\dots,J. \tag{6.19}$$

Thus, a function $\phi \in M^0_{-1}(\mathrm{S}_h)$ has one degree of freedom per subdomain $S \in \mathrm{S}_h$, which is equal to its constant value in S.

By definition, functions u_h and ϕ_h belong to $RT^0_{0,N}(\mathrm{S}_h)$ and $M^0_{-1}(\mathrm{S}_h)$, respectively, if, and only if, they can be expressed as

$$u_h(x) = \sum_{i=1}^{\tilde{I}} \tilde{u}_i\, \tilde{v}_i(x),\ \phi_h(x) = \sum_{j=1}^{J} \phi_j\, \psi_j(x),\ x \in \Omega. \tag{6.20}$$

Introducing (6.20) in (6.8), we obtain the system of linear equations

$$\tilde{A}\tilde{U} + \tilde{B}\Phi = \tilde{F}_1, \tag{6.21a}$$

$$\tilde{B}^T\tilde{U} = F_2, \tag{6.21b}$$

where $\tilde{U} = \left(\tilde{u}_1,\dots,\tilde{u}_{\tilde{I}}\right)^T$, $\Phi = (\phi_1,\dots,\phi_J)^T$, and

$$\underset{(\tilde{I}\times\tilde{I})}{\tilde{A}} = \int_\Omega (C\tilde{v}_i)\cdot\tilde{v}_j\, dx,\quad \underset{(\tilde{I}\times J)}{\tilde{B}} = -\int_\Omega \nabla\cdot\tilde{v}_i\, \psi_j\, dx, \tag{6.22a}$$

$$\underset{(\tilde{I})}{\tilde{F}_1} = -\int_{\partial\Omega} g_D\, n\cdot\tilde{v}_j\, dx - \int_\Omega (Cu_h^*)\cdot\tilde{v}_j\, dx,\ \underset{(J)}{F_2} = -\int_\Omega f\,\psi_j\, dx, \tag{6.22b}$$

where $u_h^* \in RT^0_{0,*}(\mathrm{S}_h)$, such that $\int_{\tilde{e}_i} \tilde{n}_i\cdot u_h^*\, ds = 0,\ i = 1,\dots,\tilde{I}$.

Since $\tilde{A}$ is a symmetric positive definite matrix (6.21a) yields

$$\tilde{U} = \tilde{A}^{-1}(\tilde{F}_1 - \tilde{B}\Phi). \tag{6.23}$$

Using (6.23) in (6.21b), we get

$$\tilde{B}^T\tilde{A}^{-1}\tilde{B}\Phi = \tilde{B}^T\tilde{A}^{-1}\tilde{F}_1 - F_2. \tag{6.24}$$

The next theorem is stated in [32: page IX-25] and follows immediately from the discrete inf-sup condition (6.10). For clarification we give an alternative proof.

THEOREM 6.1 (see also [8: proposition V-1]): $\tilde{B}^T\tilde{A}^{-1}\tilde{B}$ *is a symmetric positive definite matrix.*

PROOF: $\tilde{A}$ is symmetric positive definite and therefore $\tilde{B}^T\tilde{A}^{-1}\tilde{B}$ is symmetric positive semi-definite. Since

$$\begin{aligned}
&\tilde{B}\Phi = 0 \Leftrightarrow V^T B\Phi = 0 && \forall V \in \mathbb{R}^{\tilde{I}} \\
&\Leftrightarrow \int_\Omega \phi_h \, \nabla \cdot v_h \, dx = 0 && \forall v_h \in RT^0_{0,N}(S_h) \\
&\Leftrightarrow \int_\Omega \phi_h \, \psi_h \, dx = 0 && \forall \psi_h \in M^0_{-1}(S_h) \Leftrightarrow \phi_h = 0 \text{ in } \Omega \Leftrightarrow \Phi = 0,
\end{aligned}$$

where $\phi_h(x) = \sum_{j=1}^{J} \phi_j \psi_j(x)$, $x \in \Omega$, we have $\Phi^T\tilde{B}^T\tilde{A}^{-1}\tilde{B}\Phi = (\tilde{B}\Phi)^T\tilde{A}^{-1}(\tilde{B}\Phi) > 0$ $\forall \Phi \in \mathbb{R}^J\backslash\{0\}$. □

Modelling a practical elliptic boundary value problem (3.1) using the mixed finite element method may result in very large and sparse matrices $\tilde{A}$ and $\tilde{B}$, especially if the domain Ω is three-dimensional. Therefore a fast and efficient iterative method is required to solve the resulting system of linear equations. However, the matrix of the system (6.21) is not positive definite and the matrix $\tilde{B}^T\tilde{A}^{-1}\tilde{B}$ is not sparse. Fast and efficient methods are not yet known for these situations (for some iterative methods that have been used to solve (6.21), see [30: section 17]).

7. Hybridization of the mixed method.

The solution of the system of linear equations resulting from (6.8) can be simplified by enlarging the Raviart-Thomas space in which $\boldsymbol{u}_h$ is sought and introducing a Lagrange multiplier to enforce the continuity of the normal component of $\boldsymbol{u}_h$ across the interelement boundaries.

Recall that E_h is the collection of edges ($d = 2$) or faces ($d = 3$) of subdomains $S \in \mathrm{S}_h$ and $\mathrm{E}_h^{\partial} = \{e \in \mathrm{E}_h \,|\, e \subset \partial\Omega\}$. Let $g_{D,h} \in L^2(\Omega)$ be a piecewise constant approximation of g_D, such that

$$\text{(7.1)} \qquad \int_e (g_{D,h} - g_D)\, ds = 0 \quad \forall e \in \mathrm{E}_h^{\partial}.$$

Let $M^0(e)$, $e \in \mathrm{E}_h$, be the space of constant functions on e. Define the multiplier spaces

$$\text{(7.2)} \qquad M^0_{-1}(\mathrm{E}_h) \;= \{\lambda \in H^{1/2}(\bigcup_{e\in\mathrm{E}_h} e) \,|\, \lambda|_e \in M^0(e) \;\; \forall e \in \mathrm{E}_h\},$$

$$\text{(7.3)} \qquad M^0_{-1,D}(\mathrm{E}_h) = \{\lambda \in M^0_{-1}(\mathrm{E}_h) \,|\, \lambda = 0 \text{ on } \partial\Omega_D\},$$

$$\text{(7.4)} \qquad M^0_{-1,*}(\mathrm{E}_h) \;= \{\lambda \in M^0_{-1}(\mathrm{E}_h) \,|\, \lambda = g_{D,h} \text{ on } \partial\Omega_D\}.$$

Now, it follows immediately that if $\boldsymbol{u} \in RT^0_{-1}(\mathrm{S}_h)$, then $\boldsymbol{u} \in RT^0_{0,N}(\mathrm{S}_h)$ if, and only if,

$$\text{(7.5)} \qquad \sum_{S\in\mathrm{S}_h} \int_{\partial S} \boldsymbol{n}_S \cdot \boldsymbol{u}\, \mu \, ds = 0 \quad \forall \mu \in M^0_{-1,D}(\mathrm{E}_h),$$

where $\boldsymbol{n}_S$ is the outward normal to ∂S. Thus, the hybrid version of the lowest order Raviart-Thomas mixed method for problem (3.4) reads as follows:

Find $(\boldsymbol{u}_h, \phi_h, \lambda_h) \in RT^0_{-1}(\mathrm{S}_h) \times M^0_{-1}(\mathrm{S}_h) \times M^0_{-1,*}(\mathrm{E}_h)$, such that

$$\text{(7.6a)} \quad \int_\Omega (C\boldsymbol{u}_h)\cdot\boldsymbol{v}_h\,dx - \sum_{S\in \mathrm{S}_h} \{\int_S \phi_h\,\nabla\cdot\boldsymbol{v}_h\,dx - \int_{\partial S} \lambda_h\,\boldsymbol{n}_S\cdot\boldsymbol{v}_h\,ds\} = 0 \quad \forall \boldsymbol{v}_h \in RT^0_{-1}(\mathrm{S}_h),$$

$$\text{(7.6b)} \quad -\sum_{S\in \mathrm{S}_h} \int_S \nabla\cdot\boldsymbol{u}_h\,\psi_h\,dx = -\int_\Omega f\,\psi_h\,dx \quad \forall\psi_h \in M^0_{-1}(\mathrm{S}_h),$$

$$\text{(7.6c)} \quad \sum_{S\in \mathrm{S}_h} \int_{\partial S} \boldsymbol{n}_S\cdot\boldsymbol{u}_h\,\mu_h\,ds = \int_{\partial\Omega} g_N\,\mu_h\,ds \quad \forall\mu_h \in M^0_{-1,D}(\mathrm{E}_h).$$

The problem (7.6) has a unique solution (see [1: lemma 1.3]). Moreover, $\boldsymbol{u}_h = \tilde{\boldsymbol{u}}_h$ and $\phi_h = \tilde{\phi}_h$, where $(\tilde{\boldsymbol{u}}_h,\tilde{\phi}_h) \in RT^0_{0,*}(\mathrm{S}_h) \times M^0_{-1}(\mathrm{S}_h)$ is the unique solution of (6.8).

NOTE 7.1: Let $(\boldsymbol{u},\phi) \in H_*(\mathrm{div};\Omega) \times L^2(\Omega)$ be the solution of problem (3.4), then from (3.4a) it follows that

$$\text{(7.7)} \quad \int_\Omega (C\boldsymbol{u})\cdot\boldsymbol{v}\,dx - \sum_{S\in \mathrm{S}_h} \left\{\int_S \phi\,\nabla\cdot\boldsymbol{v}\,dx - \int_{\partial S} \phi\,\boldsymbol{n}_S\cdot\boldsymbol{v}\,ds\right\} = 0 \quad \forall\boldsymbol{v} \in H(\mathrm{div};\Omega).$$

When this equality is compared with (7.6a) the multiplier $\lambda_h \in M^0_{-1,*}(\mathrm{E}_h)$ seems to be an approximation of the trace of ϕ on $\bigcup_{e\in \mathrm{E}_h} e$. This conjecture is proved in [1: corollary 1.5]. □

Next, an equivalent system of linear equations for the variational problem (7.6) is introduced. Let S_j, $j = 1,\dots,J$, be the numbered subdomains of S_h and $e_i^{(S)}$, $i = 1,\dots,I^{(S)}$, be the edges ($d = 2$) or faces ($d = 3$) of S for each $S \in \mathrm{S}_h$.

The finite-dimensional space $RT^0_{-1}(\mathrm{S}_h)$ is spanned by the linearly independent vectorial basis functions $\boldsymbol{v}_i^{(S)}$, $i = 1,\dots,I^{(S)}$, $S \in \mathrm{S}_h$, such that $\boldsymbol{v}_i^{(S)}$ has its support in S, and

$$\text{(7.8)} \quad \int_{e_j^{(S)}} \boldsymbol{n}_S\cdot\boldsymbol{v}_i^{(S)}\,ds = \delta_{ij},\ i,j = 1,\dots,I^{(S)}.$$

Thus, a function $\boldsymbol{u} \in RT^0_{-1}(\mathrm{S}_h)$ has $I^{(S)}$ degrees of freedom per subdomain $S \in \mathrm{S}_h$, which are equal to $\int_{e_i^{(S)}} \boldsymbol{n}_S \cdot \boldsymbol{u}\, ds$, $i = 1,...,I^{(S)}$, i.e. the outward (with respect to S) fluxes across $e_i^{(S)}$.

Again, the finite-dimensional space $M^0_{-1}(\mathrm{S}_h)$ is spanned by the linearly independent scalar basis functions ψ_j, $j = 1,...,J$, such that (6.19) holds.

Let e_k, $k = 1,...,K$, be the numbered edges ($d = 2$) or faces ($d = 3$) of $\{e \in \mathrm{E}_h \mid e \not\subset \overline{\partial\Omega_D}\}$. The finite-dimensional space $M^0_{-1,D}(\mathrm{E}_h)$ is spanned by the linearly independent scalar basis functions μ_k, $k = 1,...,K$, such that

$$\mu_i(x) = \delta_{ij},\ x \in e_j,\ i,j = 1,...,K. \tag{7.9}$$

Thus, a function $\lambda \in M^0_{-1,D}(\mathrm{E}_h)$ has one degree of freedom per edge e_k, $k = 1,...,K$, which is equal to its constant value on e_k.

By definition, functions $\boldsymbol{u}_h, \phi_h$ and λ_h belong to $RT^0_{-1}(\mathrm{S}_h)$, $M^0_{-1}(\mathrm{S}_h)$ and $M^0_{-1,D}(\mathrm{E}_h)$, respectively, if, and only if, they can be expressed as

$$\boldsymbol{u}_h(x) = \sum_{i=1}^{I} u_i\, \boldsymbol{v}_i(x),\ \phi_h(x) = \sum_{j=1}^{J} \phi_j\, \psi_j(x),\ x \in \Omega, \tag{7.10a}$$

$$\lambda_h(x) = \sum_{k=1}^{K} \lambda_k\, \mu_k(x),\ x \in \bigcup_{e \in \mathrm{E}_h} e, \tag{7.10b}$$

where $I = \sum_{S \in \mathrm{S}_h} I^{(S)}$. Note that the first equality in (7.10a) is equivalent to

$$\boldsymbol{u}_h(x) = \sum_{S \in \mathrm{S}_h} \sum_{i=1}^{I^{(S)}} u_i^{(S)}\, \boldsymbol{v}_i^{(S)}(x),\ x \in \Omega.$$

Introducing (7.10) in (7.6) we obtain the system of linear equations

$$AU + B\Phi + C\Lambda = F_1, \tag{7.11a}$$
$$B^T U = F_2, \tag{7.11b}$$
$$C^T U = F_3, \tag{7.11c}$$

where $U = (u_1,\dots,u_I)^T$, $\Phi = (\phi_1,\dots,\phi_J)^T$, $\Lambda = (\lambda_1,\dots,\lambda_K)^T$, and

$$\text{(7.12a)}\qquad \underset{(I\times I)}{A} = \int_\Omega (Cu_i)\cdot v_j\, dx, \quad \underset{(I\times J)}{B} = -\int_{S_j} \nabla\cdot v_i\, dx, \quad \underset{(I\times K)}{C} = \int_{e_k} n_i\cdot v_i\, ds,$$

$$\text{(7.12b)}\qquad \underset{(I)}{F_1} = -\int_{\partial\Omega} g_D\, n\cdot v_i\, dx, \quad \underset{(J)}{F_2} = -\int_\Omega f\,\psi_j\, dx, \quad \underset{(K)}{F_3} = \int_{\partial\Omega} g_N\, \mu_k\, ds,$$

where, if v_i has its support in a certain subdomain $S \in \mathrm{S}_h$, then n_i is the outward normal to ∂S.

The advantage of system (7.11) compared with (6.21) is the block-diagonality of the symmetric positive definite matrix A. Hence A can be inverted at the finite element level. Thus, (7.11a) yields

$$\text{(7.13)}\qquad U = A^{-1}(F_1 - B\Phi - C\Lambda).$$

Using (7.13) in (7.11b) and (7.11c), we get

$$\text{(7.14a)}\qquad B^TA^{-1}B\Phi + B^TA^{-1}C\Lambda = B^TA^{-1}F_1 - F_2,$$
$$\text{(7.14b)}\qquad C^TA^{-1}B\Phi + C^TA^{-1}C\Lambda = C^TA^{-1}F_1 - F_3.$$

THEOREM 7.1 (see also [8: proposition V-3]): $(B\,|\,C)^TA^{-1}(B\,|\,C)$ *is a symmetric positive definite matrix.*

PROOF: A is symmetric positive definite and therefore $(B\,|\,C)^TA^{-1}(B\,|\,C)$ is symmetric positive semi-definite. Since

$$B\Phi + C\Lambda = 0 \Leftrightarrow V^T(B\Phi + C\Lambda) = 0 \quad \forall V \in \mathbb{R}^I$$

$$\Leftrightarrow \sum_{S\in \mathrm{S}_h} \left\{\int_S \phi_h\, \nabla\cdot v_h\, dx - \int_{\partial S} \lambda_h\, n_S\cdot v_h\, ds\right\} = 0 \quad \forall v_h \in RT^0_{-1}(\mathrm{S}_h)$$

$$\Leftrightarrow \sum_{S\in \mathrm{S}_h} \int_{\partial S} (\phi_h - \lambda_h)\, n_S\cdot v_h\, ds = 0 \quad \forall v_h \in RT^0_{-1}(\mathrm{S}_h)$$

$$\Leftrightarrow \phi_h = \lambda_h \text{ on } \partial S \;\; \forall S \in \mathrm{S}_h \Leftrightarrow \phi_h = 0 \text{ in } \Omega,\ \lambda_h = 0 \text{ on } \bigcup_{e\in \mathrm{E}_h} e$$

$$\Leftrightarrow \Phi = 0,\ \Lambda = 0,$$

where $\phi_h(x) = \sum_{j=1}^{J} \phi_j\, \psi_j(x)$, $x \in \Omega$, and $\lambda_h(x) = \sum_{k=1}^{K} \lambda_k\, \mu_k(x)$, $x \in \bigcup_{e\in \mathrm{E}_h} e$, we have $(B\Phi + C\Lambda)^T A^{-1}(B\Phi + C\Lambda) > 0 \;\; \forall \Lambda \in \mathbb{R}^K \backslash \{0\}$. □

Now, $B^TA^{-1}B$ is a diagonal matrix. Thus, (7.14a) yields

$$(7.15) \qquad \Phi = (B^TA^{-1}B)^{-1}\, (B^TA^{-1}(F_1 - C\Lambda) - F_2).$$

Using (7.15) in (7.14b) we get

$$(7.16) \qquad D\Lambda = F,$$

where

$$(7.17a) \qquad D = C^T(A^{-1} - A^{-1}B(B^TA^{-1}B)^{-1}\, B^TA^{-1})\, C,$$

$$(7.17b) \qquad F = C^TA^{-1}(F_1 - B(B^TA^{-1}B)^{-1}(B^TA^{-1}F_1 - F_2)) - F_3.$$

From theorem 7.1 it follows that $\Lambda^T D\Lambda = (B\hat{\Phi} + C\Lambda)^T A^{-1}(B\hat{\Phi} + C\Lambda) > 0$ $\forall \Lambda \in \mathbb{R}^K\backslash\{0\}$, where $\hat{\Phi} = -\,(B^TA^{-1}B)\; B^TA^{-1}C\Lambda$. Hence D is a symmetric positive definite matrix.

The sparsity pattern of the matrix D follows immediately from (7.17a). Let $D = [d_{ij}]_{i,j=1,\dots,K}$, then $d_{ij} \neq 0$ if, and only if, $e_i, e_j \subset \partial S$ for some $S \in \mathrm{S}_h$. Thus, D has the same sparsity pattern as the resulting matrix of the lowest order nonconforming Ritz- Galerkin method for problem (3.2) (see, e.g., [6: section 5.5], [9: section 4.2]).

8. The system of linear equations.

Before the system (7.11) can be assembled, the element contributions to the matrices and right-hand sides in (7.12) need to be computed.

First, the element contributions to the matrix A, i.e.

$$(8.1) \qquad \underset{(I^{(S)}\times I^{(S)})}{A^{(S)}} = \int_S (Cv_i^{(S)})\cdot v_j^{(S)}\, dx = \int_{\hat{S}} (\hat{C}\,\hat{v}_i)\cdot\hat{v}_j\, d\hat{x},$$

where $\hat{C}(\hat{x}) = B^T C(x)\; B/J$ and $\hat{v}_i(\hat{x}) = B^{-1}\; v_i^{(S)}(x)\; J,\; x \in S$, are considered (see section 5). Note that the local basis functions $\hat{v}_i$, $i = 1,\ldots,I^{(S)}$, only depend on the reference element $\hat{S}$.

If the tensor $A \in L^{\infty}(\Omega,\mathbb{R}^{d\times d})$, and therefore also $C = A^{-1}$, is constant on each subdomain $S \in \mathrm{S}_h$, then (8.1) can be computed exactly (see also [31: section 2.2]). Define $C^{(S)} = C(x)$, $x \in S$, and $\hat{C}^{(S)} = B^T C^{(S)} B/J$. If the notations $\hat{v}_i(\hat{x}) = [\hat{v}_{ik}(\hat{x})]_{k=1,\ldots,d}$ and $\hat{C}^{(S)} = [\hat{c}_{k\ell}^{(S)}]_{k,\ell=1,\ldots,d}$ are used, then

$$(8.2) \qquad \hat{A}^{(S)} = \sum_{k=1}^{d} \hat{c}_{k\ell}^{(S)}\, A_{k\ell},$$

where the matrices

$$(8.3) \qquad \underset{(I^{(S)}\times I^{(S)})}{A_{k\ell}} = \int_{\hat{S}} \hat{v}_{ik}\; \hat{v}_{j\ell}\; d\hat{x}$$

only depend on the reference element $\hat{S}$. Note that $\hat{c}_{k\ell}^{(S)} = \hat{c}_{\ell k}^{(S)}$ and $A_{k\ell} = A_{\ell k}^T$, thus

$$(8.4) \qquad A^{(S)} = \sum_{k=1}^{d} \hat{c}_{kk}^{(S)}\, A_{kk} + \sum_{k=1}^{d}\; \sum_{\ell=1}^{k-1} \hat{c}_{k\ell}^{(S)}(A_{k\ell} + A_{\ell k}),$$

where all matrices in the summations are symmetric.

An overview of the matrices A_{kk} and $A_{k\ell} + A_{\ell k}$, $k \neq \ell$, corresponding to various choices of the reference element $\hat{S}$ is given below.

Triangle ($d = 2$, $I^{(S)} = 3$):

$$A_{11} = \frac{1}{12}\begin{bmatrix} 1 & -1 & 1 \\ -1 & 3 & -1 \\ 1 & -1 & 1 \end{bmatrix}, \quad A_{12} + A_{21} = \frac{1}{12}\begin{bmatrix} 1 & -1 & -1 \\ -1 & -3 & 3 \\ -1 & 3 & -3 \end{bmatrix},$$

$$A_{22} = \frac{1}{12}\begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 3 \end{bmatrix}.$$

Parallelogram ($d = 2$, $I^{(S)} = 4$):

$$A_{11} = \frac{1}{6}\begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_{12} + A_{21} = \frac{1}{4}\begin{bmatrix} 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix},$$

$$A_{22} = \frac{1}{6}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 2 \end{bmatrix}.$$

Tetrahedron ($d = 3$, $I^{(S)} = 4$):

$$A_{11} = \frac{1}{30}\begin{bmatrix} 2 & -3 & 2 & 2 \\ -3 & 12 & -3 & -3 \\ 2 & -3 & 2 & 2 \\ 2 & -3 & 2 & 2 \end{bmatrix}, \quad A_{12} + A_{21} = \frac{1}{30}\begin{bmatrix} 2 & -3 & -3 & 2 \\ -3 & -8 & 12 & -3 \\ -3 & 12 & -8 & -3 \\ 2 & -3 & -3 & 2 \end{bmatrix},$$

$$A_{22} = \frac{1}{30}\begin{bmatrix} 2 & 2 & -3 & 2 \\ 2 & 2 & -3 & 2 \\ -3 & -3 & 12 & -3 \\ 2 & 2 & -3 & 2 \end{bmatrix}, \quad A_{23} + A_{32} = \frac{1}{30}\begin{bmatrix} 2 & 2 & -3 & -3 \\ 2 & 2 & -3 & -3 \\ -3 & -3 & -8 & 12 \\ -3 & -3 & 12 & -8 \end{bmatrix},$$

$$A_{33} = \frac{1}{30}\begin{pmatrix} 2 & 2 & 2 & -3 \\ 2 & 2 & 2 & -3 \\ 2 & 2 & 2 & -3 \\ -3 & -3 & -3 & 12 \end{pmatrix}, A_{31} + A_{13} = \frac{1}{30}\begin{pmatrix} 2 & -3 & 2 & -3 \\ -3 & -8 & -3 & 12 \\ 2 & -3 & 2 & -3 \\ -3 & 12 & -3 & -8 \end{pmatrix}.$$

Prism ($d = 3,\ I^{(S)} = 5$):

$$A_{11} = \frac{1}{12}\begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, A_{12} + A_{21} = \frac{1}{12}\begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ -1 & -3 & 3 & 0 & 0 \\ -1 & 3 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$A_{22} = \frac{1}{12}\begin{pmatrix} 1 & 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, A_{23} + A_{32} = \frac{1}{6}\begin{pmatrix} 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -2 & 2 \\ 1 & 1 & -2 & 0 & 0 \\ -1 & -1 & 2 & 0 & 0 \end{pmatrix},$$

$$A_{33} = \frac{1}{3}\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}, \quad A_{31} + A_{13} = \frac{1}{6}\begin{pmatrix} 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -2 & 2 \\ 0 & 0 & 0 & 1 & -1 \\ 1 & -2 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \end{pmatrix}.$$

Parallelepiped ($d = 3,\ I^{(S)} = 6$):

$$A_{11} = \frac{1}{6}\begin{pmatrix} 2 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, A_{12} + A_{21} = \frac{1}{4}\begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$A_{22} = \frac{1}{6}\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, A_{23} + A_{32} = \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$A_{33} = \frac{1}{6}\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 2 \end{pmatrix}, A_{31} + A_{13} = \frac{1}{4}\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

If the tensor A, and therefore $C = A^{-1}$, is not constant in each subdomain $S \in \mathrm{S}_h$, then (8.1) can generally not be computed exactly. This difficulty can easily be circumvented by defining $C^{(S)} = C(x_S)$, where x_S is the centroid of S, and approximating the element contribution $A^{(S)}$ as follows:

$$(8.5) \qquad A^{(S)}_{(I^{(S)} \times I^{(S)})} \approx \int_S (C^{(S)}\, v_i^{(S)}) \cdot v_j^{(S)}\, dx = \int_{\hat{S}} (\hat{C}^{(S)}\, \hat{v}_i) \cdot \hat{v}_j\, d\hat{x},$$

where $\hat{C}^{(S)} = B^T C^{(S)} B / J$.

Note 8.1: Denote $C(x) = [c_{k\ell}(x)]_{k,\ell=1,\ldots,d}$, $x \in S$, and $C^{(S)} = [c_{k\ell}^{(S)}]_{k,\ell=1,\ldots,d}$. If $C \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ is smooth in S, then

$$(8.6) \qquad \int_S (Cu) \cdot v\, dx - \int_S (C^{(S)} u) \cdot v\, dx$$

$$\leq d \max_{k,\ell=1,\ldots,d} \{\operatorname*{ess\,sup}_{x \in S} | c_{k\ell}(x) - c_{k\ell}^{(S)} |\} \, \| u \|_{\mathrm{div},S} \, \| v \|_{\mathrm{div},S}$$

$$\leq Ch \, \| u \|_{\mathrm{div},S} \, \| v \|_{\mathrm{div},S} \quad \forall u, v \in RT^0(S),$$

where the constant $C > 0$ does not depend on h (see [9: theorems 3.1.2 and 3.1.4]). Using [5: theorem 3.1] it follows that (6.13) still holds if the approximation (8.5) is used, i.e. the order of convergence is not reduced. □

NOTE 8.2: Assume that all subdomains $S \in \mathrm{S}_h$ are rectangles ($d = 2$) or blocks ($d = 3$) with edges parallel to the coordinate axes, and $A(x) \in \mathbb{R}^{d\times d}$, and therefore also $C(x) = A(x)^{-1}$, is a diagonal matrix for all $x \in \Omega$. If its element contributions (see (8.1)) are approximated using the trapezoidal quadrature formula, then the matrix $\tilde{A}$, defined in (6.22a), reduces to a diagonal matrix. Thus, the symmetric positive definite matrix $\tilde{B}^T\tilde{A}^{-1}\tilde{B}$ is sparse and the system (6.24) can be solved by the preconditioned conjugate gradient method (see, e.g., [12: chapter 10]). If the element contributions of the integrals $\int_{\partial\Omega} g_D\, n\cdot\tilde{v}_j\, ds$ and $\int_\Omega f\,\psi_j\, dx$ in (6.22b) are approximated using the midpoint quadrature formula, then the mixed finite element method (see section 6) transforms into the block-centered finite difference method (for details, see [35]). □

Using Green's formula, (4.9c) and (5.3), simple formulas can be derived for the element contributions of the remaining matrices and right-hand sides in (7.12), namely

$$(8.7)\qquad \underset{(I^{(S)}\times 1)}{B^{(S)}} = -\int_S \nabla\cdot v_i^{(S)}\, dx = -1,$$

$$(8.8)\qquad \underset{(I^{(S)}\times I^{(S)})}{C^{(S)}} = \int_{e_k^{(S)}} n_S\cdot v_i^{(S)}\, ds = \delta_{ik},$$

$$(8.9)\qquad \underset{(I^{(S)})}{F_1^{(S)}} = -\int_{\partial S\cap\partial\Omega} g_D\, n_S\cdot v_i^{(S)}\, ds = \begin{cases} -\int_{e_i} g_D\, ds \Big/ \int_{e_i} ds & \\ & \text{if } e_i \subset \overline{\partial\Omega_D}, \\ 0 & \text{otherwise,} \end{cases}$$

$$(8.10)\qquad \underset{(1)}{F_2^{(S)}} = -\int_S f\, dx,$$

$$(8.11) \qquad \underset{(I^{(S)})}{F_3^{(S)}} = \int_{\partial S \cap \partial \Omega} g_N \, \mu_k \, ds = \begin{cases} \int_{e_k} g_N \, ds & \text{if } e_k \subset \overline{\partial \Omega}_N, \\ 0 & \text{otherwise.} \end{cases}$$

If the integrals in (8.10) and (8.11) are approximated using the midpoint quadrature formula, then the order of convergence is not reduced, i.e. (6.13) holds (use [5: theorem 3.1]).

The computation of $A^{-1} - A^{-1}B(B^TA^{-1}B)^{-1}\,B^TA^{-1}$ is essential when determining the matrix D (see (7.17a)). This matrix is a block-diagonal matrix and can thus be computed at the finite element level. The matrix D is obtained by assembling its element contributions according to the matrix C. The right-hand side F can be computed in an analogous way.

The symmetric positive definite matrix D is usually large and sparse, but not particularly well conditioned. This motivates the use of the preconditioned conjugate gradient method (see, e.g., [12: chapter 10], [3: section 1.4]) to solve the system (7.16). A variety of choices for the preconditioning matrix have been discussed in the literature (see, e.g., [2], [10]).

Popular methods for computing the preconditioning matrix are to use the incomplete Cholesky decomposition (see [22], [23]) or the modified incomplete Cholesky decomposition (see [13], [3: section 1.4]). Promising alternatives can be found in [21], [14], [17]).

After solving (7.16) the vectors Φ and U can be computed by (7.15) and (7.13), respectively. This computation can be performed at the finite element level. Finally, $u_h \in RT^0_{0,*}(S_h)$ can be computed from its fluxes across the element edges ($d = 2$) or faces ($d = 3$).

9. Streamline computation.

After the approximation $u_h \in RT^0_{0,*}(S_h)$ of the specific discharge $u \in H_*(\mathrm{div};\Omega)$ has been computed, the approximate velocity $w_h = u_h/p$, where p is the porosity, can be determined. Henceforth we assume that p is constant in each subdomain $S \in S_h$.

As stated before, a streamline is a curve that is everywhere tangential to the velocity $w = u/p$. Thus, streamlines indicate the direction of flow almost everywhere in Ω. The time required to flow along a streamline from one point to another is called the residence time between these two points.

Now, recall that the function f is used to represent sources and sinks. If a source or a sink is represented by some subdomains S_m, $m = 1,...,M$, then the discharge is equal to $Q = -\sum_{m=1}^{M} \int_{S_m} f\ dx$. It is meaningless to determine streamlines and residence times in these subdomains. Only streamlines starting at the boundary of a sink or ending at the boundary of a source are significant.

For determining accurate streamlines the approximation u_h must be divergence-free outside sources and sinks. For the mixed finite element method this follows immediately from (6.8b). In this section we present a method to obtain streamlines and residence times using elementwise computations. Moreover, the streamlines and the residence times are determined exactly with respect to the approximate velocity w_h.

NOTE 9.1: Usually the size of a source or a sink is very small compared with the domain Ω. Thus, a strongly refined mesh is required near a source or a sink. If it can be assumed that the flow near a source or a sink is essentially radial, then this source or this sink can be represented by some larger subdomains, using the concept of macroelements (see [7: section 4]). □

Let $S \in S_h$ be some finite element and $\hat{S}$ the corresponding reference element (see section 5). With any point $\hat{x} \in \hat{S}$ the point $x \in S$ is associated by (5.1). With any velocity $\hat{w}$ on $\hat{S}$ the function w on S is associated, where

$$(9.1) \qquad w(x) = B\hat{w}(\hat{x}),\ \hat{x} \in \hat{S}$$

(see, e.g., [24: section 1.5]). Let $\hat{w}(\hat{x}) = \hat{u}(\hat{x})/\hat{p}$, $x \in \hat{S}$, and $w(x) = u(x)/p$, $x \in S$. From (4.8) and (9.1) it follows immediately that

$$(9.2) \qquad p = \hat{p}/J.$$

Triangle (see figure 1a)

Let $\hat{u} \in RT^0(\hat{S})$, then $\hat{w} = \hat{u}/\hat{p}$ is of the form

$$(9.3) \qquad \hat{w}(\hat{x}) = \binom{\alpha}{\beta} + \gamma\hat{x},\ \hat{x} \in \hat{S}.$$

Depending on the sign of $\gamma = \nabla \cdot \hat{u}/\hat{p}$ three cases can be distinguished:

(i) $\gamma = 0$: all streamlines are parallel straight lines;
(ii) $\gamma < 0$: all streamlines are straight lines ending in the point $(-\alpha/\gamma,-\beta/\gamma)^T$;
(iii) $\gamma > 0$: all streamlines are straight lines starting in the point $(-\alpha/\gamma,-\beta/\gamma)^T$.

If $\gamma < 0$ $(\gamma > 0)$, then $(-\alpha/\gamma,-\beta/\gamma)^T \in \hat{S}$ if, and only if, the outward normal component of $\hat{w}$ is negative (positive) on the entire boundary $\partial\hat{S}$.

One can consider (9.3) as an ordinary differential equation, i.e.

$$(9.4a) \qquad \hat{x}'(t) = \binom{\alpha}{\beta} + \gamma\, \hat{x}(t),\ t > 0.$$

The initial value is given by the point of entrance, i.e.

$$(9.4b) \qquad \hat{x}(0) = \hat{x}_{in} = (\hat{x}_{in},\hat{y}_{in})^T.$$

The solution of (9.4) is equal to

$$(9.5) \qquad \hat{x}(t) = e^{\gamma t}\, \hat{x}_{in} + t\, E(\gamma t)\, \binom{\alpha}{\beta},\ t \geq 0,$$

where the function E is defined as

$$(9.6) \qquad E(\varsigma) = \begin{cases} (e^{\varsigma} - 1)/\varsigma & \text{if } \varsigma \neq 0, \\ 1 & \text{if } \varsigma = 0. \end{cases}$$

If $\gamma = 0$, or $\gamma \neq 0$ and $(-\alpha/\gamma, -\beta/\gamma) \notin \hat{S}$, then the residence time Δt between $\hat{x}_{in}$ and the point of exit $\hat{x}_{out} = (\hat{x}_{out}, \hat{y}_{out})^T$ can be determined by intersecting the streamline through $\hat{x}_{in}$ with each of the edges $\hat{e}_i$ of $\hat{S}$. If, e.g., $\hat{x}_{out} \in \hat{e}_1$, then

$$(9.7) \qquad \Delta t = \frac{1 - \hat{x}_{in} - \hat{y}_{in}}{\alpha + \beta + \gamma(\hat{x}_{in} + \hat{y}_{in})} \; L \left(\frac{\gamma(1 - \hat{x}_{in} - \hat{y}_{in})}{\alpha + \beta + \gamma(\hat{x}_{in} + \hat{y}_{in})} \right),$$

where the function L is defined as

$$(9.8) \qquad L(\varsigma) = \begin{cases} \log(1 + \varsigma)/\varsigma & \text{if } \varsigma > -1 \text{ and } \varsigma \neq 0, \\ 1 & \text{if } \varsigma = 0. \end{cases}$$

Clearly, $\hat{x}_{out} = e^{\gamma\Delta t} + \Delta t \; E(\gamma\Delta t) \binom{\alpha}{\beta}$. Similar expressions can be derived for $\hat{e}_2$ and $\hat{e}_3$. Obviously, if $\hat{x}_{in} \in \hat{e}_1$, then $t = 0$ and $\hat{x}_{out} = \hat{x}_{in}$. Since $\hat{x}_{in} \in \partial\hat{S}$ and therefore $\hat{x}_{in} + \hat{y}_{in} \leq 1$, it follows immediately from (9.7) that $\Delta t > 0$ if, and only if, $\alpha + \beta + \gamma > 0$ and $\alpha + \beta + \gamma(\hat{x}_{in} + \hat{y}_{in}) > 0$.

In order to determine $\Delta t > 0$ two conditions have to be verified per edge $\hat{e}_i$. After deducing similar formulas for $\hat{e}_2$ and $\hat{e}_3$ we obtain

Algorithm "triangle" :

```
if ($\alpha + \beta + \gamma > 0$) and ($\alpha + \beta + \gamma(\hat{x}_{in} + \hat{y}_{in}) > 0$) then "$\hat{e}_{out} = \hat{e}_1$"
    $\Delta t := \frac{1 - \hat{x}_{in} - \hat{y}_{in}}{\alpha + \beta + \gamma(\hat{x}_{in} + \hat{y}_{in})} \; L \left( \frac{\gamma(1 - \hat{x}_{in} - \hat{y}_{in})}{\alpha + \beta + \gamma(\hat{x}_{in} + \hat{y}_{in})} \right)$
    $\hat{x}_{out} := e^{\gamma\Delta t} x_{in} + \Delta t \; E(\gamma\Delta t) \; \alpha$
    $\hat{y}_{out} := 1 - \hat{x}_{out}$
    if ($0 \leq \hat{x}_{out} \leq 1$) then stop
if ($\alpha < 0$) and ($\alpha + \gamma\hat{x}_{in} < 0$) then "$\hat{e}_{out} = \hat{e}_2$"
    $\Delta t := \frac{-\hat{x}_{in}}{\alpha + \gamma\hat{x}_{in}} \; L \; \left(\frac{-\gamma\hat{x}_{in}}{\alpha + \gamma\hat{x}_{in}}\right)$
    $\hat{x}_{out} := 0$
    $\hat{y}_{out} := e^{\gamma\Delta t} \, \hat{y}_{in} + \Delta t \; E(\gamma\Delta t) \; \beta$
    if ($0 \le y_{out} \le 1$) then stop
if ($\beta < 0$) and ($\beta + \gamma\hat{y}_{in} < 0$) then "$\hat{e}_{out} = \hat{e}_3$"
    $\Delta t := \frac{-\hat{y}_{in}}{\beta + \gamma\hat{y}_{in}} \; L \; \left(\frac{-\gamma\hat{y}_{in}}{\beta + \gamma\hat{y}_{in}}\right)$
    $\hat{x}_{out} := e^{\gamma\Delta t} \, \hat{x}_{in} + \Delta t \; E(\gamma\Delta t) \; \alpha$
    $\hat{y}_{out} := 0$
    if ($0 \le \hat{x}_{out} \le 1$) then stop.
```

If u_h is divergence-free in a certain subdomain $S \in \mathrm{S}_h$, then $\gamma = 0$, and the algorithm can be simplified considerably. Recall that from a physical point of view it is reasonable to assume that u_h is divergence-free in most of the subdomains.

Parallelogram (see figure 1b)

Let $\hat{u} \in RT^0(\hat{S})$, then $\hat{w} = \hat{u}/\hat{p}$ is of the form

$$(9.9) \qquad \hat{w}(\hat{x}) = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}\gamma & 0\\ 0 & \delta\end{pmatrix} \hat{x}, \; \hat{x} \in \hat{S}.$$

One can consider (9.9) as an ordinary differential equation, i.e.

$$(9.10a) \qquad \hat{x}'(t) = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}\gamma & 0\\ 0 & \delta\end{pmatrix} \hat{x}(t), \; t > 0,$$

$$(9.10b) \qquad \hat{x}(0) = \hat{x}_{in}.$$

The solution of (9.10) is equal to

$$\hat{x}(t) = \begin{bmatrix} e^{\gamma t}\,\hat{x}_{in} \\ e^{\delta t}\,\hat{y}_{in} \end{bmatrix} + t \begin{bmatrix} E(\gamma t)\,\alpha \\ E(\delta t)\,\beta \end{bmatrix}, \; t \geq 0. \tag{9.11}$$

If $\gamma = 0$, or $\gamma \neq 0$ and $0 \leq -\alpha/\gamma \leq 1$, and $\delta = 0$, or $\delta \neq 0$ and $0 \leq -\beta/\delta \leq 1$, then the residence time Δt between $\hat{x}_{in}$ and $\hat{x}_{out}$ can be determined by intersecting the streamline through $\hat{x}_{in}$ with each of the edges $\hat{e}_i$ of $\hat{S}$.

In order to determine $\Delta t > 0$ two conditions have to be verified per edge $\hat{e}_i$. We obtain

```
Algorithm "parallelogram":
if $(\alpha + \gamma > 0)$ and $(\alpha + \gamma \hat{x}_{in} > 0)$ then "$\hat{e}_{out} = \hat{e}_1$"
    $\Delta t := \frac{1 - \hat{x}_{in}}{\alpha + \gamma \hat{x}_{in}} \; L\left(\frac{\gamma(1 - \hat{x}_{in})}{\alpha + \gamma \hat{x}_{in}}\right)$
    $\hat{x}_{out} := 1$
    $\hat{y}_{out} := e^{\delta \Delta t}\, \hat{y}_{in} + \Delta t\, E(\delta \Delta t)\, \beta$
    if $(0 \leq \hat{y}_{out} \leq 1)$ then stop
if $(\beta + \delta > 0)$ and $(\beta + \delta \hat{y}_{in} > 0)$ then "$\hat{e}_{out} = \hat{e}_2$"
    $\Delta t := \frac{1 - \hat{y}_{in}}{\beta + \delta \hat{y}_{in}} \; L\left(\frac{\delta(1 - \hat{y}_{in})}{\beta + \delta \hat{y}_{in}}\right)$
    $\hat{x}_{out} := e^{\gamma \Delta t}\, \hat{x}_{in} + \Delta t\, E(\gamma \Delta t)\, \alpha$
    $\hat{y}_{out} := 1$
    if $(0 \leq \hat{x}_{out} \leq 1)$ then stop
if $(\alpha < 0)$ and $(\alpha + \gamma \hat{x}_{in} < 0)$ then "$\hat{e}_{out} = \hat{e}_3$"
    $\Delta t := \frac{- \hat{x}_{in}}{\alpha + \gamma \hat{x}_{in}} \; L\left(\frac{- \gamma \hat{x}_{in}}{\alpha + \gamma \hat{x}_{in}}\right)$
    $\hat{x}_{out} := 0$
    $\hat{y}_{out} := e^{\delta \Delta t}\, \hat{y}_{in} + \Delta t\, E(\delta \Delta t)\, \beta$
    if $(0 \leq \hat{y}_{out} \leq 1)$ then stop
if ($\beta < 0$) and ($\beta + \delta\hat{y}_{in} < 0$) then "$\hat{e}_{out} = \hat{e}_4$"
    $\Delta t := \frac{-\hat{y}_{in}}{\beta + \delta\hat{y}_{in}} \; L\left(\frac{-\delta\hat{y}_{in}}{\beta + \delta\hat{y}_{in}}\right)$
    $\hat{x}_{out} := e^{\gamma\Delta t} + \Delta t\; E(\gamma\Delta t)\; \alpha$
    $\hat{y}_{out} := 0$
    if ($0 \le \hat{x}_{out} \le 1$) then stop.
```

If u_h is divergence-free in a certain subdomain $S \in S_h$, then $\gamma + \delta = 0$. Now the algorithm in general cannot be simplified.

Three-dimensional elements (see figure 2)

Since algorithms for tetrahedrons, prisms and parallelepipeds can be derived analogously, they will not be presented here.

Global streamlines

So far, we have only considered the computation of streamlines and residence times on the reference element $\hat{S}$. On an arbitrary element $S \in S_h$ one has to solve the ordinary differential equation

$$(9.12a) \quad x'(t) = w(x(t)),\ t > 0,$$

$$(9.12b) \quad x(0) = x_{in}.$$

Since $x(t) = B\hat{x}(t) + b$, $w(x(t)) = B\hat{w}(\hat{x}(t))$ for $t \ge 0$, and $x_{in} = B\hat{x}_{in} + b$, the differential equation (9.12) is equivalent to

$$(9.13a) \quad \hat{x}'(t) = \hat{w}(\hat{x}(t)),\ t > 0,$$

$$(9.13b) \quad \hat{x}(0) = \hat{x}_{in}.$$

Therefore,

$$(9.14)\quad x_{out} = B\hat{x}_{out} + b.$$

By computing streamlines and residence times for a chain of elements an array of points and residence times is obtained. A global streamline is obtained by interconnecting the consecutive points and the total residence time by summing all residence times.

10. Numerical experiments.

In the preceding sections the mixed finite element method, its hybridization and the computation of streamlines and residence times were presented. In this section the applicability and advantages of the mixed finite element method and the convergence of the preconditioned conjugate gradient method are illustrated by some numerical experiment.

We are especially interested in potential flow problems with sources and sinks or with large jumps in the tensor of hydraulic conductivity. We are also interested in triangulations of the domain into very flat subdomains.

All the computations presented below were done in double precision on an Alliant FX/40.

Square model problems

In this section four two-dimensional elliptic boundary value problems of the form (3.1) are considered on the unit square $\Omega = \{x = (x_1,x_2)^T \mid 0 < x_i < 1, i = 1,2\}$. These problems were solved using the lowest order mixed-hybrid finite element method and, to serve as a comparison, the lowest order conforming finite element method (see [9: chapter 2]).

A regular triangulation of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of $2M^2$ triangles or M^2 squares, where M is an even natural number (see

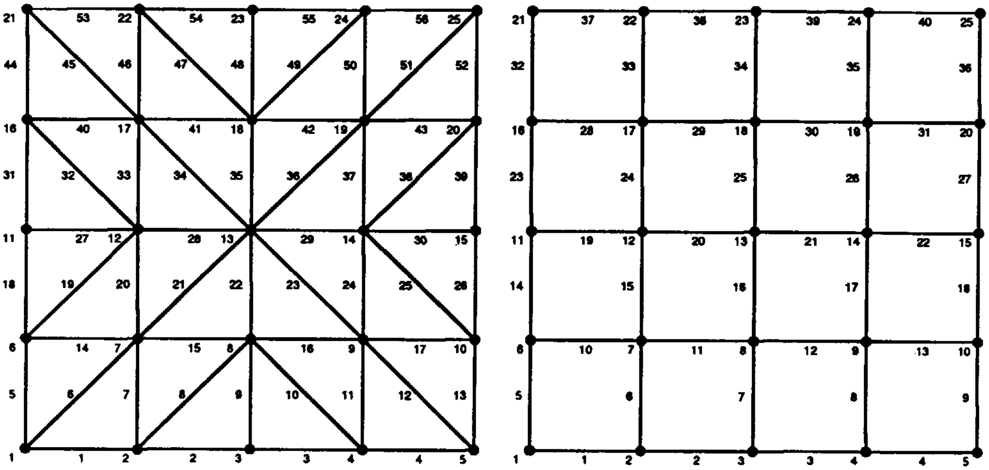


Figure 3. Triangulation of the unit square ($M = 4$).

Table I. Amount of work (flops) per iteration of the preconditioned conjugate gradient method for the square model problems.

		n	m	$7m + 4n$
conforming method	triangles squares	$(M+1)^2 \sim M^2$ $(M+1)^2 \sim M^2$	$M(3M+2) \sim 3M^2$ $2M(2M+1) \sim 4M^2$	$\sim 19M^2$ $\sim 23M^2$
mixed-hybrid method	triangles squares	$M(3M+2) \sim 3M^2$ $2M(M+1) \sim 2M^2$	$6M^2$ $6M^2$	$\sim 45M^2$ $\sim 38M^2$

figure 3). The conforming method results in an approximation of the potential ϕ, which has one degree of freedom per vertex of the mesh. These vertices were numbered lexicographically from bottom to top and from left to right. The mixed-hybrid method results in the Lagrange multiplier λ_h, which has one degree of freedom per edge of the mesh. These edges were also numbered lexicographically.

The resulting systems of linear equations were solved by the preconditioned conjugate gradient method (see, e.g., [12: chapter 10]), where the preconditioning matrices were constructed by the incomplete Cholesky decomposition (see [22], [23]) or the modified incomplete Cholesky decomposition (see [13], [3: section 1.4]). Let a certain resulting system be denoted as $Ax = b$ for the time being. A sequence $x_1,x_2,...$ was generated, starting with the vector $x_0 = 0$. We terminated if $\| x - x_i \|_A / \| x \|_A \leq 10^{-6}$ (for the corresponding termination criterion, see [20]). The amount of work per iteration is equal to three inner products, two vector updates and two matrix-vector multiplications, i.e. $7n + 4m$ flops, where n is the number of unknowns and $n + 2m$ is the number of nonzero entries of the matrix A (see table I). About twice as much work is required per iteration for the mixed-hybrid method as for the conforming method.

If the medium is isotropic, i.e. if the tensor A is equal to a scalar function times the unit tensor, then the resulting matrix is a symmetric weakly diagonally dominant M-matrix for the conforming method using triangles or squares and for the mixed-hybrid method using triangles. Therefore, preconditioning matrices can be constructed by the incomplete Cholesky decomposition and the modified incomplete Cholesky decomposition.

For the mixed-hybrid method using squares the resulting matrix is not an M-matrix. However, in the following experiments it appeared that preconditioning matrices can still be constructed using the incomplete Cholesky decomposition and the modified incomplete Cholesky decomposition.

Using the conforming finite element method, a piecewise linear or bilinear approximation of ϕ is determined. This approximation is differentiated in each subdomain and multiplied by the tensor A to obtain an approximation of $\boldsymbol{u}$. At this stage the L^2-norm of the differences of ϕ and $\boldsymbol{u}$ and their respective approximations were computed if ϕ and u were known explicitly. The divergence of the approximation of u is equal to zero in each subdomain, even if the source term f is not equal to zero in such a subdomain. Moreover, the normal component of this approximation is generally not continuous across the interelement boundaries.

Using the mixed-hybrid finite element method, the Lagrange multiplier λ_h is determined. From λ_h the approximations ϕ_h and u_h are obtained using (7.15) and (7.13), respectively. We computed $\| \phi - \phi_h \|_{0,\Omega}$ and $\| u - u_h \|_{0,\Omega}$ if ϕ and u were known explicitly. Now $\nabla \cdot u_h = f$ if the function f is piecewise constant (use (6.8b)). Moreover, the normal component of u_h is continuous across the interelement boundaries.

Streamlines were computed by the method that is exposed in section 9. Using the conforming method, first the approximation of u was improved by introducing a quasi Lagrange multiplier. This quasi multiplier is defined on the union of all edges and is constant on each edge, as in the mixed-hybrid case. On each edge its value is equal to the value of the conforming approximation of ϕ at the midpoint of the edge. From this quasi multiplier outward fluxes across the edges of all subdomains were obtained using (7.15) and (7.13). On each interelement edge we took the mean of the two corresponding fluxes. Finally, an approximation of u was computed from these fluxes. Of course, the normal component of this approximation is continuous across the interelement boundaries. Still, its divergence is generally not equal to f. However, the computed streamlines appeared to be improved by this ad hoc technique.

The Tóth model problem

Let $\Omega = \{x = (x_1,x_2)^T \mid 0 < x_i < 1,\ i = 1,2\}$. Consider the Tóth model problem (see [33])

$$\text{(10.1)}\qquad \boxed{\begin{array}{l} -\Delta\phi = 0 \ \text{ in } \Omega, \\ \phi(x) = \cos(\pi x_1) \ \text{ if } x \in \partial\Omega_D, \ -\dfrac{\partial\phi}{\partial n} = 0 \ \text{ on } \partial\Omega_N, \end{array}}$$

where

$$\text{(10.2)}\qquad \partial\Omega_D = \{x \in \partial\Omega \mid x_2 = 1\},\ \partial\Omega_N = \partial\Omega\backslash\partial\Omega_D.$$

The solution of this boundary value problem is equal to

$$\text{(10.3)}\qquad \phi(x) = [\cosh(\pi(1 - x_2)) - \tanh(\pi)\sinh(\pi(1 - x_2))]\cos(\pi x_1),\ x \in \Omega.$$

Therefore, $\boldsymbol{u} = -\nabla\phi$ is given by

$$\text{(10.4)}\quad \boldsymbol{u}(x) = \begin{bmatrix} \pi\,[\cosh(\pi(1-x_2)) - \tanh(\pi)\sinh(\pi(1-x_2))]\sin(\pi x_1) \\ \pi\,[\sinh(\pi(1-x_2)) - \tanh(\pi)\cosh(\pi(1-x_2))]\cos(\pi x_1) \end{bmatrix},\ x \in \Omega.$$

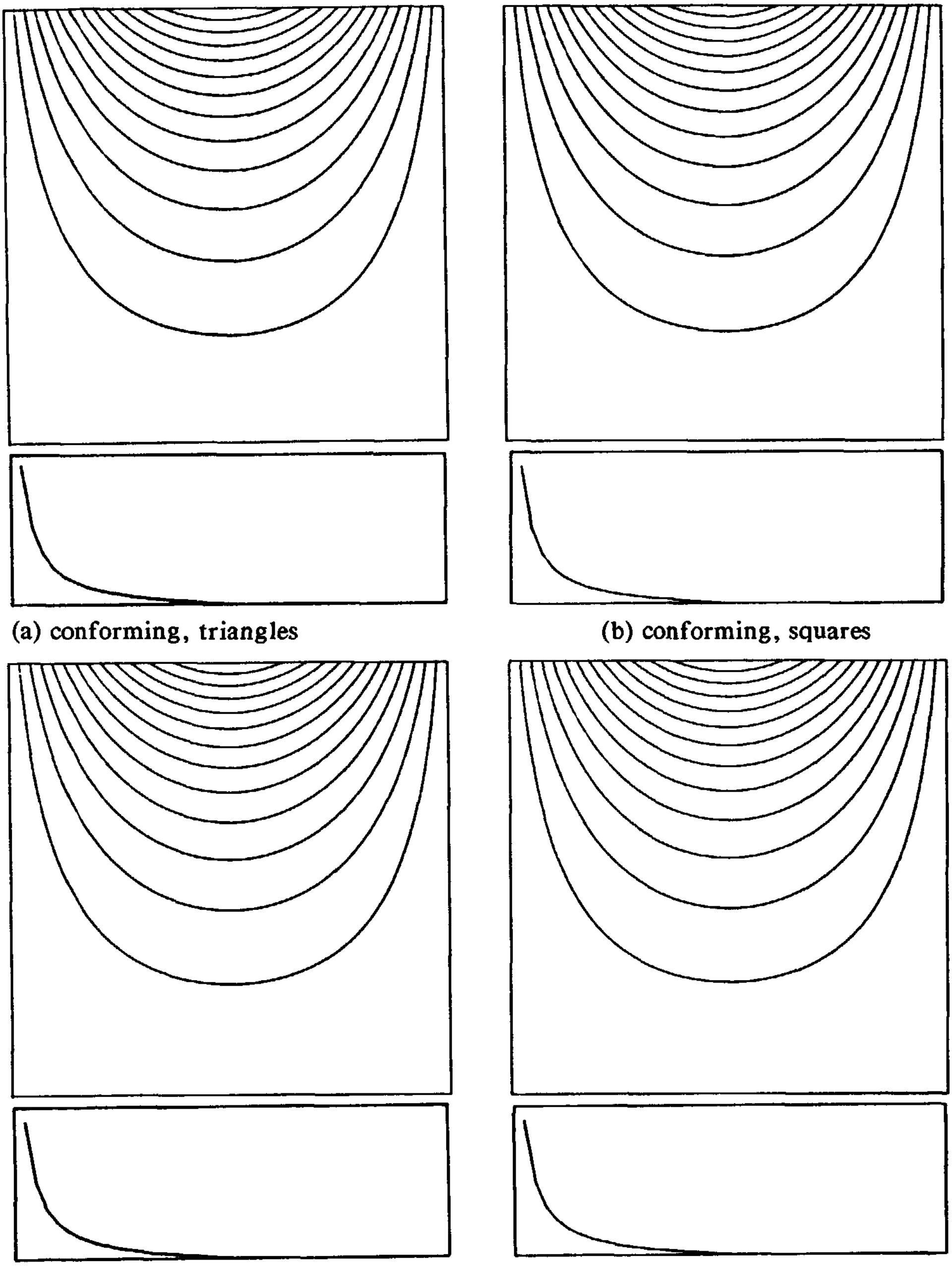
A regular triangulation of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of $2M^2$ triangles or M^2 squares (see figure 3), where $M = 2^k$, $k = 2,...,6$. The Tóth model problem was solved using the lowest order conforming and mixed-hybrid finite element method. The resulting systems of linear equations were solved by the preconditioned conjugate gradient method, where the preconditioning matrices were constructed by the incomplete Cholesky decomposition (IC) and the modified incomplete Cholesky decomposition (MIC).

The discretization errors and the numbers of iterations necessary to fulfil the desired termination criterion are displayed in table II. Note that $h = \sqrt{2}/M$ using triangles and $h = 1/M$ using squares.

The numerical results confirm the theoretical error estimates for the conforming finite element method (see [9: section 3.2]) and the mixed finite element method (see note 6.2). Further, it appears that the number of iterations is proportional to h^{-1} for the incomplete Cholesky decomposition and to $h^{-1/2}$ for the modified incomplete Cholesky decomposition (for discussions on this phenomenon,

Table II. Discretization errors and numbers of iterations for the Tóth model problem.

		M	n	$\Vert u - u_h \Vert_{0,\Omega}$	$\Vert \phi - \phi_h \Vert_{0,\Omega}$	IC	MIC
Conforming method	triangles	4	25	0.473	$0.23 \cdot 10^{-1}$	8	8
		8	81	0.242	$0.62 \cdot 10^{-2}$	12	11
		16	289	0.122	$0.16 \cdot 10^{-2}$	21	17
		32	1089	0.061	$0.39 \cdot 10^{-3}$	40	25
		64	4225	0.031	$0.99 \cdot 10^{-4}$	79	37
	squares	4	25	0.390	$0.15 \cdot 10^{-1}$	5	6
		8	81	0.199	$0.37 \cdot 10^{-2}$	10	9
		16	289	0.100	$0.93 \cdot 10^{-3}$	16	13
		32	1089	0.050	$0.23 \cdot 10^{-3}$	31	19
		64	4225	0.025	$0.58 \cdot 10^{-4}$	58	28
Mixed-hybrid method	triangles	4	56	0.464	0.0833	12	10
		8	208	0.243	0.0419	19	16
		16	800	0.123	0.0209	35	23
		32	3136	0.062	0.0105	69	32
		64	12416	0.031	0.0052	139	47
	squares	4	40	0.282	0.0877	8	8
		8	144	0.140	0.0448	5	11
		16	544	0.070	0.0225	28	16
		32	2112	0.035	0.0113	54	22
		64	8320	0.018	0.0056	106	31



(a) conforming, triangles

(b) conforming, squares

(c) mixed, triangles

(d) mixed, squares

Figure 4. Streamlines and residence times for the Tóth model problem.

see [14: section 4], [3: section 7.2]).

Streamlines were computed for $M = 50$. Note that $\boldsymbol{n}\cdot\boldsymbol{u}(x) = -\pi \tanh(\pi) \cos(\pi x_1)$, $x \in \partial\Omega_D$. Thus, $\boldsymbol{n}\cdot\boldsymbol{u} < 0$ on $\partial\Omega_D^- = \{x \in \partial\Omega_D \,|\, x_1 < 1/2\}$ and $\boldsymbol{n}\cdot\boldsymbol{u} > 0$ on $\partial\Omega_D^+ = \{x \in \partial\Omega_D \,|\, x_1 > 1/2\}$. The outward flux across $\partial\Omega_D^-$ is equal to $\int_{\partial\Omega_D^-} \boldsymbol{n}\cdot\boldsymbol{u}\, ds = -\tanh(\pi)$. We computed 12 streamlines starting on $\partial\Omega_D^-$ and ending on $\partial\Omega_D^+$, such that the flow rate is equal between all adjacent streamlines. The residence times between $\partial\Omega_D^-$ and $\partial\Omega_D^+$ were also computed. The streamlines and residence times are displayed in figure 4. Note that the residence time is plotted as a function of the x_1-coordinate of the starting point of the corresponding streamline. There are only minor differences.

The Muskat model problem

Let $\Omega = \{x = (x_1,x_2)^T \,|\, 0 < x_i < 1,\ i = 1,2\}$. Consider the Muskat model problem (see [25: figure 12.9])

$$(10.5)\qquad \boxed{\begin{array}{l} -\Delta\phi = f \text{ in } \Omega, \\ \phi = 0 \text{ on } \partial\Omega_D,\ \ -\dfrac{\partial\phi}{\partial n} = 0 \text{ on } \partial\Omega_N, \end{array}}$$

where $\partial\Omega_D$ and $\partial\Omega_N$ are defined in (10.2). Here $f \in H^{-1}(\Omega)$ is defined as

$$(10.6)\qquad f(x) = -2\delta\,(x_1 - 1/2, x_2),\ x \in \Omega,$$

where δ is the Dirac function. The function f represents a sink with a discharge $Q = 2$. Thus, the outward flux across $\partial\Omega_D$ is equal to -1. Now, a function $\phi \in H_D^1(\Omega)$ (see (2.5a)) is the unique weak solution of (10.5) if

$$(10.7)\qquad \int_\Omega \nabla\phi\cdot\nabla\psi\, dx = -2\psi\,(1/2,0)\ \ \forall\psi \in H_D^1(\Omega)$$

(see [15: chapter 7]). Since $f \notin L^2(\Omega)$, a mixed variational formulation of (10.5) of the form (3.4) cannot be stated. Therefore, for $\epsilon > 0$ small, we replaced definition (10.6) by

$$(10.8) \qquad f(x) = \begin{cases} \dfrac{-1}{2\epsilon^2} & \text{if } x \in \Omega_\epsilon = \{x \in \Omega \mid 1/2 - \epsilon < x_1 < 1/2 + \epsilon,\ x_2 < \epsilon\}, \\ 0 & \text{if } x \in \Omega \backslash \Omega_\epsilon. \end{cases}$$

For the numerical experiments we chose $\epsilon = 1/50$. A regular triangulation of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of 5000 triangles or 2500 squares, i.e. $M = 50$. The Muskat model problem was solved using the lowest order conforming and mixed-hybrid finite element method. The resulting systems of linear equations were solved by the preconditioned conjugate gradient method.

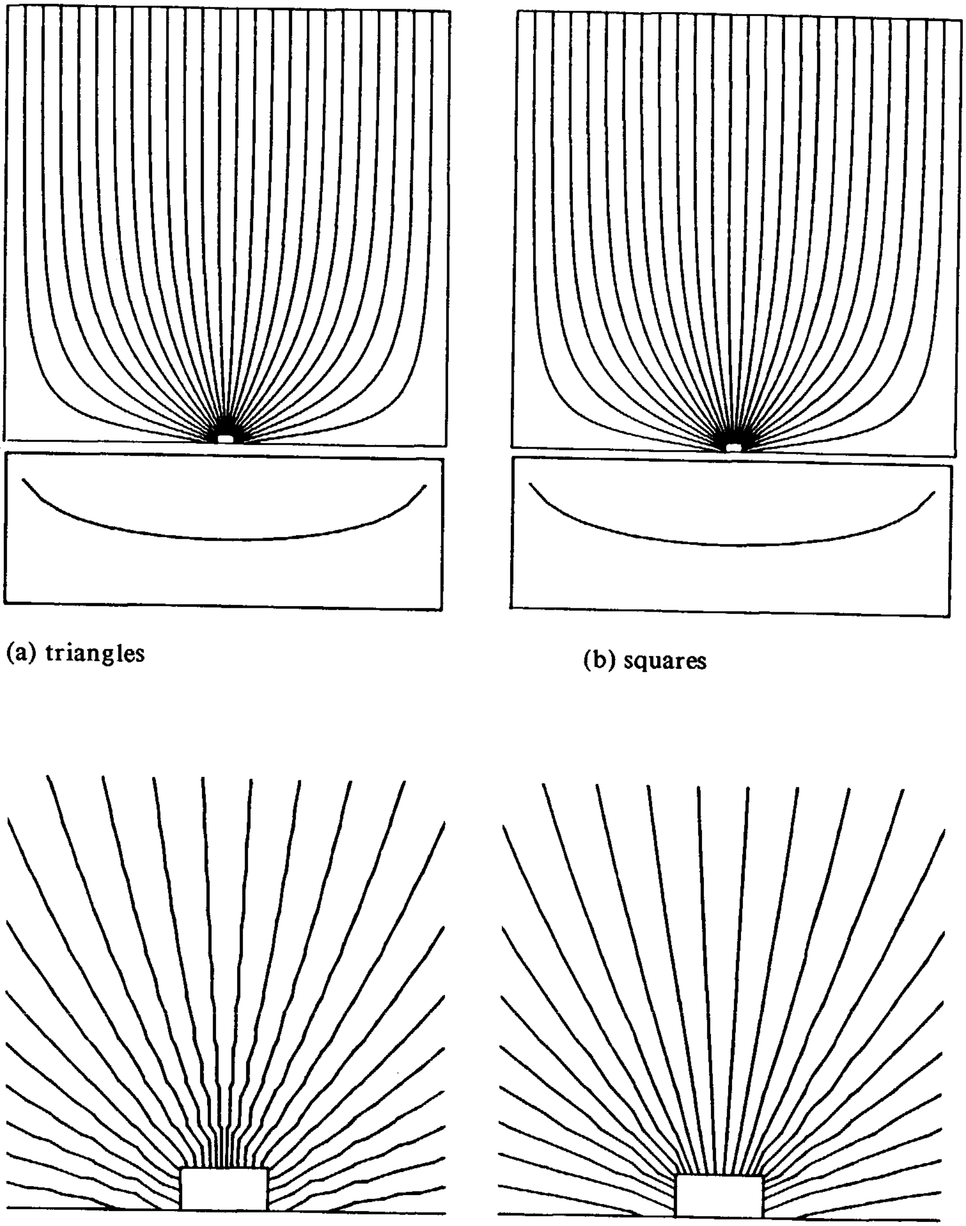
The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table III.

Table III. Numbers of iterations for the Muskat model problem.

		n	IC	MIC
conforming method	triangles squares	2601 2601	64 50	36 27
mixed-hybrid method	triangles squares	7601 5100	112 93	45 32

We have computed 24 streamlines starting on $\partial\Omega_D$ and ending on the boundary of $\Omega_{1/50}$, such that the flow rate is equal between all adjacent streamlines. The corresponding residence times were also computed. The streamlines and residence times are displayed in figures 5 and 6.

Using the conforming finite element method, some streamlines end on $\partial\Omega_N$



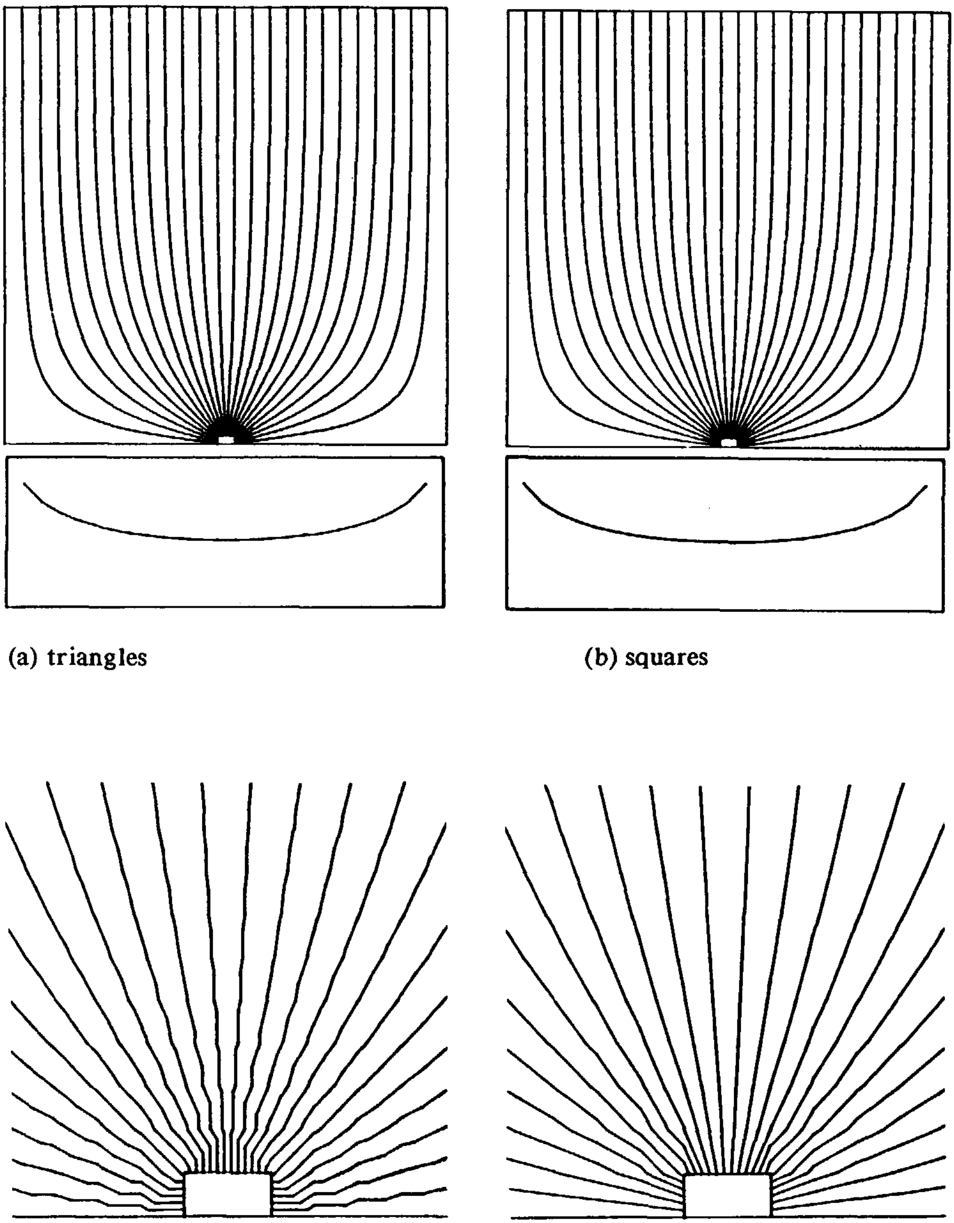
(a) triangles

(b) squares

(c) triangles, enlargement

(d) squares, enlargement

Figure 5. Streamlines and residence times for the Muskat model problem using the conforming finite element method.



(a) triangles

(b) squares

(c) triangles, enlargement

(d) squares, enlargement

Figure 6. Streamlines and residence times for the Muskat model problem using the mixed finite element method.

instead of the boundary of $\Omega_{1/50}$. Using the mixed finite element method, all streamlines end on the boundary of $\Omega_{1/50}$. Of course, one should choose $\epsilon > 0$ smaller, and refine near the well to obtain superior results. An alternative approach is to use macroelements (see note 9.1). There are only minor differences between the residence times.

The Philip model problem

Let $\Omega = \{x = (x_1,x_2)^T \mid 0 < x_i < 1,\ i = 1,2\}$. For $0 \le \epsilon < 1$, consider the Philip model problem (see [28: section 2.2])

$$(10.9)\qquad \boxed{\begin{array}{l} -\nabla\cdot(a\nabla\phi) = 0 \quad \text{in } \Omega, \\ \phi = g_D \quad \text{on } \partial\Omega_D, \quad -\boldsymbol{n}\cdot(a\nabla\phi) = 0 \quad \text{on } \partial\Omega_N, \end{array}}$$

where

$$(10.10)\qquad a(x) = [1 + 2\epsilon\cos(\pi x_1)\cos(\pi x_2) + \epsilon^2\cos^2(\pi x_2)]^{-1},\ x \in \Omega,$$

$$(10.11)\qquad \partial\Omega_D = \{x \in \partial\Omega_D \mid x_2 = 0 \text{ or } x_2 = 1\},\ \partial\Omega_N = \partial\Omega\backslash\partial\Omega_D,$$

$$(10.12)\qquad g_D(x) = \pi(1 - x_2),\ x \in \partial\Omega.$$

Note that a has values from $(1 - \epsilon)^{-2}$ to $(1 + \epsilon)^{-2}$. The solution of this boundary value problem is equal to

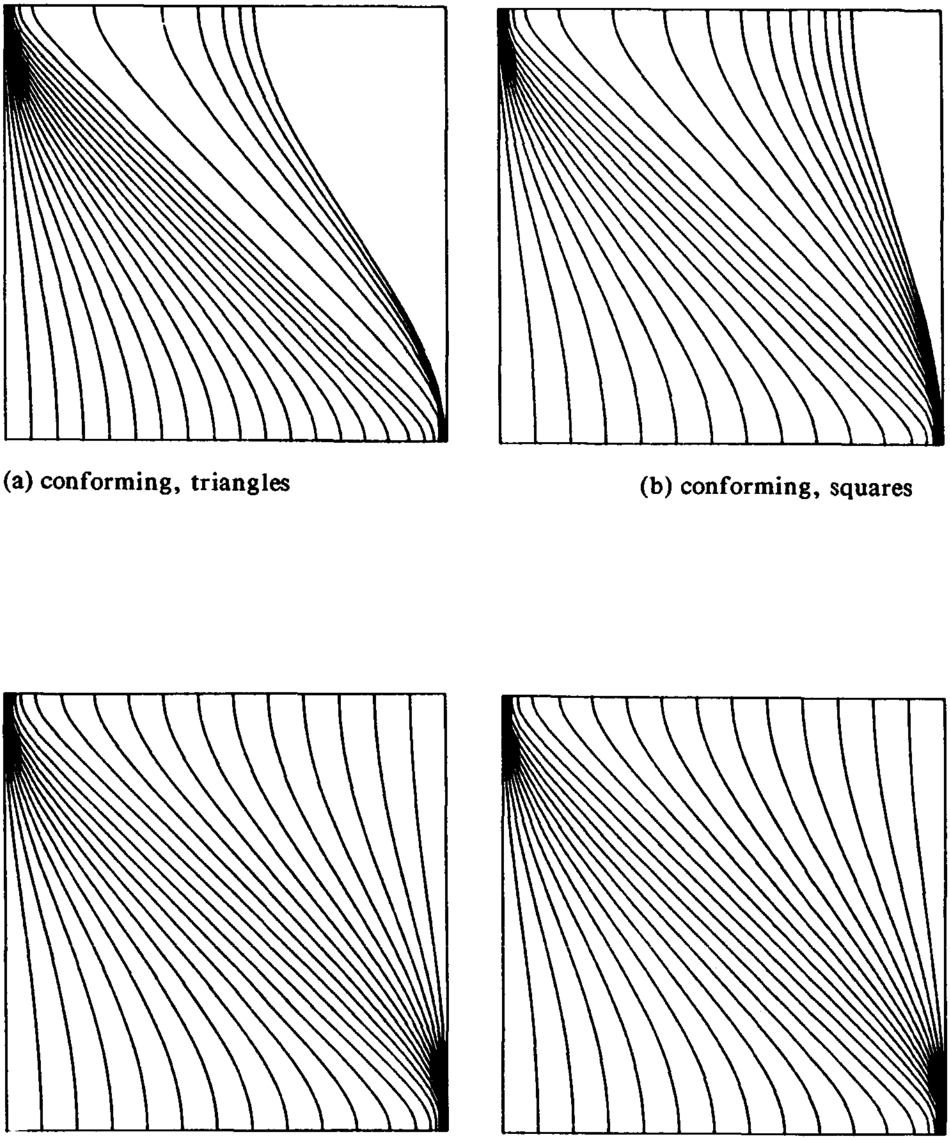
$$(10.13)\qquad \phi(x) = \pi(1 - x_2) - \epsilon\cos(\pi x_1)\sin(\pi x_2),\ x \in \partial\Omega.$$

Therefore $\boldsymbol{u} = -a\nabla\phi$ is given by

$$(10.14)\qquad \boldsymbol{u}(x) = -a(x)\begin{bmatrix} \pi\epsilon\sin(\pi x_1)\sin(\pi x_2) \\ -\pi\,[1 + \epsilon\cos(\pi x_1)\cos(\pi x_2)] \end{bmatrix},\ x \in \Omega.$$

Table IV. Discretization errors and numbers of iterations for the Philip model problem.

		M	n	$\Vert u - u_h \Vert_{0,\Omega}$			$\epsilon = 0.999$		
				ϵ=0.9	ϵ=0.99	ϵ=0.999	$\Vert \phi - \phi_h \Vert_{0,\Omega}$	IC	MIC
Conforming method	triangles	4	25	1.88	2.31	2.44	0.32	6	6
		8	81	1.43	2.91	3.14	$0.90 \cdot 10^{-1}$	9	9
		16	289	0.82	3.24	3.23	$0.26 \cdot 10^{-1}$	15	12
		32	1089	0.41	3.24	3.56	$0.78 \cdot 10^{-2}$	28	17
		64	4225	0.20	2.58	4.39	$0.24 \cdot 10^{-2}$	54	24
	squares	4	25	2.07	2.43	2.51	0.23	4	4
		8	81	1.54	2.63	2.64	$0.71 \cdot 10^{-1}$	6	6
		16	289	0.86	3.09	3.00	$0.24 \cdot 10^{-1}$	12	9
		32	1089	0.42	3.53	3.48	$0.79 \cdot 10^{-2}$	21	12
		64	4225	0.21	3.07	4.11	$0.27 \cdot 10^{-2}$	44	18
Mixed-hybrid method	triangles	4	56	2.24	2.85	2.96	0.730	9	8
		8	208	1.65	3.41	3.60	0.357	14	13
		16	800	0.96	3.78	3.71	0.175	26	18
		32	3136	0.47	3.80	4.10	0.087	50	24
		64	12416	0.23	3.07	5.04	0.043	98	34
	squares	4	40	2.06	2.45	2.53	0.887	6	6
		8	144	1.51	2.61	2.62	0.441	11	7
		16	544	0.82	3.08	2.99	0.219	21	10
		32	2112	0.39	3.52	3.47	0.109	41	13
		64	8320	0.19	3.07	4.11	0.055	84	19



(a) conforming, triangles

(b) conforming, squares

(c) mixed, triangles

(d) mixed, squares

Figure 7. Streamlines for the Philip model problem.

For the numerical experiments we chose $\epsilon = 0.9$, $\epsilon = 0.99$ and $\epsilon = 0.999$. A regular partitioning of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of $2M^2$ triangles or M^2 squares, where $M = 2^k$, $k = 2,...,6$. The Philip model problem was solved using the lowest order conforming and mixed-hybrid finite element method. The resulting systems of linear equations were solved by the preconditioned conjugate gradient method.

The discretization errors of $\boldsymbol{u}$ for $\epsilon = 0.9$, $\epsilon = 0.99$ and $\epsilon = 0.999$, and the discretization errors of ϕ and the numbers of iterations necessary to fulfil the desired termination criterion only for $\epsilon = 0.999$ are displayed in table IV.

The numerical results confirm the theoretical error estimates for the conforming and mixed finite element method, but it seems that if $\epsilon \approx 1$ convergence for $\boldsymbol{u}_h$ is only achieved for very small mesh sizes (see note 6.2). Therefore, one should refine near the upper left and lower right corners of Ω to obtain superior results, because the largest variation of a occurs in these regions (see also figure 7). Again, it appears that the number of iterations is proportional to h^{-1} for the incomplete Cholesky decomposition and to $h^{-1/2}$ for the modified incomplete Cholesky decomposition.

Streamlines were computed for $M = 50$. Note that $\boldsymbol{n} \cdot \boldsymbol{u} < 0$ on the bottom boundary and $\boldsymbol{n} \cdot \boldsymbol{u} > 0$ on the top boundary. The outward flux across the bottom boundary is equal to -1 (this follows immediately by computing the flux across $\{x \in \Omega \,|\, x_2 = 1/2\}$ and bearing in mind that $\nabla \cdot \boldsymbol{u} = 0$ in Ω). We computed 24 streamlines starting on the bottom boundary and ending on the top boundary, such that the flow rate is equal between all adjacent streamlines. The streamlines are displayed in figure 7. Clearly, the results of the mixed finite element method are superior.

The square layer model problem

Let $\Omega = \{x = (x_1,x_2)^T \,|\, 0 < x_i < 1,\ i = 1,2\}$. Consider the square layer model problem

$$(10.15)\qquad \boxed{\begin{array}{l} -\nabla\cdot(a\nabla\phi) = 0 \ \text{ in } \Omega \\ \phi = g_D \ \text{ on } \partial\Omega_D, \ -\boldsymbol{n}\cdot(a\nabla\phi) = 0 \ \text{ on } \partial\Omega_N, \end{array}}$$

where

$$(10.16)\qquad a(x) = \begin{cases} 10^{-5} & \text{if } x \in \Omega,\ x_1 < 0.8 \text{ and } 0.58 < x_2 < 0.60, \\ & \quad\text{or } x_1 > 0.2 \text{ and } 0.40 < x_2 < 0.42, \\ 1 & \text{if } x \text{ is elsewhere in } \Omega, \end{cases}$$

$$(10.17)\qquad g_D(x) = 1 - x_2,\ x \in \partial\Omega,$$

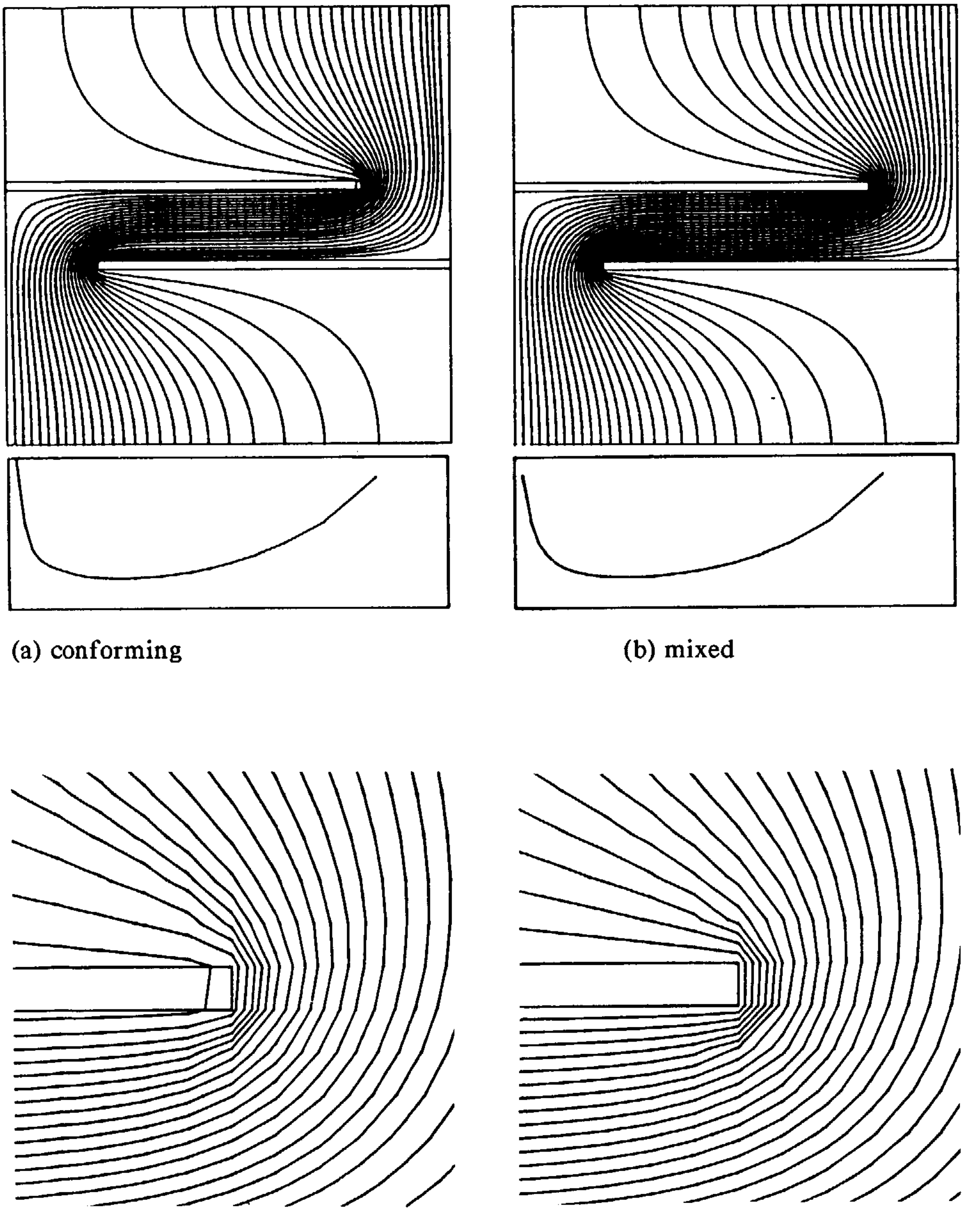
$\partial\Omega_D$ and $\partial\Omega_N$ are defined in (10.11). One can interpret (10.16) as the description of a sandy porous medium with two clay layers.

A regular triangulation of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of 2500 squares, i.e. $M = 50$. The square layer problem was solved using the lowest order conforming and mixed-hybrid finite element method. The resulting systems of linear equations were solved by the preconditioned conjugate gradient method.

The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table V.

Table V. Numbers of iterations for the square layer model problem.

	n	IC	MIC
conforming method	2601	52	22
mixed-hybrid method	5100	107	257



(a) conforming

(b) mixed

(c) conforming, enlargement

(d) mixed, enlargement

Figure 8. Streamlines and residence times for the square layer model problem.

Surprisingly, for the mixed-hybrid method the number of iterations when using the modified incomplete Cholesky decomposition is larger than when using the incomplete Cholesky decomposition (for a discussion on this phenomenon, see [34]).

We computed 24 streamlines starting on the bottom boundary and ending on the top boundary, such that the flow rate is equal between all adjacent streamlines. The corresponding residence times were also computed. The streamlines and residence times are displayed in figure 8. The results of the mixed finite element method are slightly superior, particularly near the tips of the clay layers.

The flat layer model problem

Let $\Omega = \{x = (x_1, x_2)^T \mid 0 < x_1 < 10,\ 0 < x_2 < 1\}$. Consider the flat layer model problem

$$\text{(10.18)} \quad \boxed{\begin{array}{l} -\nabla\cdot(a\nabla\phi) = 0 \ \text{ in } \Omega, \\ \phi = g_D \ \text{ on } \partial\Omega_D, \ \ -\boldsymbol{n}\cdot(a\nabla\phi) = 0 \ \text{ on } \partial\Omega_N, \end{array}}$$

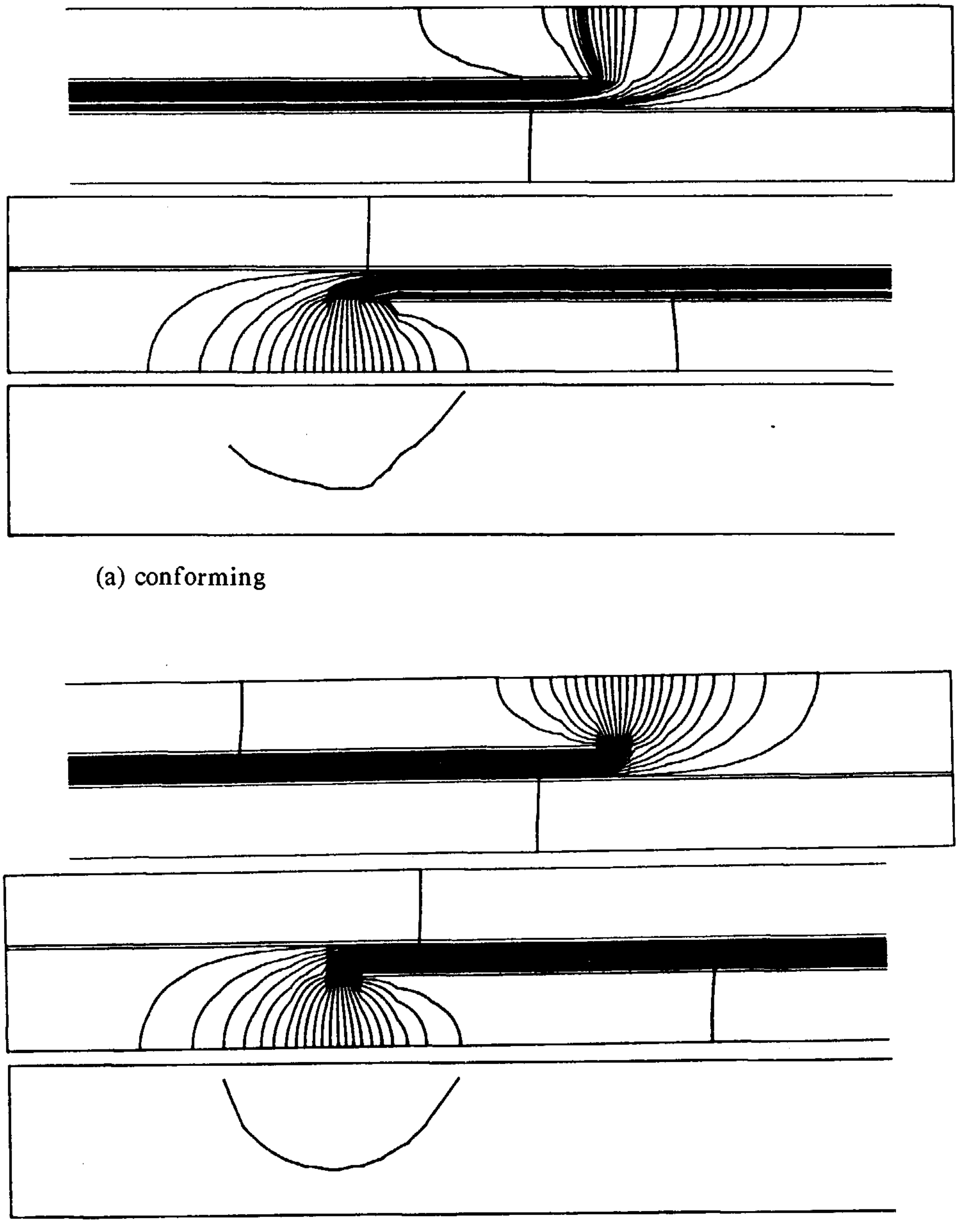
where

$$\text{(10.19)} \quad a(x) = \begin{cases} 10^{-5} & \text{if } x \in \Omega,\ x_1 < 8 \text{ and } 0.58 < x_2 < 0.60, \\ & \quad\ \text{or } x_1 > 2 \text{ and } 0.40 < x_2 < 0.42, \\ 1 & \text{if } x \text{ is elsewhere in } \Omega, \end{cases}$$

$$\text{(10.20)} \quad \partial\Omega_D = \{x \in \partial\Omega \mid x_2 = 0 \text{ or } x_2 = 1\},\ \partial\Omega_N = \partial\Omega\backslash\partial\Omega_D,$$

$$\text{(10.21)} \quad g_D(x) = 1 - x_2,\ x \in \partial\Omega.$$

A regular triangulation of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of 2500 rectangles. The flat layer model problem was solved using the lowest order conforming and mixed-hybrid finite element method. The resulting systems



(a) conforming

(b) mixed

Figure 9. Streamlines and residence times for the flat layer model problem.

of linear equations were solved by the preconditioned conjugate gradient method.

The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table VI.

Table VI. Numbers of iterations for the flat layer model problem
(† indicates that the preconditioning matrix does not exist).

	n	IC	MIC
conforming method	2601	17	7
mixed-hybrid method	5100	72	†

Although the flat layer model problem seems to be tougher than the square layer model problem, fewer iterations were required. For the mixed-hybrid method, a preconditioning matrix cannot be constructed by the modified incomplete Cholesky decomposition. If in definition (10.19) the number 10^{-5} is replaced by 10^{-3}, then for the mixed-hybrid method 66 iterations are required when using the incomplete Cholesky decomposition and 62 iterations when using the modified incomplete Cholesky decomposition.

We computed 24 streamlines starting on the bottom boundary and ending on the top boundary, such that the flow rate is equal between all adjacent streamlines. The corresponding residence times were also computed. The streamlines and residence times are displayed in figure 9. The results of the mixed finite element method are clearly superior. The symmetry of the problem is preserved very well by the mixed method.

The cubic layer model problem

Let $\Omega = \{x = (x_1,x_2,x_3)^T \,|\, 0 < x_i < 1,\ i = 1,2,3\}$. Consider the cubic layer model problem

$$(10.22)\quad \boxed{\begin{array}{l} -\nabla\cdot(a\nabla\phi) = 0 \ \text{ in } \Omega, \\ \phi = g_D \ \text{ on } \partial\Omega_D, \quad -\boldsymbol{n}\cdot(a\nabla\phi) = 0 \ \text{ on } \partial\Omega_N, \end{array}}$$

where

$$(10.23)\quad a(x) = \begin{cases} 10^{-5} & \text{if } x \in \Omega,\ x_1 < 0.8 \text{ or } x_2 < 0.8, \text{ and } 0.60 < x_3 < 0.64, \\ & \quad \text{or, } x_1 > 0.2 \text{ or } x_2 > 0.2, \text{ and } 0.36 < x_3 < 0.40, \\ 1 & \text{if } x \text{ is elsewhere in } \Omega, \end{cases}$$

$$(10.24)\quad \partial\Omega_D = \{x \in \partial\Omega \,|\, x_3 = 0 \text{ or } x_3 = 1\},\ \partial\Omega_N = \partial\Omega\backslash\partial\Omega_D,$$

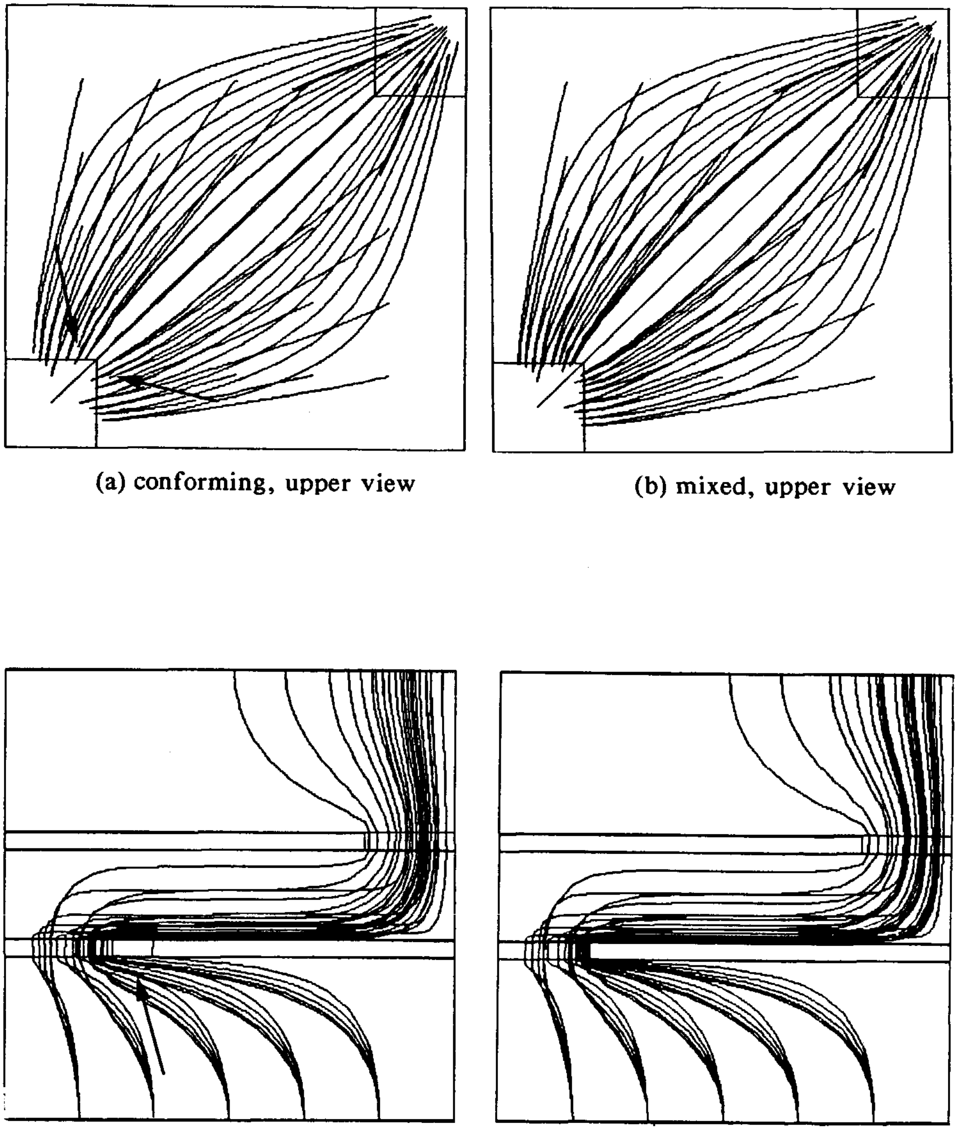
$$(10.25)\quad g_D(x) = 1 - x_3,\ x \in \partial\Omega.$$

One can interpret (10.22) as the description of a sandy porous medium with two clay layers in which there are two holes. A regular triangulation of $\overline{\Omega}$ was constructed by subdividing $\overline{\Omega}$ into a collection of 25^3 cubes. The vertices and faces of the mesh were numbered lexicographically. The cubic layer model problem was solved using the lowest order conforming and mixed-hybrid finite element method. The resulting systems of linear equations were solved by the preconditioned conjugate gradient method.

The numbers of iterations necessary to fulfil the desired termination criterion are displayed in table VII.

Table VII. Numbers of iterations for the cubic layer model problem.

	n	IC	MIC
conforming method	17576	40	39
mixed-hybrid method	48750	99	1501



(a) conforming, upper view

(b) mixed, upper view

(c) conforming, side view

(d) mixed, side view

Figure 10. Streamlines for the cubic layer model problem.

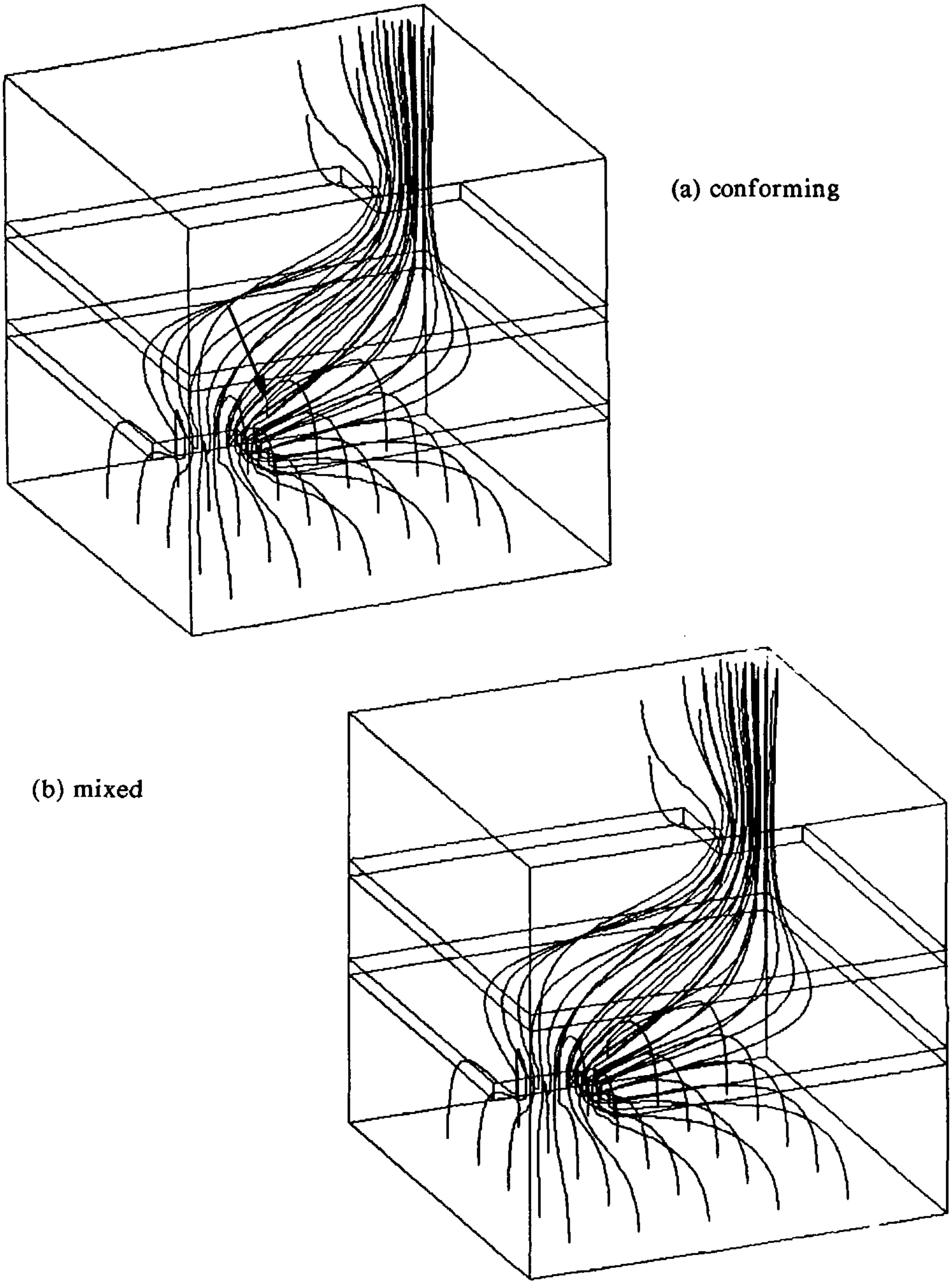


Figure 11. Streamlines for the cubic layer model problem.

For the conforming method, the incomplete Cholesky decomposition and the modified incomplete Cholesky decomposition are competitive. Surprisingly few iterations were required. For the mixed-hybrid method, very many iterations were required when using the modified incomplete Cholesky decomposition.

We computed 25 streamlines starting at the points $(i/6,j/6,0)^T$, $i,j = 1,\dots,5$, and ending on the top boundary. The streamlines are displayed in figures 10 and 11. The results of the mixed finite element method are superior. Using the conforming method, some streamlines intersect the lower clay layer (indicated by arrows in the figures), whereas using the mixed method the corresponding streamlines pass through the holes.

11. Conclusions.

From the numerical experiments it is clear that an accurate approximation of the specific discharge $\boldsymbol{u} = -A\nabla\phi$ can be determined by the mixed finite element method. The benefits of the mixed method are apparent for problems with rough tensors of hydraulic conductivity and especially if the domain is subdivided into very flat subdomains. In the flat layer model problem the subdomains can be transformed into squares, but then the transformed tensor A is not equal to a scalar function times the unit tensor, i.e. the porous medium of the transformed problem is anisotropic.

Of course, if one is interested in an accurate approximation of the potential ϕ, then the conforming finite element method is preferable.

Using the hybridization technique, the mixed finite element method results in a system of linear equations with a sparse and symmetric positive definite coefficient matrix. This system can be solved efficiently by the preconditioned conjugate gradient method, where the preconditioning matrix is constructed by the incomplete Cholesky decomposition or the modified incomplete Cholesky decomposition. For a smooth tensor A the modified incomplete Cholesky decomposition results in fewer iterations, but it appears to be very sensitive for rough tensors of hydraulic conductivity.

After an approximation of *u* has been computed by the mixed finite element method, streamlines and residence times can be determined efficiently and accurately using elementwise computations at the element level.

Acknowledgements.

We wish to thank Henk van der Vorst and Jos Jansen for critically reading the manuscript and earlier drafts.

REFERENCES

[1] Arnold, D.N. and F. Brezzi, *Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates*, Mathematical Modelling and Numerical Analysis 19 (1985), pp. 7-32.

[2] Axelsson, O., *A survey of preconditioned iterative methods for linear systems of algebraic equations*, BIT 25 (1985), pp. 166-187.

[3] Axelsson, O. and V.A. Barker, *Finite Element Solution of Boundary Value Problems*, Academic Press, New York (1984).

[4] Bear, J., *Hydraulics of Groundwater*, McGraw-Hill, New York (1979).

[5] Brezzi, F., *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Revue Française d'Automatique, Informatique et Recherche Opérationelle 8-R2 (1974), pp. 129-151.

[6] Carey, G.F. and J.T. Oden, *Finite Elements, Volume II: A Second Course*, Prentice-Hall, Englewood Cliffs (1983).

[7] Chavent, G., G. Cohen and J. Jaffré, *Discontinuous upwinding and mixed finite elements for two-phase flows in reservoir simulation*, Computer Methods in Applied Mechanics and Engineering 47 (1984), pp. 93-118.

[8] Chavent, G. and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation*, North-Holland, Amsterdam (1986).

[9] Ciarlet, P.G., *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam (1978).

[10] Concus, P., G.H. Golub and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM Journal of Scientific and Statistical Computation 6 (1985), pp. 220-252.

[11] Ewing, R.E., J.V. Koebbe, R. Gonzalez and M.F. Wheeler, *Mixed finite element methods for accurate fluid velocities*, in: R.H. Gallagher, G.F. Carey, J.T. Oden and O.C. Zienkiewicz (eds.), *Finite Elements in Fluids, Volume 6*, Wiley, New York (1985), pp. 233-249.

[12] Golub, G.H. and C.F. van Loan, *Matrix Computations*, North Oxford Academic, Oxford (1983).

[13] Gustafsson, I., *Modified incomplete Cholesky (MIC) methods*, in: D.J. Evans (ed.), *Preconditioning Methods, Theory and Applications*, Gordon and Breach, New York (1983), pp. 265-293.

[14] Gustafsson, I. and G. Lindskog, *A preconditioning technique based on element matrix factorizations*, Computer Methods in Applied Mechanics and Engineering 55 (1986), pp. 201-220.

[15] Hackbush, W., *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner, Stuttgart (1986).

[16] Hecht, F., *Construction d'une base de fonctions P_1 non conforme à divergence nulle dans* $\mathbb{R}^3$, RAIRO Analyse Numérique 15 (1981), pp. 119-150.

[17] Hughes, T.J.R., I. Levit and J. Winget, *An element-by-element solution algorithm for problems of structural and solid mechanics*, Computer Methods in Applied Mechanics and Engineering 36 (1983), pp. 241-254.

[18] Huijben, A.J.M., *Een Hybride Gemengde Eindige-Elementenmethode met Postprocessing, Toegepast op het Hall-Probleem*, Masters Thesis, Eindhoven University of Technology (1987).

[19] Jaffré, J., *Mixed finite elements for the water flooding problems*, in: E. Hinton, P. Bettess and R.W. Lewis (eds.), *Numerical Methods for Coupled Problems*, Pineridge, Swansea (1981), pp. 968-976.

[20] Kaasschieter, E.F., *A practical termination criterion for the conjugate gradient method*, BIT 28 (1988), pp. 308-322.

[21] Kaasschieter, E.F., *A general finite element preconditioning for the conjugate gradient method*, to appear in BIT.

[22] Meijerink, J.A. and H.A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Mathematics of Computation 31 (1977), pp. 148-162.

[23] Meijerink, J.A. and H.A. van der Vorst, *Guidelines for the usage of incomplete decompositions in solving sets of linear systems as they occur in practical problems*, Journal of Computational Physics 44 (1981), pp. 131-155.

[24] Morse, P.M. and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, New York (1953).

[25] Muskat, M., *Physical Principles of Oil Production*, McGraw-Hill, New York (1949).

[26] Nedelec, J.C., *Mixed finite elements in* $\mathbb{R}^3$, Numerische Mathematik 35 (1980), pp. 315-341.

[27] Nedelec, J.C., *A new family of mixed finite elements in* $\mathbb{R}^3$, Numerische Mathematik 50 (1986), pp. 57-81.

[28] Philip, J.R., *Issues in flow and transport in heterogeneous porous media*, Transport in Porous Media 1 (1986), pp. 319-338.

[29] Raviart, P.-A. and J.-M. Thomas, *A mixed finite element method for 2-nd order elliptic problems*, in: I. Galligani and E. Magenes (eds.), *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Mathematics 606, Springer, Berlin (1977), pp. 292-315.

[30] Roberts, J.E. and J.-M. Thomas, *Mixed and hybrid finite element methods*, to appear in P.G. Ciarlet and J.-L. Lions (eds.), *Handbook of Numerical Analysis, Volume II: Finite Element Methods*, North-Holland, Amsterdam.

[31] Schwarz, H.R., *The Finite Element Method*, Academic Press, New York (1988).

[32] Thomas, J.-M., *Sur l'Analyse Numérique des Méthodes d'Eléments Finis Hybrides et Mixtes*, Ph.D. Thesis, University Pierre et Marie Curie, Paris (1977).

[33] Tóth, J., *A theoretical analysis of groundwater flow in small drainage basins*, Journal of Geophysical Research 68 (1963), pp. 4795-4812.

[34] van der Vorst, H.A., *The convergence behaviour of preconditioned CG and CG-S*, to appear in the *Proceedings of the Conference on Preconditioned Conjugate Gradient Methods*, Nijmegen, June 19-21, 1989.

[35] Weiser, A. and M.F. Wheeler, *On convergence of block-centered finite differences for elliptic problems*, SIAM Journal on Numerical Analysis 25 (1988), pp. 351-375.

SAMENVATTING

Dit proefschrift gaat over problemen, die zich kunnen voordoen, wanneer potentiaalstromingsproblemen numeriek worden opgelost.

Potentiaalstromingsproblemen zijn fundamenteel in verschillende gebieden van de mathematische fysica, zoals de warmtegeleiding en de electrostatica. De voornaamste inspiratie van dit proefschrift is echter de hydraulica van grondwater.

Alleen in bijzondere gevallen kan een exacte oplossing van een potentiaalstromingsprobleem bepaald worden. Daarom zijn numerieke methoden het belangrijkste middel ter oplossing van dergelijke problemen.

De eindige elementenmethode is zeer geschikt ter bepaling van een benadering van de oplossing van een potentiaalstromingsprobleem. Volgens deze methode wordt het stromingsdomein onderverdeeld in een aantal geometrisch eenvoudige deelgebieden, die eindige elementen worden genoemd. In ieder deelgebied wordt de oplossing benaderd door een polynomiale functie. Deze stuksgewijs polynomiale benadering moet aan zekere continuïteitseisen langs de interelementranden voldoen.

De conforme eindige elementenmethode bepaalt een stuksgewijs polynomiale benadering van de potentiaal, die continu dient te zijn langs de interelementranden. Uiteindelijk wordt een stelsel lineaire vergelijkingen met een symmetrisch positief definiete matrix verkregen.

Deze matrix is in het algemeen groot, ijl en slecht geconditioneerd. Iteratieve methoden zijn noodzakelijk ter oplossing van een dergelijk stelsel lineaire vergelijkingen. Als de matrix symmetrisch positief definiet is, dan is de gepreconditioneerde geconjugeerde gradiëntenmethode een uitstekende keuze.

Hoewel de conforme eindige elementenmethode zeer geschikt is ter bepaling van een nauwkeurige benadering van de potentiaal, is dit niet het geval voor de volumestroomdichtheid. Een nauwkeurige benadering van de volumestroomdichtheid kan worden verkregen met de gemengde eindige elementenmethode. Deze methode bepaalt stuksgewijs polynomiale benaderingen van de volumestroomdichtheid en de potentiaal, waarbij de normale component van de eerste benadering continu dient te zijn door de interelementranden.

Wederom wordt een stelsel lineaire vergelijkingen verkregen. De keuze van

een numerieke methode ter oplossing van dit stelsel is beperkt, omdat de matrix niet definiet is. Dit bezwaar kan omzeild worden door een implementatietechniek, die hybridisatie wordt genoemd. Aldus wordt een symmetrisch positief definiet stelsel lineaire vergelijkingen verkregen. Aangezien dit stelsel ijl is, kan het met behulp van de gepreconditioneerde geconjugeerde gradiëntenmethode efficiënt worden opgelost.

In dit proefschrift worden verschillende aspecten van eindige elementenmethoden en gepreconditioneerde geconjugeerde gradiëntenmethoden behandeld.

Hoofdstuk I gaat over de geconjugeerde gradiëntenmethode voor de iteratieve oplossing van een stelsel vergelijkingen $Ax = b$. Er wordt getoond, hoe de kleinste actieve eigenwaarde van A goedkoop benaderd kan worden, en de bruikbaarheid van deze benadering voor een praktisch stopcriterium voor de geconjugeerde gradiëntenmethode wordt bestudeerd. Bewezen wordt, dat dit stopcriterium betrouwbaar is in vele relevante situaties.

In hoofdstuk II wordt de gepreconditioneerde geconjugeerde gradiëntenmethode gebruikt ter oplossing van het stelsel vergelijkingen $Ax = b$, waarbij A een singuliere matrix is. De methode divergeert, als b niet precies in het bereik van A ligt. Als de nulruimte van A expliciet bekend is, dan kan deze divergentie vermeden worden door van b de orthogonale projectie op de nulruimte af te trekken. Naast de analyse van deze aftrekking worden voldoende voorwaarden voor de existentie van de incomplete Cholesky decompositie gegeven. Tenslotte wordt de theorie toegepast op het gediscretiseerde potentiaalstromingsprobleem met Neumann randvoorwaarden.

De discretisatie van een symmetrisch elliptisch randwaardeprobleem door middel van de eindige elementenmethode leidt tot een stelsel lineaire vergelijkingen met een symmetrisch positief definiete matrix. In hoofdstuk III wordt een preconditioneringsmatrix voorgesteld, die voor alle eindige elementenmethoden geconstrueerd kan worden, mits aan een milde eis voor de knooppuntsnummering is voldaan. Zo'n nummering kan gecreëerd worden door middel van een variant van het Cuthill-McKee algoritme.

In hoofdstuk IV wordt de laagste orde gemengd-hybride eindige elementenmethode besproken voor algemene potentiaalstromingsproblemen. De elementsgewijze berekening van stroomlijnen en verblijftijden wordt uiteengezet.

CURRICULUM VITAE

De schrijver van dit proefschrift werd op 27 augustus 1960 te Buenos Aires geboren. Nadat hij in juni 1978 het eindexamen Atheneum aan de Rembrandt Scholengemeenschap te Leiden had afgelegd, begon hij in hetzelfde jaar met de studie wiskunde aan de Rijksuniversiteit te Leiden. In januari 1981 werd het kandidaatsexamen wiskunde (toegepaste richting) met bijvak inleiding natuur- en sterrenkunde afgelegd, in september van hetzelfde jaar gevolgd door het kandidaatsexamen wiskunde (toegepaste richting) met tweede hoofdvak natuurkunde. Op 21 september 1984 slaagde hij voor het doctoraalexamen wiskunde met bijvakken theoretische natuurkunde en geschiedenis en maatschappelijke functie van de wiskunde. Het afstuderen bestond uit de actieve deelname aan een seminarium numerieke wiskunde onder leiding van Prof.dr. M.N. Spijker en Dr. J.A. van de Griend. Tevens werden colleges in de wijsbegeerte der natuurwetenschappen, in de wetenschap, techniek, beleid en maatschappij en in de muziekgeschiedenis gevolgd.

Van 1 oktober 1984 tot 19 april 1986 vervulde de auteur de vervangende dienstplicht bij de subgroep Numerieke Analyse van de vakgroep Toegepaste Analyse aan de Technische Hogeschool te Delft. Hij kwam daar in aanraking met de gepreconditioneerde geconjugeerde gradiëntenmethode en de eindige elementenmethode. Sinds 19 april 1986 is de auteur verbonden aan de Dienst Grondwaterverkenning TNO te Delft. Aldaar heeft hij onder supervisie van Prof.dr. H.A. van der Vorst en Dr.ir. W. Zijl het onderzoek verricht, dat heeft geleid tot de in dit proefschrift beschreven resultaten.