

J.M. Burgerscentrum

Research School for Fluid Mechanics

Finite element methods for  
the incompressible Navier-Stokes equations

Ir. A. Segal

2011



Delft University of Technology  
Faculty of Electrical Engineering, Mathematics and Computer Science  
Delft Institute of Applied Mathematics

Copyright © 2011 by Delft Institute of Applied Mathematics, Delft, The Netherlands.

No part of this work may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission from Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Introduction to the finite element method</b>	<b>8</b>
2.1	Differential equations and boundary conditions . . . . .	8
2.2	Weak formulation . . . . .	9
2.3	The Galerkin method . . . . .	11
2.4	The finite element method . . . . .	12
2.5	Computation of the element matrix and element vector . . . . .	15
2.6	Higher order elements . . . . .	17
2.7	Structure of the large matrix . . . . .	18
<b>3</b>	<b>Convection-diffusion equation by the finite element method</b>	<b>22</b>
3.1	Formulation of the equations . . . . .	22
3.2	Standard Galerkin . . . . .	22
3.3	Solution of the system of ordinary differential equations . . . . .	24
3.4	Accuracy aspects of the SGA . . . . .	27
3.5	Streamline Upwind Petrov Galerkin . . . . .	30
3.6	Some classical benchmark problems for convection-diffusion solvers . . . . .	35
<b>4</b>	<b>Discretization of the incompressible Navier-Stokes equations by standard Galerkin</b>	<b>39</b>
4.1	The basic equations of fluid dynamics . . . . .	39
4.2	Initial and boundary conditions . . . . .	40
4.3	Axisymmetric flow . . . . .	42
4.4	The weak formulation . . . . .	43
4.5	The standard Galerkin method . . . . .	45
4.6	Treatment of the non-linear terms . . . . .	46
4.7	Necessary conditions for the elements . . . . .	48
4.8	Examples of admissible elements . . . . .	50
4.9	Solution of the system of linear equations due to the discretization of Navier-Stokes . . . . .	53
<b>5</b>	<b>The penalty function method</b>	<b>56</b>
5.1	Introduction . . . . .	56
5.2	The discrete penalty functions approach . . . . .	57
5.3	The continuous penalty function method . . . . .	58
5.4	Practical aspects of the penalty function method . . . . .	60
<b>6</b>	<b>Divergence-free elements</b>	<b>62</b>
6.1	Introduction . . . . .	62
6.2	The construction of divergence-free basis functions for 2D elements . . . . .	63
6.3	The construction of element matrices and vectors for (approximate) divergence-free basis functions . . . . .	67
6.4	Boundary conditions with respect to the divergence-free elements . . . . .	69
6.5	Computation of the pressure . . . . .	70
6.6	Practical aspects of the divergence-free elements . . . . .	71

<b>7</b>	<b>The instationary Navier-Stokes equations</b>	<b>72</b>
7.1	Introduction . . . . .	72
7.2	Solution of the instationary Navier-Stokes equations by the method of lines .	72
7.3	The pressure-correction method . . . . .	74
<b>A</b>	<b>Derivation of the integration by parts for the momentum equations</b>	<b>78</b>

# 1 Introduction

In this second course on numerical flow problems we shall focus our attention to two specific subjects:

- the solution of the incompressible Navier-Stokes equations by finite elements;
- the efficient solution of large systems of linear equations.

Although, at first sight there seems no connection between both subjects, it must be remarked that the numerical solution of partial differential equations and hence also of the Navier-Stokes equations, always results in the solving of a number of large systems of sparse linear systems. Since, in general the solution of these systems and the storage of the corresponding matrices take the main part of the resources (CPU time and memory), it is very natural to study efficient solution methods.

In the first part of these lectures we shall be concerned with the discretization of the incompressible Navier-Stokes equations by the finite element method (FEM). First we shall give a short introduction of the FEM itself. The application of the FEM to potential problems is considered and the extension to convection-diffusion type problems is studied. The Galerkin method is introduced as a natural extension of the so-called weak formulation of the partial differential equations. One of the reasons why finite elements have been less popular in the past than finite differences, is the lack of upwind techniques. In the last decade, however, accurate upwind methods have been constructed. The most popular one, the so-called streamline upwind Petrov-Galerkin method (SUPG), will be treated in Chapter 3. It is shown that upwinding may increase the quality of the solution considerably. Another important aspect of upwinding is that it makes the systems of equations more appropriate for the iterative methods treated in part II. As a consequence both the number of iterations and the computation time decrease.

In Chapter 4 the discretization of the incompressible Navier-Stokes equations is considered. Since the pressure is an unknown in the momentum equations but not in the continuity equation, the discretization must satisfy some special requirements. In fact one is no longer free to choose any combination of pressure and velocity approximation but the finite elements must be constructed such that the so-called Brezzi-Babuška (or BB) condition is satisfied. This condition makes a relation between pressure and velocity approximation. In finite differences and finite volumes the equivalent of the BB condition is satisfied if staggered grids are applied. Even if the BB condition is satisfied we are still faced with a problem with respect to the solution of the linear systems of equations. The absence of the pressure in the continuity equation induces zeros at some of the diagonal elements of the matrix. In general linear solvers may be influenced by such zeros, some iterative solvers even do not allow non-positive diagonal elements. For that reason alternative solution methods have been developed, which all try to segregate the pressure and velocity computation.

Chapter 5 treats the most popular segregated method, the so-called penalty function formulation. In this approach the continuity equation is perturbed with a small compressibility including the pressure.

From this perturbed equation the pressure is expressed in terms of velocity and this is substituted into the momentum equations. In this way the velocity can be computed first and afterwards the pressure. A disadvantage of this method is that the perturbation parameter

introduces extra complications.

In Chapter 6 an alternative formulation is derived, the so-called solenoidal approach. In this method, the elements are constructed in such a way that the approximate divergence freedom is satisfied explicitly. To that end it is necessary to introduce the stream function as help unknown. This method seems very attractive, however, the extension to three-dimensional problems is very difficult.

Finally, in Chapter 7, methods for the time-dependent incompressible Navier-Stokes equations are treated. For this type of methods an alternative segregation is possible, the so-called pressure-correction method. This method is also popular in finite differences and finite volumes.

In the second part we consider the efficient solution of large system of linear equations. As applications and examples we mostly use systems of equations resulting from the discretization of 2- and 3 dimensional partial differential equations.

We start in Chapter 2.1 by considering direct methods as there are: Gaussian elimination and Cholesky decomposition. These methods are used if the dimension of the system is not too large. Since these methods are implemented on computers we consider the behavior of the methods with respect to rounding errors. For large problems the memory requirements of direct methods are a bottle neck, using special methods for banded, profile or "general sparse" matrices we can solve much larger systems. These methods only use the non zero part of the resulting decomposition.

In Chapter 2 we consider classical iterative methods for linear equations. These methods are very cheap with respect to memory requirements. However, convergence can be very slow so the computing time may be much larger than for direct methods.

In Chapter 3 and 4 we consider modern iterative methods of Krylov subspace type. In Chapter 3 the conjugate gradient method for symmetric positive definite matrices, and in Chapter 4 Krylov methods for general matrices. The rate of convergence is much better than for basic iterative methods, whereas no knowledge of the spectrum is needed in contrast with some basic iterative methods. A drawback for general matrices is that there are many methods proposed and until now there is no clear winner. We shall summarize the most successful ones and try to give some guidelines for choosing a method depending on the properties of the problem.

The Krylov subspace methods become much faster if they are combined with a preconditioner. For the details we refer to Chapter 5. We only note that in essence a preconditioned Krylov method is a combination of a method given in chapter 3 and 4 and a basic iterative method or an incomplete direct method.

In many applications eigenvalues give information of physical properties (like eigenmodes) or they are used to analyze, and enhance mathematical methods for solving a physical problem. If only a small number of eigenvalues are needed for a very large matrix it is a good idea to use iterative methods, which are given in Chapter 6. We start with the Power method which is easy to understand, and approximate the largest eigenvalue. Thereafter we consider the Lanczos method for symmetric matrices, which is closely related to the CG method. Again for general matrices different methods are proposed and it is not always clear, which one is

the best.

Finally in Chapter 7 we give a summary of present day supercomputers. There are mainly two types: vector- and parallel computers. For the problems considered in this report supercomputers are necessary to obtain results for large 3 dimensional problems. At this moment vector computers give the best results, with respect to computing time and memory. However, we expect that in the near future parallel computers (especially those based on a clustering of very fast nodes) will beat them for real live problems.

## 2 Introduction to the finite element method

### 2.1 Differential equations and boundary conditions

The finite element method (FEM) may be considered as a general discretization tool for partial differential equations. In this sense the FEM forms an alternative for finite difference methods (FDM) or finite volume methods (FVM). The main reason to use the FEM is its ability to tackle relatively easily, problems that are defined on complex geometries. However, the programming of finite element methods is more complicated than that of finite differences, and hence in general requires standard software packages.

In this lecture we shall restrict ourselves to the application of the FEM to two general types of differential equations: the convection-diffusion equation and the incompressible Navier-Stokes equations. The last type of equations are the subject of Chapter 4, and convection-diffusion type of equations will be the subject of Chapter 3. In this introductory chapter we shall neglect the convective terms and focus ourselves to diffusion type problems:

$$\rho \frac{\partial c}{\partial t} - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left( a_{ij} \frac{\partial c}{\partial x_j} \right) + \beta c = f \quad (2.1)$$

where  $c$  denotes the unknown, for example the potential, temperature or concentration. The matrix  $A$  with elements  $a_{ij}$  represents the diffusion tensor and is supposed to be symmetric and positive definite. The coefficient  $\beta$  is zero in many practical problems, but is added for the sake of generality.  $f$  represents a source term and  $\rho \frac{\partial c}{\partial t}$  the time-derivative part, where  $\rho$  must be positive in the instationary case. All coefficients  $\rho$ ,  $a_{ij}$ ,  $\beta$  and  $f$  may depend on time and the space variable  $\mathbf{x}$ .  $n$  is the dimension of space which in our applications varies from 1 to 3.

If the coefficients also depend on the solution, the equations become non-linear. In this chapter we restrict ourselves to linear problems only.

Equation 2.1 is usually written in vector notation:

$$\rho \frac{\partial c}{\partial t} - \operatorname{div}(A \nabla c) + \beta c = f, \quad (2.2)$$

where  $\nabla$  denotes the gradient operator.

In this chapter we shall only consider stationary problems, so equation (2.2) reduces to

$$-\operatorname{div}(A \nabla c) + \beta c = f \quad (2.3)$$

In order that equation (2.3) has a unique solution, and to make the problem well posed it is necessary to prescribe exactly one boundary condition at each part of the boundary. In the sequel the region at which the differential equation is defined is called  $\Omega$  and its boundary is denoted by  $\Gamma$ . Common boundary conditions in equation (2.3) are:

$$c = g_1 \quad \text{on } \Gamma_1, \quad (2.4)$$

$$A \nabla c \cdot \mathbf{n} = g_2 \quad \text{on } \Gamma_2, \quad (2.5)$$

$$\sigma c + A \nabla c \cdot \mathbf{n} = g_3 \quad \text{on } \Gamma_3 \quad (\sigma \geq 0), \quad (2.6)$$

where the boundary  $\Gamma$  is subdivided into three parts  $\Gamma_1$ ,  $\Gamma_2$  and  $\Gamma_3$ .

Boundary conditions of type (2.4) are called Dirichlet boundary conditions, boundary conditions of type (2.5) Neumann conditions and boundary conditions of type (2.6) are called Robbins boundary conditions. In fact (2.5) may be considered as a special case of (2.6).

Other types of boundary conditions may also be applied, but they will not be studied in this lecture.

It can be shown that equation (2.3) with boundary conditions (2.4) to (2.6) has a unique solution provided the coefficients and the boundary of the region are sufficiently smooth. Only in the special case  $\Gamma_1 = \phi$ ,  $\Gamma_3 = \phi$  and  $\beta = 0$ , the function  $\varphi$  is determined up to an additive constant. In that case the functions  $f$  and  $g_2$  must satisfy the compatibility condition

$$\int_{\Omega} f \, d\Omega = - \int_{\Gamma} g_2 \, d\Gamma . \quad (2.7)$$

(2.7) can be derived by integrating equation (2.3) over the domain  $\Omega$  and applying the Gauss divergence theorem.

## 2.2 Weak formulation

Before applying the FEM to solve equation (2.3) under the boundary conditions (2.4) to (2.6), it is necessary to transform the equation into a more suitable form. To do that there are two alternatives:

1. one can derive an equivalent minimization problem, which has exactly the same solution as the differential equation.
2. one can derive a so-called weak formulation.

Both methods lead finally to exactly the same result, however, since for the general equations to be treated in Chapters 2 and 4, no equivalent minimization problem exists, we shall restrict ourselves to method 2.

Originally the weak formulation has been introduced by pure mathematicians to investigate the behavior of the solution of partial differential equations, and to prove existence and uniqueness of the solution. Later on numerical schemes have been based on this formulation which lead to an approximate solution in a constructive way.

The weak formulation of equation (2.3) can be derived by multiplying (2.3) by a so-called test function  $v$  and integrating over the domain. So:

$$\int_{\Omega} (-\operatorname{div}(A\nabla c) + \beta c) v \, d\Omega = \int_{\Omega} f v \, d\Omega . \quad (2.8)$$

The choice of the class of functions to which  $v$  belongs, determines whether (2.8) has a solution and whether this solution is unique.

It is common practice to apply integration by parts to equation (2.8) in order to get rid of the second derivative term. integration by parts is derived by applying the Gauss divergence theorem:

$$\int_{\Omega} \operatorname{div} \mathbf{a} \, d\Omega = \int_{\Gamma} \mathbf{a} \cdot \mathbf{n} \, d\Gamma \quad (2.9)$$

to the function

$$\mathbf{a} = v A \nabla c . \quad (2.10)$$

Hence

$$\operatorname{div} \mathbf{a} = \operatorname{div}(v A \nabla c) = \nabla v \cdot A \nabla c + v \operatorname{div} A \nabla c . \quad (2.11)$$

Substitution of (2.11) in (2.9) yields

$$\int_{\Omega} v \operatorname{div} A \nabla c \, d\Omega = - \int_{\Omega} \nabla v \cdot A \nabla c \, d\Omega + \int_{\Gamma} v A \nabla c \cdot \mathbf{n} \, d\Gamma \quad (2.12)$$

and so (2.8) can be written as

$$\int_{\Omega} \{A \nabla c \cdot \nabla v + \beta c v\} \, d\Omega - \int_{\Gamma} v A \nabla c \cdot \mathbf{n} \, d\Gamma = \int_{\Omega} f v \, d\Omega . \quad (2.13)$$

The boundary conditions (2.4) to (2.6) are applied by evaluating the boundary integral at (2.13) if possible. This boundary integral can be subdivided into three parts:

$$\int_{\Gamma} v A \nabla c \cdot \mathbf{n} \, d\Gamma = \int_{\Gamma_1} v A \nabla c \cdot \mathbf{n} \, d\Gamma + \int_{\Gamma_2} v A \nabla c \cdot \mathbf{n} \, d\Gamma + \int_{\Gamma_3} v A \nabla c \cdot \mathbf{n} \, d\Gamma . \quad (2.14)$$

On boundary  $\Gamma_1$  we have the boundary condition  $c = g_1$ . Since this boundary condition can not be incorporated explicitly in (2.14) we demand that the function  $c$  in (2.13) satisfies (2.4) and furthermore in order to get rid of the boundary integral over  $\Gamma_1$ :

$$v = 0 \text{ at } \Gamma_1 . \quad (2.15)$$

On boundary  $\Gamma_2$  we can substitute the boundary condition (2.5) and the same is true for boundary  $\Gamma_3$ . For that reason we do not demand anything for the solution  $c$  or the test function  $v$  at these boundaries.

The boundary condition (2.4) is called essential, since it should be satisfied explicitly. The boundary conditions (2.5), (2.6) are called natural, since they are implicitly satisfied by the formulation. These terms are in first instance motivated by the corresponding minimization problem.

The weak formulation corresponding to equation (2.3) under the boundary conditions (2.4) to (2.6) now becomes:

find  $c$  with  $c|_{\Gamma_1} = g_1$  such that

$$\int_{\Omega} \{A \nabla c \cdot \nabla v + \beta c v\} \, d\Omega + \int_{\Gamma_3} \sigma c v \, d\Gamma = \int_{\Omega} f v \, d\Omega + \int_{\Gamma_2} g_2 v \, d\Gamma + \int_{\Gamma_3} g_3 v \, d\Gamma , \quad (2.16)$$

for all functions  $v$  satisfying  $v|_{\Gamma_1} = 0$ .

Furthermore it is necessary to demand some smoothness requirements for the functions  $c$  and  $v$ . One can prove that it is sufficient to require that all integrals in (2.16) exist, which means that both  $\nabla c$  and  $\nabla v$  must be square integrable.

We see in this expression that an essential boundary condition automatically implies that the

corresponding test function is equal to zero, whereas the natural boundary conditions do not impose any condition either to the unknown or to the test function. It is not immediately clear whether a boundary condition is essential or natural, except in the case where we have a corresponding minimization problem. In general, however, one can say that for second order differential equations, all boundary conditions containing first derivatives are natural, and a given function at the boundary is essential.

In fourth order problems the situation is more complex. However, for physical problems, in general, one can state that if the boundary conditions contain second or third derivatives they are natural, whereas boundary conditions containing only the function or first order derivatives are essential. The easiest way to check whether a boundary condition is essential or natural is to consider the boundary integrals. If in some way the boundary condition can be substituted, the boundary condition is natural. Otherwise the condition is essential and the test functions must be chosen such that the boundary integral vanishes.

### 2.3 The Galerkin method

Formulation (2.16) is one of the various possible weak formulations. However, it is the most common one and also the most suitable for our purpose. In the FEM we use formulation (2.16) instead of (2.3)-(2.6) to derive the discretization. Starting point is the so-called Galerkin method. In this method the solution  $c$  is approximated by a linear combination of expansion functions the so-called basis functions:

$$c^n(\mathbf{x}) = \sum_{j=1}^n c_j \varphi_j(\mathbf{x}) + c_0(\mathbf{x}) \quad (2.17)$$

where the parameters  $c_j$  are to be determined. The basis functions  $\varphi_j(\mathbf{x})$  must be linearly independent.

Furthermore they must be such that an arbitrary function in the solution space can be approximated with arbitrary accuracy, provided a sufficient number of basis functions is used in the linear combination (2.17). The function  $c_0(\mathbf{x})$  must be chosen such that  $c^n(\mathbf{x})$  satisfies the essential boundary conditions. In general this means that

$$c_0(\mathbf{x}) = g \quad \text{at } \Gamma_1 \quad (2.18)$$

$$\varphi_j(\mathbf{x}) = 0, \quad \text{at } \Gamma_1 \quad (2.19)$$

In order to determine the parameters  $c_j (j = 1, 2, \dots, n)$  the test functions  $v$  are chosen in the space spanned by the basis functions  $\varphi_1(\mathbf{x})$  to  $\varphi_n(\mathbf{x})$ .

It is sufficient to substitute

$$v(\mathbf{x}) = \varphi_i(\mathbf{x}) \quad i = 1(1)n \quad (2.20)$$

into equations (2.16). This leads to a linear system of  $n$  equations with  $n$  unknowns. The choice (2.20) implies immediately that  $v$  satisfies the essential boundary conditions for  $v$ .

After substitution of (2.17) and (2.20) into (2.16) we get the so-called Galerkin formulation

$$\begin{aligned} & \int_{\Omega} \{A \nabla c^n \cdot \nabla \varphi_i + \beta c^n \varphi_i\} d\Omega + \int_{\Gamma_3} \sigma c^n \varphi_i d\Gamma \\ &= \int_{\Omega} f \varphi_i d\Omega + \int_{\Gamma_2} g_2 \varphi_i d\Gamma + \int_{\Gamma_3} g_3 \varphi_i d\Gamma, \end{aligned} \quad (2.21)$$

with  $c^n = \sum_{j=1}^n c_j \varphi_j + c_0$ .

Hence

$$\sum_{j=1}^n c_j \left\{ \int_{\Omega} \{A \nabla \varphi_j \cdot \nabla \varphi_i + \beta \varphi_i \varphi_j\} d\Omega + \int_{\Gamma_3} \sigma \varphi_j \varphi_i d\Gamma \right\} = \int_{\Omega} f \varphi_i d\Omega + \int_{\Gamma_2} g_2 \varphi_i d\Gamma + \int_{\Gamma_3} g_3 \varphi_i d\Gamma - \int_{\Omega} \{A \nabla c_0 \cdot \nabla \varphi_i + \beta c_0 \varphi_i\} d\Omega. \quad (2.22)$$

Clearly (2.22) is a system of  $n$  linear equations with  $n$  unknowns, which can be written in matrix-vector notation as

$$\mathbf{S} \mathbf{c} = \mathbf{F}, \quad (2.23)$$

with

$$s_{ij} = \int_{\Omega} \{A \nabla \varphi_j \cdot \nabla \varphi_i + \beta \varphi_i \varphi_j\} d\Omega + \int_{\Gamma_3} \sigma \varphi_j \varphi_i d\Gamma, \quad (2.24a)$$

$$F_i = \int_{\Omega} f \varphi_i d\Omega + \int_{\Gamma_2} g_2 \varphi_i d\Gamma + \int_{\Gamma_3} g_3 \varphi_i d\Gamma - \int_{\Omega} \{A \nabla c_0 \cdot \nabla \varphi_i + \beta c_0 \varphi_i\} d\Omega. \quad (2.24b)$$

## 2.4 The finite element method

The FEM offers us a constructive way to create the basis functions  $\varphi_i$  and to compute the integrals in (2.24a-2.24b) in a relatively simple way. To that end the region  $\Omega$  is subdivided into simple elements. In  $\mathbb{R}^1$  these elements are intervals, in  $\mathbb{R}^2$  usually triangles or quadrilaterals and in  $\mathbb{R}^3$  tetrahedra and hexahedra are very popular. The subdivision of a region in elements is performed by a so-called mesh generator.

In each element a number of nodal points are chosen and the unknown function is approximated by a polynomial. Although other types of approximations are permitted it is common practice to restrict one selves to lower degree polynomials (linear or quadratic). These polynomial approximations implicitly define the basis functions  $\varphi_i$ .

For example a piecewise linear polynomial in  $\mathbb{R}^1$  defined on  $n$  elements  $e_i(x_{i-1}, x_i)$  (see Figure 2.1),

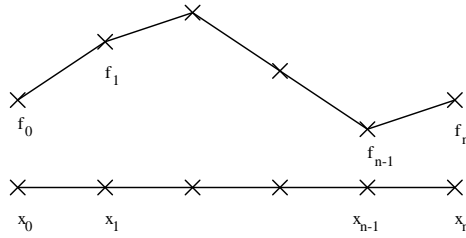


Figure 2.1: Piecewise linear approximation of function  $f(x)$

can be written as

$$f^n(x) = \sum_{j=0}^n f(x_j) \lambda_j(x), \quad (2.25)$$

where  $\lambda_j(x)$  is defined as follows:

$$\begin{cases} \lambda_j(x) \text{ is linear on each element,} \\ \lambda_j(x_k) = \delta_{jk} . \end{cases} \quad (2.26)$$

So in terms of the Galerkin method the function values  $f(x_j)$  take the role of the parameters and the shape functions  $\lambda_j(x)$  the role of the basis functions. Figure 2.2 shows a typical linear basis function  $\varphi_i(x)$ .

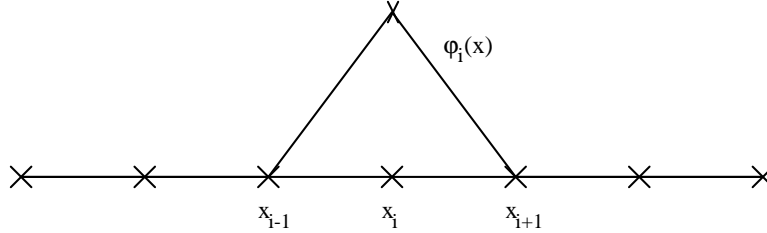


Figure 2.2: Example of a linear basis function

In the case of so-called quadratic elements in  $\mathbb{R}^1$  we define the vertices as well as the centroid as nodal points and the basis functions  $\varphi_i(x)$  are defined by

$$\begin{cases} \varphi_i(x) \text{ is quadratic on each element,} \\ \varphi_i(x_j) = \delta_{ij} \end{cases} . \quad (2.27)$$

Figure 2.3 shows the two types of quadratic basis functions we may expect in  $\mathbb{R}^1$ .

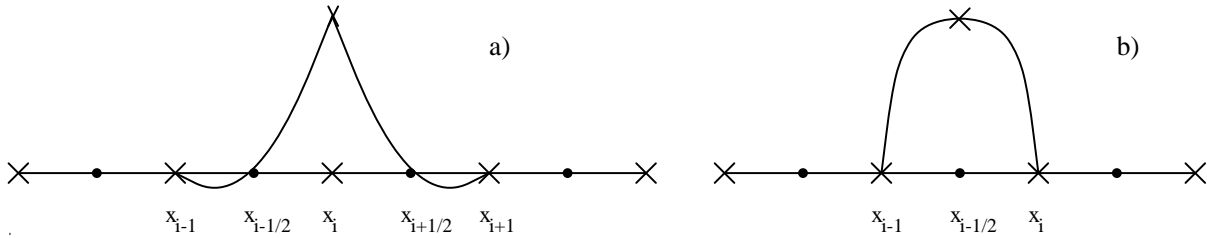


Figure 2.3: Quadratic basis functions in  $\mathbb{R}^1$  a) corresponding to vertex b) corresponding to midpoint

In  $\mathbb{R}^2$  and  $\mathbb{R}^3$  the basis functions are merely extensions of the one-dimensional basis functions. Typical elements in  $\mathbb{R}^2$  have been sketched in Figure 2.4.

With respect to the linear elements, the boundaries of the elements are usually straight, however for quadratic elements, the boundaries of the elements may be quadratic in order to get a good approximation of the boundary. In general one can state that the boundary must be approximated with the same type of polynomials as the solution. Once the basis functions have been constructed it is necessary to compute the integrals (2.24a) and (2.24b) in order to build the matrix and right-hand side of the system of equations (2.23). For a typical finite element grid as the one depicted in Figure 2.5, these computations seem rather complicated. For that reason the integrals over the region are split into integrals over the elements, i.e.

$$\int_{\Omega} \{A \nabla \varphi_i \cdot \nabla \varphi_i + \beta \varphi_i \varphi_j\} d\Omega = \sum_{k=1}^{n_e} \int_{\Omega^k} \{A \nabla \varphi_i \cdot \nabla \varphi_i + \beta \varphi_i \varphi_j\} d\Omega, \quad (2.28)$$

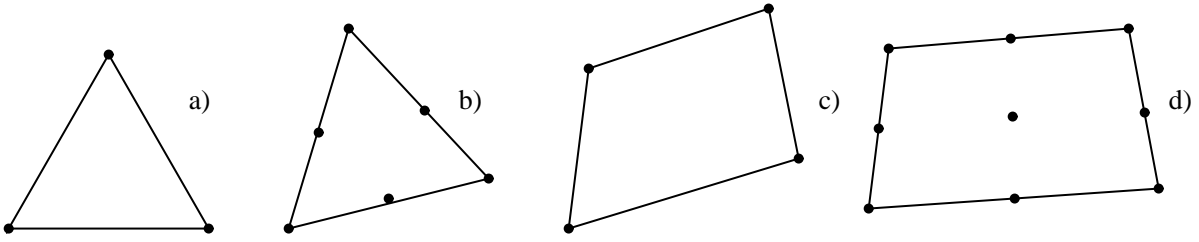


Figure 2.4: Examples of elements in  $\mathbb{R}^2$ : a) linear triangle, b) quadratic triangle, c) bilinear quadrilateral, d) biquadratic quadrilateral

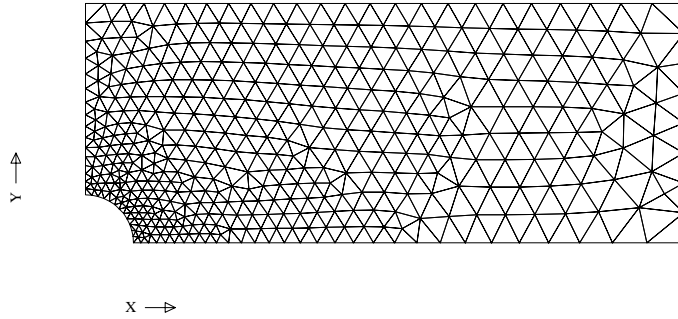


Figure 2.5: Typical example of a two-dimensional finite element mesh.

and so on. In (2.28)  $n_e$  is the number of elements and  $\Omega^{e_k}$  is the area of element  $e_k$ . Since the mesh generator produces automatically the topological information of the mesh, it is an easy task for the computer to carry out the additions. If we restrict ourselves to a typical element  $e_k$ , then it is clear that only a very little number of the possible integrals

$$\int_{\Omega^{e_k}} \{A \nabla \varphi_i \cdot \nabla \varphi_i + \beta \varphi_i \varphi_j\} d\Omega \quad (2.29)$$

are unequal to zero. Only those basis functions corresponding to nodal points in the element  $e_k$  have a non-zero contribution to the integrals. So it is sufficient to compute only those integrals that are non-zero on the element and store them in a so-called element matrix. For a linear triangle such an element matrix is for example of size  $3 \times 3$ .

In exactly the same way it is natural to introduce the so-called element vector, which in a linear triangle reduces to a  $3 \times 1$  vector, with elements

$$\int_{\Omega^{e_k}} f \varphi_i d\Omega \quad (2.30)$$

In order to compute the boundary integrals in (2.24a) and (2.24b) so-called boundary elements or line elements are introduced. These boundary elements are identical to the intersection

of the internal elements with the boundaries  $\Gamma_2$  and  $\Gamma_3$  and have no other purpose then to evaluate the boundary integrals. Here we have assumed that the boundary is identical to the outer boundary of the elements.

Hence we get:

$$\begin{aligned} \int_{\Gamma_3} \sigma \varphi_i \varphi_j d\Gamma &= \sum_{k=1}^{n_{be3}} \int_{\Gamma_3^{e_k}} \sigma \varphi_i \varphi_j d\Gamma \\ \int_{\Gamma_2} g_2 \varphi_i d\Gamma &= \sum_{k=1}^{n_{be2}} \int_{\Gamma_2^{e_k}} g_2 \varphi_i d\Gamma \\ \int_{\Gamma_3} g_3 \varphi_i d\Gamma &= \sum_{k=1}^{n_{be3}} \int_{\Gamma_3^{e_k}} g_3 \varphi_i d\Gamma \end{aligned} \quad (2.31)$$

## 2.5 Computation of the element matrix and element vector

The evaluation of the system of equations (2.23), (2.24a-2.24b) is now reduced to the computation of some integrals over an arbitrary element. For the sake of simplicity we shall restrict ourselves to  $\mathbb{R}^2$ . As an example we consider the computation of the integral given by (2.29):

$$S_{ij}^{e_k} = \int_{\Omega^{e_k}} \{A \nabla \varphi_j \cdot \nabla \varphi_i + \beta \varphi_j \varphi_i\} d\Omega \quad (2.32)$$

Before we are able to compute this integral it is necessary to compute the basis functions  $\varphi_i$ . For a so-called linear triangle (see Figure 2.6),  $\varphi_i(x) = \lambda_i(x)$  is defined by (2.26).

From (2.26,(1)) it follows that:

$$\lambda_i(x) = a_0^i + a_1^i x + a_2^i y \quad (2.33)$$

and from (2.26,(2)):

$XA = I$  , with

$$A = \begin{bmatrix} a_0^1 & a_0^2 & a_0^3 \\ a_1^1 & a_1^2 & a_1^3 \\ a_2^1 & a_2^2 & a_2^3 \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \quad (2.34)$$

and hence  $A = X^{-1}$

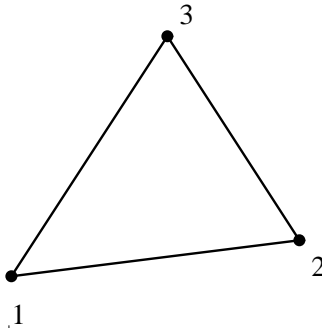


Figure 2.6: Linear triangle with local numbering

where the local numbering of Figure 2.6 is used. A necessary condition for the existence of

$\lambda_i(x)$  is that the determinant of the matrix to be inverted is unequal to zero. This determinant is given by:

$$\Delta = (x_2 - x_1)(y_3 - y_2) - (y_2 - y_1)(x_3 - x_2) \quad (\text{cyclic}) \quad (2.35)$$

One can prove that  $|\Delta|$  is equal to two times the area of the triangle, hence the triangle may not deform to a line. In practice it is necessary that the largest angle of the triangle is limited by some angle (for example  $150^\circ$ ).

The coefficients  $a_1^i$  and  $a_2^i$  are easily computed from (2.34) and one immediately verifies that they are given by:

$$\begin{aligned} a_1^1 &= \frac{1}{\Delta}(y_2 - y_3) & a_1^2 &= \frac{1}{\Delta}(y_3 - y_1) & a_1^3 &= \frac{1}{\Delta}(y_1 - y_2) \\ a_2^1 &= \frac{1}{\Delta}(x_3 - x_2) & a_2^2 &= \frac{1}{\Delta}(x_1 - x_3) & a_2^3 &= \frac{1}{\Delta}(x_2 - x_1) . \end{aligned} \quad (2.36)$$

Since  $\nabla \lambda_i = \begin{bmatrix} a_1^i \\ a_2^i \end{bmatrix}$ , (2.36) immediately defines the gradient of  $\varphi_i$ .

In general the integral (2.32) can not be computed exactly and some quadrature rule must be applied. A quadrature rule has usually the shape:

$$\int_{\Omega^{e_k}} \text{Int}(x) d\Omega = \sum_{k=1}^m w_k \text{Int}(\mathbf{x}_k), \quad (2.37)$$

where  $m$  is the number of quadrature points in the element,  $w_k$  are the weights and  $\mathbf{x}_k$  the co-ordinates of the quadrature points. We distinguish between the so-called Newton-Cotes rule based upon exact integration of the basis functions and so-called Gauss quadrature. The weights and quadrature points of the Gauss rules can be found in the literature, see for example Strang and Fix (1973) or Hughes (1987).

The Newton-Cotes rule can be derived by:

$$\text{Int}(x) = \sum_{k=1}^{n+1} \text{Int}(\mathbf{x}_k) \varphi_k(x), \quad (2.38)$$

where  $n + 1$  is the number of basis functions in the element, and application of the general rule:

$$\int_{\text{simplex}} \lambda_1^{m_1} \lambda_2^{m_2} \dots \lambda_{n+1}^{m_{n+1}} d\Omega = \frac{m_1! m_2! \dots m_{n+1}!}{(m_1 + m_2 + \dots + m_{n+1} + n)!} |\Delta|, \quad (2.39)$$

where  $n$  denotes the dimension of space. For a proof, see Holand and Bell (1969), page 84.

From (2.28) and (2.39) it follows that the Newton-Cotes rule for the linear element is defined by:

$$\int_{\Omega^{e_n}} \text{Int}(x) d\Omega = \frac{|\Delta|}{6} \sum_{k=1}^3 \text{Int}(\mathbf{x}_k), \quad (2.40)$$

where  $\mathbf{x}_k$  is the  $k^{\text{th}}$  vertex of the triangle.

Application of (2.40) to (2.32) gives

$$S_{ij}^{e_k} = \frac{|\Delta|}{6} \sum_{k=1}^3 A(\mathbf{x}_k) \nabla \varphi_i \cdot \nabla \varphi_j + \beta(\mathbf{x}_k) \delta_{ij} \quad (2.41)$$

In the same way (2.30) may be approximated by

$$f_i^{e_k} = \frac{|\Delta|}{6} f(\mathbf{x}_i). \quad (2.42)$$

In order to evaluate the boundary integrals (2.31) we use linear boundary elements as sketched in Figure 2.7.

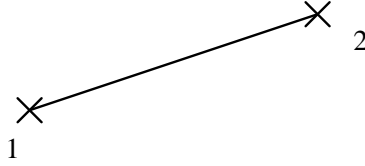


Figure 2.7: Linear boundary element in  $\mathbb{R}^2$ , with local numbering.

One easily verifies that the Newton-Cotes rule for this element is identical to the trapezoid rule:

$$\int_{\Gamma^e} \text{Int}(\mathbf{x}) = \frac{h}{2} (\text{Int}(\mathbf{x}_1) + \text{Int}(\mathbf{x}_2)), \quad (2.43)$$

where  $h$  is the length of the element:

$$h = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2.44)$$

Application of (2.43) to the three integrals (2.31) gives

$$\begin{aligned} \int_{\Gamma_3^{e_k}} \sigma \varphi_i \varphi_i d\Gamma &= \frac{h}{2} \sigma(\mathbf{x}_i) \delta_{ij}, \\ \int_{\Gamma_2^{e_k}} g_2 \varphi_i d\Gamma &= \frac{h}{2} g_2(\mathbf{x}_i), \\ \int_{\Gamma_3^{e_k}} g_3 \varphi_i d\Gamma &= \frac{h}{2} g_3(\mathbf{x}_i). \end{aligned} \quad (2.45)$$

## 2.6 Higher order elements

In (2.5) we have derived the element matrix and vector for linear triangles. However, in practice also other types of elements are used. A simple extension of the linear triangle is for example the quadratic triangle. For that element both the vertices and the mid-side points are used as nodal points. See Figure 2.8 for the definition of the nodes. One can verify that the basis functions  $\varphi_i(\mathbf{x})$  may be expressed in terms of the linear basis function  $\lambda_i(\mathbf{x})$  by:

$$\begin{aligned} \varphi_i(\mathbf{x}) &= \lambda_i(2\lambda_i - 1), \quad i = 1, 2, 3, \\ \varphi_{ij}(\mathbf{x}) &= 4\lambda_i\lambda_j, \quad 1 \leq i < j \leq 3. \end{aligned} \quad (2.46)$$

See for example Cuvelier et al (1986) for a derivation.

Quadrilateral elements are not so easy to derive. Nodal points will be either the vertices (bi-linear elements) or the vertices and midside points (bi-quadratic elements). However to derive the basis function the quadrilateral is mapped onto a square reference element. Such a mapping is plotted in Figure 2.9.

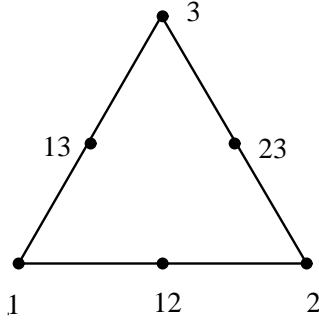


Figure 2.8: Quadratic triangle with nodal points and local numbering.

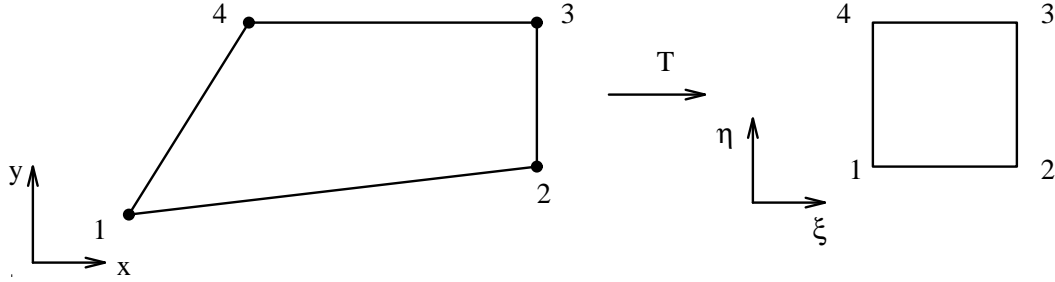


Figure 2.9: Mapping  $T$  from quadrilateral in  $(x, y)$ -plane onto square in  $(\xi, \eta)$  plane.

All basis functions are derived in the reference element by choosing products of one-dimensional basis functions. Also all integrals to be computed over the quadrilateral are transformed to integrals over the reference element.

For example:

$$\int_{\Omega_{xy}^{e_k}} A \nabla \varphi_i \cdot \nabla \varphi_j + \beta \varphi_i \varphi_j \, d\Omega_{xy} = \int_{\Omega_{\xi\eta}^{e_k}} \{A \nabla \varphi_i \cdot \nabla \varphi_j + \beta \varphi_i \varphi_j\} |J| \, d\Omega_{\xi\eta}, \quad (2.47)$$

where  $J$  is the Jacobian of the transformation. The transformation itself is a so-called isoparametric transformation defined by the basis functions in the following way:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \sum_{k=1}^4 \varphi_k(\xi, \eta) \begin{pmatrix} x_k \\ y_k \end{pmatrix}. \quad (2.48)$$

For details of the derivation of basis functions and element matrices and vectors the reader is referred to Cuvelier et al (1986).

## 2.7 Structure of the large matrix

The finite element method applied to the linear differential equation (2.3) leads to linear systems of equations of the form (2.23):

$$\mathbf{S} \mathbf{c} = \mathbf{F} \quad (2.49)$$

The matrix is often referred to as the stiffness matrix. From the relation (2.24a) it is clear that this matrix is symmetric. Furthermore one can prove that the matrix  $S$  is positive definite,

except in the case of Neumann boundary conditions and  $\beta = 0$ , in which case the matrix is singular because the original problem is singular. If we consider a number of adjacent triangles in a mesh as sketched in Figure 2.10, then it is clear that the basis function corresponding to nodal point  $i$  is only non-zero in those elements

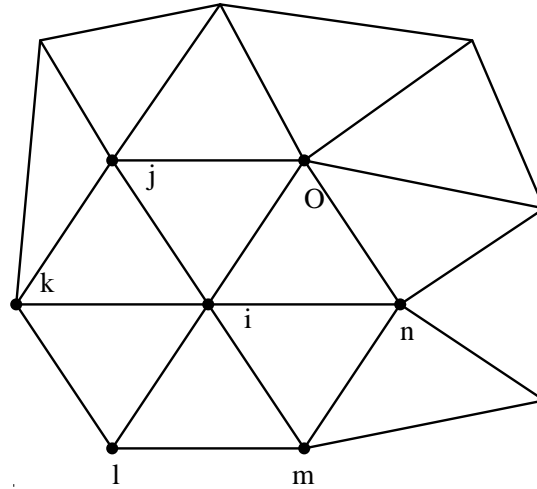


Figure 2.10: Nodal point  $i$  with direct neighbors  $j - o$  as part of a triangular mesh.

containing node  $i$ . As a consequence products with basis functions that correspond to nodes not in these triangles are zero.

If we identify row  $i$  in matrix  $\mathbf{S}$  with nodal point  $i$ , it is clear that only the entries  $s_{ii}, s_{ij}, \dots, s_{io}$  may be unequal to zero. All other matrix elements in row  $i$  are identical to zero. So we see that in the matrix  $\mathbf{S}$ , only a very limited number of entries is non-zero. Such a matrix is called sparse.

If the numbering of the nodal points is chosen in a clever way, the sparse matrix  $\mathbf{S}$  may have a band structure. For example if we consider the mesh in Figure 2.11 with rectangular elements, then a natural numbering is to use either a horizontal or a vertical numbering of

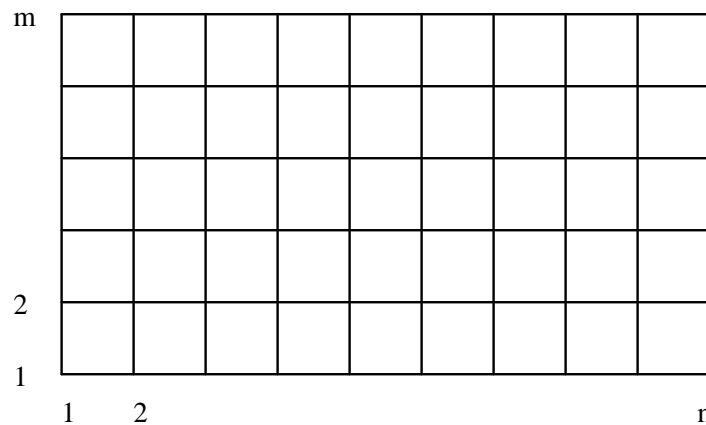


Figure 2.11: Rectangular elements in a rectangle, with  $n$  nodes in horizontal and  $m$  nodes in vertical direction.

the nodes. Figure 2.12a shows that in case of a horizontal numbering the band width is equal

to  $2n + 3$ , where  $n$  is the number of nodes in horizontal direction. Figure 2.12b shows that a vertical numbering leads to a band width of  $2m + 3$ , with  $m$  the number of nodes in vertical direction. Hence an optimal band width is achieved if nodes are numbered in the shortest direction.

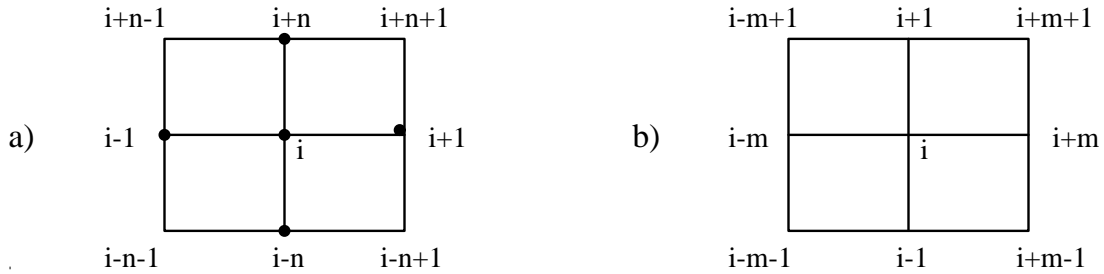


Figure 2.12: Connections of node  $i$  with neighbors a) horizontal numbering b) vertical numbering.

In general finite element meshes are not so structured as the one in Figure 2.11. and so locally a larger band width may be present. A typical example is sketched in Figure 2.13

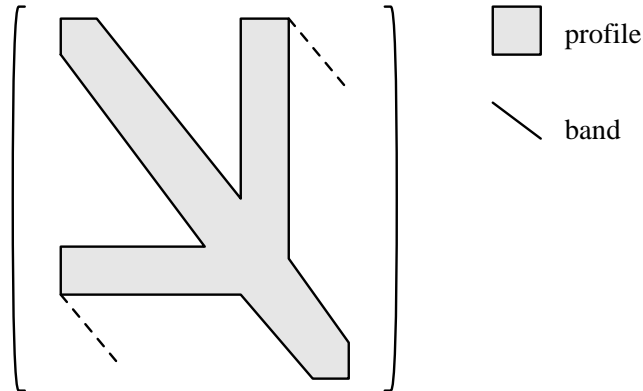


Figure 2.13: Example of a matrix with a local wide profile.

The external non-zero elements in this matrix define the so-called profile. A very simple example of a profile matrix is created by a one-dimensional example with periodical boundary conditions as sketched in Figure 2.14. In this example point  $i$  is connected to points  $i - 1$  and  $i + 1$  leading to a band width of 3. However, because of the periodical boundary conditions,

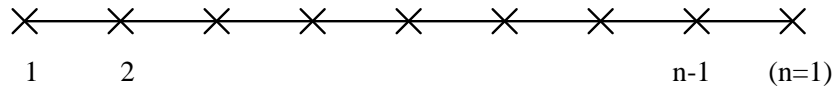


Figure 2.14: One-dimensional mesh, for problem with periodical boundary conditions.

point  $n$  and 1 have the same unknown and point 1 is connected to both  $n - 1$  and 2. Point  $n - 1$  connected to  $n - 2$  and 1. The corresponding matrix gets the structure as sketched in Figure 2.15. The band width of this matrix is equal to  $n - 1$ , which means that in case of a band storage, the matrix is full. The profile sketched in Figure 2.15b is much smaller.

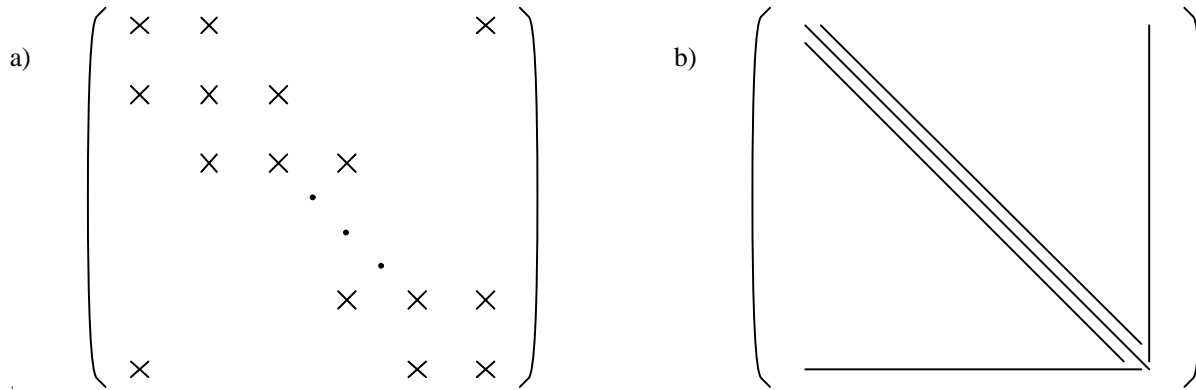


Figure 2.15: a) non-zero pattern of one-dimensional problem with periodical boundary conditions, b) corresponding profile.

Methods employing the band-structure are called band methods, whereas methods using the profile of the matrix only are called profile methods. Both methods belong to the class of direct solvers. Iterative methods fully utilize the sparsity pattern of the matrix and are therefore recommended in case of problems with many unknowns.

A good numbering may reduce the band width or the profile of the matrix considerably. For finite element methods various renumbering algorithms have been constructed. Many of them are variants of the so-called Cuthill-Mckee renumbering algorithm. See for example George and Liu (1981).

Part II of this book is devoted to efficient methods for the solution of systems of equations of the form (2.49).

### 3 Convection-diffusion equation by the finite element method

#### 3.1 Formulation of the equations

In this chapter we shall investigate the solution of convection-diffusion equations of the shape:

$$\rho \frac{\partial c}{\partial t} - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} (a_{ij} \frac{\partial c}{\partial x_j}) + \sum_{i=1}^n u_i \frac{\partial c}{\partial x_i} + \beta c = f . \quad (3.1)$$

Compared to equation (2.1), equation (3.1) is extended with the convective term

$$\sum_{i=1}^n u_i \frac{\partial c}{\partial x_i} , \quad (3.2)$$

or in vector notation

$$(\mathbf{u} \cdot \nabla c) , \quad (3.3)$$

where  $\mathbf{u}$  denotes the velocity.

In the stationary case, (3.1) reduces to:

$$-\text{div} (A \nabla c) + (\mathbf{u} \cdot \nabla c) + \beta c = f . \quad (3.4)$$

For a unique solution of (3.4) it is necessary to prescribe exactly one boundary condition at each part of the boundary. Theoretically exactly the same type of boundary conditions as for equation (2.3) may be used. In many practical flow problems, however, the convection term strongly dominates the diffusive terms. Numerically this means that the character of the equations resembles more that of the pure convection equation than that of the diffusion equation. For a pure convection equation, boundary conditions should only be given at inflow not at outflow. Since for the convection-diffusion equation, boundary conditions must be given at outflow, it is advised to use those boundary conditions that influence the solution as little as possible. In general this means that at outflow one usually applies natural boundary conditions; Dirichlet boundary conditions may result in unwanted wiggles.

With respect to the instationary equation it is not only necessary to prescribe boundary conditions, but also to define an initial condition.

In the remainder of this chapter we shall study the discretization of the convection-diffusion equation by finite elements and standard Galerkin. It will be shown that this discretization might introduce inaccurate solutions in the case of dominant convection. For that reason an upwinding technique is introduced. It will be shown that this upwinding improves the accuracy considerably.

#### 3.2 Standard Galerkin

In order to apply the standard Galerkin approach (SGA) the weak formulation of (3.1) under the boundary conditions (2.4) - (2.6) is derived.

Multiplication of (3.1) by a time-independent test function  $v$  and integration over the domain yields:

$$\int_{\Omega} \rho \frac{\partial c}{\partial t} v d\Omega + \int_{\Omega} \{-div (A\nabla c) + (\mathbf{u} \cdot \nabla c) + \beta c\} v d\Omega = \int_{\Omega} f v d\Omega . \quad (3.5)$$

After application of the Gauss divergence theorem, which results in relation (2.12), (3.5) can be written as

$$\int_{\Omega} \rho \frac{\partial c}{\partial t} v d\Omega + \int_{\Omega} \{(A\nabla c \cdot \nabla v + \beta c v + \mathbf{u} \cdot \nabla c) v\} d\Omega - \int_{\Gamma} v A\nabla c \cdot \mathbf{n} d\Gamma = \int_{\Omega} f v d\Omega . \quad (3.6)$$

Substitution of the boundary conditions in the same way as is performed in Chapter 2 results in the weak formulation:

Find  $c(\mathbf{x}, t)$  with  $c(\mathbf{x}, 0)$  given and  $c|_{\Gamma_1} = g_1$  such that

$$\begin{aligned} & \int_{\Omega} \rho \frac{\partial c}{\partial t} v d\Omega + \int_{\Omega} \{(A\nabla c \cdot \nabla v) + \mathbf{u} \cdot \nabla c v + \beta c v\} d\Omega + \int_{\Gamma_3} \sigma c v d\Gamma = \\ & \int_{\Omega} f v d\Omega + \int_{\Gamma_2} g_2 v d\Gamma + \int_{\Gamma_3} g_3 v d\Gamma , \end{aligned} \quad (3.7)$$

for all functions  $v(\mathbf{x})$  satisfying  $v|_{\Gamma_1} = 0$ .

In the SGA the weak form (3.7) is approximated by a finite dimensional subspace. To that end we define time-independent basis functions in exactly the same way as for the potential problem. The solution  $c$  is approximated by a linear combination of the basis functions:

$$c_h(\mathbf{x}, t) = \sum_{j=1}^n c_j(t) \varphi_j(\mathbf{x}) + c_0(\mathbf{x}, t) . \quad (3.8)$$

The basis functions  $\varphi_j(\mathbf{x})$  and the function  $c_0(\mathbf{x}, t)$  must satisfy the same requirements as in Chapter 2, i.e. (2.18) and (2.19) are still necessary.

For the test functions  $v(\mathbf{x})$  again the basis functions  $\varphi_i(\mathbf{x})$  ( $i = 1, 2, \dots, n$ ) are substituted. So finally we arrive at the Galerkin formulation:

$$\begin{aligned} & \sum_{j=1}^n \frac{\partial c_j}{\partial t} \int_{\Omega} \varphi_i \varphi_j d\Omega + \sum_{j=1}^n c_j \left\{ \int_{\Omega} [(A\nabla \varphi_j \cdot \nabla \varphi_i) + (\mathbf{u} \cdot \nabla \varphi_j) \varphi_i + \beta \varphi_i \varphi_j] d\Omega \right. \\ & \left. + \int_{\Gamma_3} \sigma \varphi_i \varphi_j d\Gamma \right\} = \int_{\Omega} f \varphi_i d\Omega + \int_{\Gamma_2} g_2 \varphi_i d\Gamma + \int_{\Gamma_3} g_3 \varphi_i d\Gamma \\ & - \int_{\Omega} \{(A\nabla c_0 \cdot \nabla \varphi_i) + \beta c_0 \varphi_i + (\mathbf{u} \cdot \nabla c_0) \varphi_i\} d\Omega - \int_{\Omega} \frac{\partial c_0}{\partial t} \varphi_i d\Omega , \quad i = 1(1)n . \end{aligned} \quad (3.9)$$

Clearly (3.9) forms a system of  $n$  linear ordinary differential equations with  $n$  unknowns, which can be written in matrix-vector notation as

$$\mathbf{M}\dot{\mathbf{c}} + \mathbf{S}\mathbf{c} = \mathbf{F} , \quad (3.10)$$

where  $\cdot$  denotes differentiation with respect to time,  $M$  is the so-called mass matrix and  $S$  the stiffness matrix. The elements of the matrices and right-hand side are defined by:

$$m_{ij} = \int_{\Omega} \varphi_i \varphi_j d\Omega \quad (3.11a)$$

$$s_{ij} = \int_{\Omega} \{(A \nabla \varphi_j \cdot \nabla \varphi_j) + (\mathbf{u} \cdot \nabla \varphi_j) \varphi_i + \beta \varphi_i \varphi_j\} d\Omega + \int_{\Gamma_3} \sigma \varphi_i \varphi_j d\Gamma \quad (3.11b)$$

$$\begin{aligned} F_i &= \int_{\Omega} f \varphi_i d\Omega - \int_{\Omega} \{(A \nabla c_0 \cdot \nabla \varphi_i) + (\mathbf{u} \cdot \nabla c_0) \varphi_i + \beta c_0 \varphi_i\} d\Omega \\ &\quad - \int_{\Omega} \frac{\partial c_0}{\partial t} \varphi_i d\Omega + \int_{\Gamma_2} g_2 \varphi_i d\Gamma + \int_{\Gamma_3} g_3 \varphi_i d\Gamma . \end{aligned} \quad (3.11c)$$

The construction of the basis functions and the computation of the integrals is exactly the same as for the potential problem. The only extra parts are the time-derivative with the mass matrix and the extra convective terms in the stiffness matrix. Due to these extra convective terms the stiffness matrix becomes non-symmetric.

The computation of the mass-matrix can be performed exactly or by a quadrature rule. In general  $M$  has exactly the same structure as  $S$ . However, if the integrals (3.11a) are computed by the Newton Cotes rule corresponding to the basis functions, the matrix  $M$  reduces to a diagonal matrix. In that case one speaks of a lumped mass-matrix. A non-lumped mass matrix is also known as a consistent mass matrix. The (dis-)advantages of both types of matrices will be treated in Paragraphs 3.3 and 3.5.

In the next paragraph we shall consider some methods to solve the instationary equations. After that, problems in case of a dominant convective term will be investigated and a upwind technique will be introduced.

### 3.3 Solution of the system of ordinary differential equations

The discretization of the instationary convection-diffusion equation results in a system of ordinary differential equations of the shape (3.10). In order to solve this system of equations any classical method for the solution of ordinary differential equations may be used.

In general one may distinguish between explicit and implicit methods and between one-step and multi-step methods. In this chapter we shall restrict ourselves to one-step methods only. That means that to compute the solution at a certain time-step only information of the preceding time-step is used and not of older time-steps.

In most ordinary differential equation solvers the time derivative in (3.10) is replaced by a forward difference discretization:

$$\dot{\mathbf{c}} = \frac{\mathbf{c}^{k+1} - \mathbf{c}^k}{\Delta t} , \quad (3.12)$$

where  $k$  denotes the present time-level and  $k + 1$  the next time-level. A method is called explicit if the term  $\mathbf{S}\mathbf{c}$  is only taken at the time-level  $k$ . As soon as  $\mathbf{S}\mathbf{c}$  is also taken at the new time-level  $k + 1$ , the method is called implicit. The reason is that in that case always a system of equations has to be solved, even if the matrix  $M$  is the identity matrix.

Among the many available methods for solving the system (3.10) we restrict ourselves to the

so-called  $\theta$ -method:

$$\mathbf{M} \frac{\mathbf{c}^{k+1} - \mathbf{c}^k}{\Delta t} + \theta \mathbf{S} \mathbf{c}^{k+1} + (1 - \theta) \mathbf{S} \mathbf{c}^k = \theta \mathbf{F}^{k+1} + (1 - \theta) \mathbf{F}^k, \quad 0 \leq \theta \leq 1, \quad (3.13)$$

The most common values for  $\theta$  are:

$$\begin{aligned} \theta &= 0, & \text{Explicit Euler} \\ \theta &= 1, & \text{Implicit Euler and} \\ \theta &= 1/2, & \text{Implicit Heun or Crank Nicolson.} \end{aligned}$$

For  $\theta = 0$ , (3.13) reduces to

$$\mathbf{M} \mathbf{c}^{k+1} = (\mathbf{M} - \Delta t \mathbf{S}) \mathbf{c}^k + \Delta t \mathbf{F}^k. \quad (3.14)$$

Although it concerns an explicit method, we still have to solve a system of equations. However, in the case of a lumped mass matrix, the solution implies only the inversion of a diagonal matrix. In that case an explicit method is relatively cheap. A clear disadvantage of an explicit method is that the time-step must be restricted in order to get a stable solution. For example in the case of a pure time-dependent diffusion problem a stability criterion of the shape

$$\Delta t \leq C \Delta x^2 \quad (3.15)$$

is required, where  $C$  is some constant and  $\Delta x$  a local diameter of the elements.

In the case of a dominant convection, the Euler explicit method is not longer stable and one should use either an implicit method or a higher order explicit method. For such problems the classical fourth order Runge Kutta method is a good choice.

For  $\theta = 1$  (3.13) reduces to

$$(\mathbf{M} + \Delta t \mathbf{S}) \mathbf{c}^{k+1} = \mathbf{M} \mathbf{c}^k + \Delta t \mathbf{F}^{k+1},$$

which is a purely implicit method. One can show that this method is unconditionally stable, for the convection equation (see for example Cuvelier et al 1986), so the only reason to restrict the time-step is because of accuracy requirements. It can be easily verified that the accuracy of both the implicit and the explicit Euler time-discretization is of  $O(\Delta t)$ . The implicit Euler method belongs to the class of ultra-stable methods, which means that errors in time always will be damped.

The most accurate scheme is achieved for  $\theta = 1/2$  (Crank Nicolson). This scheme can be written as

$$\left(\mathbf{M} + \frac{\Delta t}{2} \mathbf{S}\right) \mathbf{c}^{k+1} = \left(\mathbf{M} - \frac{\Delta t}{2} \mathbf{S}\right) \mathbf{c}^k + \frac{\Delta t}{2} (\mathbf{F}^k + \mathbf{F}^{k+1}). \quad (3.16)$$

One can show that this scheme is also unconditionally stable and that the accuracy is one order higher, i.e. of  $O(\Delta t^2)$ . This scheme does not have the damping property of Euler implicit and as a consequence once produced errors in time will always be visible. This one usually starts in these cases with one step Euler implicit.

So the solution of the systems of ordinary differential equations is always reduced to a time-stepping algorithm in combination with matrix-vector multiplications, and sometimes the solution of a system of linear equations.

For  $\theta \neq 0$  it is easier to replace the  $\theta$ -method (3.13) by the so-called modified  $\theta$ -method:

$$\mathbf{M} \frac{\mathbf{c}^{k+\theta} - \mathbf{c}^k}{\Delta t} + \theta \mathbf{S} \mathbf{c}^{k+\theta} = \mathbf{F}^{k+\theta}, \quad 0 \leq \theta \leq 1 \quad (3.17)$$

$$\mathbf{c}^{k+1} = \frac{1}{\theta} \mathbf{c}^{k+\theta} + \frac{1-\theta}{\theta} \mathbf{c}^k \quad (3.18)$$

One can prove that equation (3.13) is equal to equation (3.17) if the system of equations to be solved is linear. In case of a non-linear system the approximation (3.17) is of the order  $\Delta t^2$ . A clear advantage of (3.17) above (3.13) is that the matrix to be solved is always independent of  $\theta$  and that no explicit matrix-vector multiplication is required.

A disadvantage of the  $\theta$ -method is the fixed  $\theta$ . It could be advantageous to combine a number of different  $\theta$ 's per time step in such a way that second order accuracy is accomplished, and some damping is ensured as well. Two methods that offer this opportunity are the fractional  $\theta$ -method and the generalized  $\theta$ -method. The latter is a generalization of the fractional  $\theta$ -method, so we will restrict ourselves to the description of the generalized  $\theta$ -method. We rewrite equation (3.13) as follows, letting  $\Sigma_k = \sum_{i=1}^k \theta_i$ :

$$\begin{aligned} c^{n+\Sigma_2} &= c^n + \Delta t \left( \theta_1 f(x, t^n) + \theta_2 f(x, t^{n+\Sigma_2}) \right) \\ c^{n+\Sigma_4} &= c^{n+\Sigma_2} + \Delta t \left( \theta_3 f(x, t^{n+\Sigma_2}) + \theta_4 f(x, t^{n+\Sigma_4}) \right) \\ &\vdots \\ c^{n+\Sigma_{2k}} &= c^{n+\Sigma_{2k-2}} + \Delta t \left( \theta_{2k-1} f(x, t^{n+\Sigma_{2k-2}}) + \theta_{2k} f(x, t^{n+\Sigma_{2k}}) \right) \end{aligned} \quad (3.19)$$

There are two necessary conditions:

1.  $\Sigma_{2k} = 1$  for a  $k$ -stage method. This gives a first order method, and is only a scaling requirement.
2.  $\sum_{i=1}^k \theta_{2i-1}^2 = \sum_{i=1}^k \theta_{2i}^2$  to guarantee second order accuracy.

A third condition is optional, but guarantees some damping:

1.  $\theta_{2i-1} = 0$  for at least one  $i \in 1, \dots, k$ .

This condition includes at least one Implicit Euler step per time step.

The generalized  $\theta$ -method is a 3-stage method, and is therefore 3 times as expensive as the Crank-Nicolson method. However, one may choose  $\Delta t_{gen\theta} = 3 \cdot \Delta t_{CN}$  to accomplish similar results for both methods. A common choice for the generalized  $\theta$ -method is the following 'optimum' for  $k = 3$ :

$$\begin{aligned} \theta_1 = \theta_5 &= \frac{\alpha}{2}, & \theta_3 &= 0, \\ \theta_2 = \theta_6 &= \alpha \frac{\sqrt{3}}{6}, & \theta_4 &= \alpha \frac{\sqrt{3}}{3} \\ \alpha &= \left( 1 + \frac{2}{\sqrt{3}} \right)^{-1}. \end{aligned} \quad (3.20)$$

A common choice for the fractional  $\theta$ -method is the following:

$$\begin{aligned}\theta_1 = \theta_5 &= \beta\theta, & \theta_3 &= \alpha(1 - 2\theta), \\ \theta_2 = \theta_6 &= \alpha\theta, & \theta_4 &= \beta(1 - 2\theta), \\ \alpha &= \frac{1 - 2\theta}{1 - \theta}, & \beta &= \frac{\theta}{1 - \theta}, \\ \theta &= 1 - \frac{1}{2}\sqrt{2}.\end{aligned}\tag{3.21}$$

### 3.4 Accuracy aspects of the SGA

One can show that the SGA in combination with the FEM yields an accuracy of  $O(h^{k+1})$ , where  $h$  is some representative diameter of the elements and  $k$  is the degree of the polynomials used in the approximation per element. However, this is only true for problems where the convection does not dominate the diffusion. As soon as the convection dominates, the accuracy strongly decreases as can be seen in Table 3.1, which shows the accuracy of the following artificial mathematical example:

$$-\varepsilon\Delta c + \mathbf{u} \cdot \nabla c = f, \quad x \in \Omega \tag{3.22a}$$

$$c(x, y) = \sin(x) \sin(y), \quad x \in \Gamma \tag{3.22b}$$

$$\text{where } \mathbf{u} = \begin{pmatrix} x \\ y \end{pmatrix} \text{ and } f = 2\varepsilon \sin(x) \sin(y) + x \cos(x) \sin(y) + y \sin(x) \cos(y). \tag{3.22c}$$

One easily verifies that the exact solution of this problem is given by

$$c(x, y) = \sin(x) \sin(y). \tag{3.23}$$

This problem has been solved on the square  $(0, 1) \times (0, 1)$  using linear and quadratic elements. Table 3.1 shows the maximal error for  $\varepsilon = 1, 10^{-3}$  respectively  $10^{-6}$  and triangular elements. The results for quadrilaterals are comparable. In the linear case a subdivision in  $6 \times 6, 11 \times 11$  respectively  $21 \times 21$  nodes has been made, in the quadratic case only  $11 \times 11$  and  $21 \times 21$  nodes have been used. From this table we may draw the following conclusions:

number of nodes	linear triangles			quadratic triangles		
	$\varepsilon = 1$	$\varepsilon = 10^{-3}$	$\varepsilon = 10^{-6}$	$\varepsilon = 1$	$\varepsilon = 10^{-3}$	$\varepsilon = 10^{-6}$
$6 \times 6$	$3.0_{10^{-4}}$	$3.3_{10^{-2}}$	$3.9_{10^1}$	-	-	-
$11 \times 11$	$7.6_{10^{-5}}$	$4.1_{10^{-3}}$	$1.5_{10^0}$	$8.0_{10^{-6}}$	$5.4_{10^{-3}}$	$2.2_{10^0}$
$21 \times 21$	$1.9_{10^{-5}}$	$1.0_{10^{-3}}$	$8.4_{10^{-2}}$	$6.3_{10^{-7}}$	$4.8_{10^{-4}}$	$9.8_{10^{-2}}$

Table 3.1: Error in max-norm of convection-diffusion problem (3.22a-3.22c) for various values of  $\varepsilon$ . Linear and quadratic triangles

- for relatively small convection the accuracy of the linear elements is  $O(h^2)$ , and for the quadratic elements at least  $O(h^3)$ ,
- for convection-dominant flow the numerical solution is very inaccurate especially for coarse grids,

- the use of quadratic elements makes only sense for problems with small convection.

Remark: the conclusions are based on an example with a very smooth solution. For problems with steep gradients the conclusion may be different, especially for the quadratic elements, in which cases the  $O(h^3)$  cannot be expected anymore.

The most important part of the conclusion is that SGA is not a good method for convection-dominant flows. This conclusion is also motivated by the following less trivial problem.

### Rotating cone problem

Consider the region  $\Omega$  sketched in Figure 3.1. The region consists of a square with a cut  $B$ . In

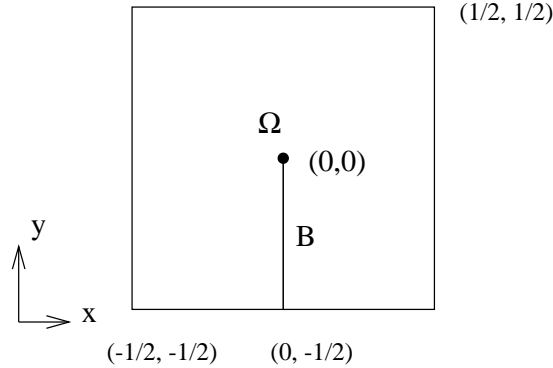


Figure 3.1: Definition region for rotating cone problem

the inner region we suppose that the concentration satisfies the convection-diffusion equation

$$-\varepsilon \Delta c + \mathbf{u} \cdot \nabla c = 0, \quad (3.24)$$

where  $\varepsilon$  is chosen equal to  $10^{-6}$  and the velocity  $\mathbf{u}$  is such that the flow rotates counter clockwise. This is achieved by setting  $\mathbf{u} = \begin{bmatrix} -y \\ x \end{bmatrix}$ . At the outer boundary we use the boundary condition

$$c|_{\Gamma} = 0. \quad (3.25)$$

On the starting curve  $B$  the concentration  $c$  is set equal to

$$c|_B = \cos\left(2\pi\left(y + \frac{1}{4}\right)\right), \quad (3.26)$$

and due to the small diffusion one expects that the concentration at the end curve is nearly the same. The end curve has the same co-ordinates as  $B$  but the nodal points differ, which means that the solution may be different from the starting one. Since no boundary condition is given at the outflow curve "B" implicitly the boundary condition

$$\varepsilon \frac{\partial c}{\partial n} \Big|_B = 0, \quad (3.27)$$

is prescribed.

Figure 3.2 shows a  $23 \times 23$  mesh consisting of triangles. The direction of the diagonals in the squares are chosen randomly. Figure 3.3 shows the lines of equal concentration. For the exact solution these should be concentration circles with levels  $0, 0.1, \dots, 10$ . However, the standard Galerkin method completely destroys the result. Finally Figure 3.4 shows a 3D plot of the concentration, which contains a large number of wiggles. It must be remarked that the solution is relatively smooth in the case of a grid consisting of squares or triangles all pointing in the same direction.

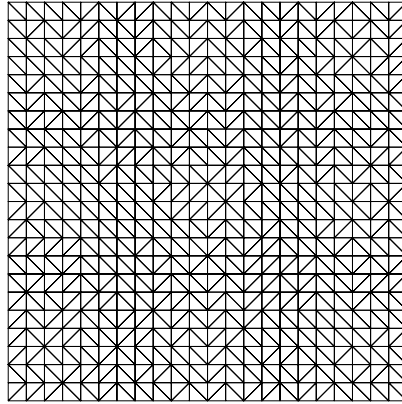


Figure 3.2: Triangular mesh for rotating cone problems random diagonals.

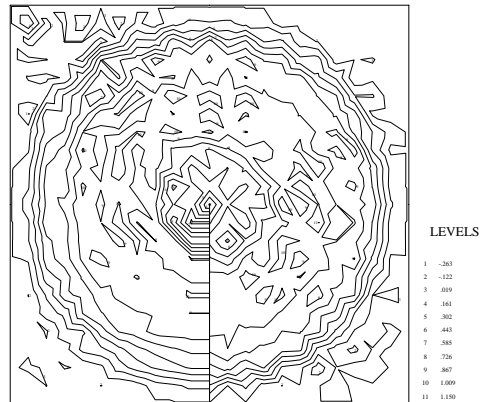


Figure 3.3: Equi-concentration lines for rotating cone problem computed by SGA.

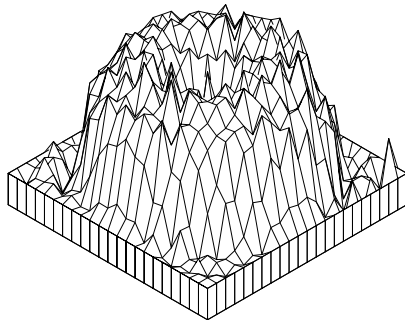


Figure 3.4: 3D plot of concentration in rotating cone problem computed by SGA.

### 3.5 Streamline Upwind Petrov Galerkin

In the previous paragraph we have seen that the SGA method may lead to wiggles and inaccurate results for convection-dominant flows. In finite differences this phenomenon is well known for a long time and one has tried to solve it by so-called upwind methods. This has motivated researchers in finite elements to construct schemes, which are comparable to classical finite difference upwind schemes. Among the various upwind techniques for the FEM, the streamline upwind Petrov-Galerkin method (SUPG) is the most popular one. This method has first been derived by Brooks and Hughes (1982) and is later on improved by a large number of authors.

Starting point for SUPG is the weak formulation (3.5). However, instead of choosing the test function in the same space as the solution a test function  $\bar{v}$  is introduced according to

$$\bar{v} = v + p, \quad (3.28)$$

where  $v$  is the classical test function and  $p$  denotes a correction in order to take care of the upwinding part. Substitution of (3.28) in (3.5) gives:

$$\int_{\Omega} \rho \frac{\partial c}{\partial t} (v + p) d\Omega + \int_{\Omega} \{-div (A \nabla c) + \mathbf{u} \cdot \nabla c + \beta c\} (v + p) d\Omega = \int_{\Omega} f (v + p) d\Omega. \quad (3.29)$$

The function  $v$  is chosen in the same space as the solution, which means that the first derivative is square integrable. However, with respect to the function  $p$ , we assume that it may be discontinuous over the elements. As a consequence Gauss' divergence theorem may only be applied to the  $v$  part of (3.29). Hence after integration by parts we get

$$\begin{aligned} & \int_{\Omega} \rho \frac{\partial c}{\partial t} v d\Omega + \int_{\Omega} \{A \nabla c \cdot \nabla v + (\mathbf{u} \cdot \nabla c) v + \beta c v\} d\Omega + \int_{\Gamma_3} \sigma c v d\Gamma \\ & + \int_{\Omega} \{\rho \frac{\partial c}{\partial t} - div A \nabla c + \mathbf{u} \cdot \nabla c + \beta c - f\} p d\Omega = \int_{\Omega} f v d\Omega + \int_{\Gamma_2} g_2 v d\Gamma + \int_{\Gamma_3} g_3 v d\Gamma. \end{aligned} \quad (3.30)$$

Actually the second derivative of  $c$  does not have to exist over the element boundaries and is certainly not integrable in the examples of elements we have given before. So we are not able to compute the integral containing the  $p$  term. In order to solve that problem the integral is split into a sum of integrals over the elements, and the inter-element contributions are

neglected. So instead of (3.30) we write:

$$\begin{aligned}
& \int_{\Omega} \rho \frac{\partial c}{\partial t} v d\Omega + \int_{\Omega} \{A \nabla c \cdot \nabla v + (\mathbf{u} \cdot \nabla c) v + \beta c v\} d\Omega + \int_{\Gamma_3} \sigma c v d\Gamma \\
& + \sum_{k=1}^{n_e} \int_{\Omega^{e_k}} \left\{ \rho \frac{\partial c}{\partial t} - \operatorname{div} A \nabla c + \mathbf{u} \cdot \nabla c + \beta c \right\} p d\Omega = \\
& \int_{\Omega} f v d\Omega + \int_{\Gamma_2} g_2 v d\Gamma + \int_{\Gamma_3} g_3 v d\Gamma + \sum_{k=1}^{n_e} \int_{\Omega^{e_k}} f p d\Omega .
\end{aligned} \tag{3.31}$$

One can see that the approximation (3.31) itself is consistent since it consists of a standard Galerkin part, which itself is consistent, and a summation of residuals of the differential equation per element multiplied by a function. With consistency we mean that at least the constant and first term of the Taylor series expansion of the solution are represented exactly.

At this moment the choice of the function  $p$  per element is completely free. However, it is clear that the choice of  $p$  actually defines the type of SUPG method used. In fact a complete class of different SUPG methods may be defined by different choices of  $p$ .

A common choice for the function  $p$  is inspired by the one-dimensional stationary diffusion equation:

$$-\varepsilon \frac{d^2 c}{dx^2} + u \frac{dc}{dx} = 0 , \tag{3.32}$$

with boundary conditions

$$c(0) = 0 \quad , \quad c(1) = 1 . \tag{3.33}$$

The solution of (3.32), (3.33) is sketched in Figure 3.5 for  $\varepsilon = 0.01$ . It has a steep gradient in the neighborhood of  $x = 1$ . The size of this gradient depends on the value of  $\varepsilon$ . The smaller  $\varepsilon$ , the steeper the gradient. If a central difference scheme:

$$\begin{aligned}
-\varepsilon \frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2} + u \frac{c_{i+1} - c_{i-1}}{2\Delta x} &= 0 , \quad (i = 1, 2, \dots, n) , \\
c_0 = 0 , \quad c_{n+1} &= 1 ,
\end{aligned} \tag{3.34}$$

is applied with an equidistant step-size  $\Delta x$ , the solution shows wiggles as long as  $\Delta x > \frac{2}{Pe}$ , where the Peclet number  $Pe$  is defined as

$$Pe = \frac{u}{\varepsilon} . \tag{3.35}$$

Figure 3.5 shows an example for  $\Delta x = 0.1$ ,  $u = 1$  and  $\varepsilon = 0.01$ .

In the classical finite difference upwind scheme one tries to get rid of these wiggles by replacing the first derivative by a backward difference scheme provided the velocity  $u$  is positive. The idea for this choice is based on the fact that for a pure convection problem all information is transported from left to right and hence the discretization of the convective term should also be based upon information from the left. Figure 3.6 shows the result of the upwinding; the wiggles have been disappeared and the numerical solution has been smoothed. This figure makes it clear that, although backward differences suppress the wiggles, it also makes the

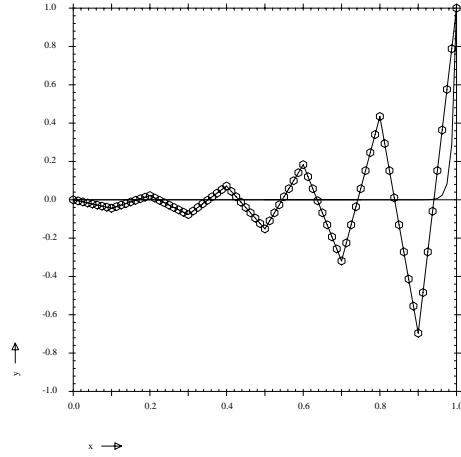


Figure 3.5: Solution of equation (3.32) for  $\varepsilon = 0.01$  and  $u = 1$ : — exact solution, - o - numerical solution for  $\Delta x = 0.1$  and central differences.

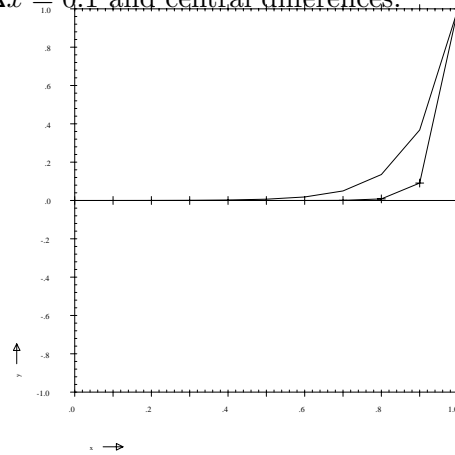


Figure 3.6: Solution of equation (3.32) for  $\varepsilon = 0.01$  and  $u = 1$ ; — exact solution, -+- numerical solution for  $\Delta x = 0.1$  and backward differences.

solution inaccurate. In the literature many upwinding schemes for finite difference methods have been derived which are much more accurate than the backward difference scheme.

The backward difference scheme for (3.32), (3.33) reads:

$$-\varepsilon \frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2} + u \frac{c_i - c_{i-1}}{\Delta x} = 0 \quad (i = 1, 2, \dots, n), \quad (3.36)$$

$$c_0 = 0, \quad c_{n+1} = 1.$$

Using Taylor series expansion one can show that (3.36) gives a local truncation error of

$$-\frac{u\Delta x}{2} \frac{d^2c}{dx^2} + O(\Delta x^2), \quad (3.37)$$

which is only of order  $\Delta x$  instead of  $\Delta x^2$ . Moreover, the second derivative appears in the truncation error, which implies that in fact one may consider (3.36) as the discretization of a

convection-diffusion equation with a diffusion of

$$\varepsilon + \frac{u\Delta x}{2} . \quad (3.38)$$

In fact (3.36) can be derived by taking the central difference scheme of the differential equation

$$-(\varepsilon + \frac{u\Delta x}{2}) \frac{d^2 c}{dx^2} + u \frac{dc}{dx} = 0 , \quad (3.39)$$

$$c(0) = 0 , \quad c(1) = 1 .$$

The term  $-\frac{u\Delta x}{2} \frac{d^2 c}{dx^2}$  is usually called artificial diffusion. Many of the upwind schemes in finite differences may be considered as a central difference scheme with artificial diffusion. For example the exact solution of (3.32) is constructed by the Il'in scheme, which may be considered as a central difference scheme with artificial diffusion equal to

$$\frac{u\Delta x}{2} \bar{\xi} \frac{d^2 c}{dx^2} , \quad (3.40a)$$

$$\bar{\xi} = \coth(\alpha) - \frac{1}{\alpha} , \quad \alpha = \frac{u\Delta x}{2\varepsilon} . \quad (3.40b)$$

Different schemes lead to different choices of  $\bar{\xi}$ . The following values of  $\bar{\xi}$  are commonly proposed:

$$\text{Classical upwind scheme} \quad \bar{\xi} = \text{sign}(\alpha) , \quad (3.41a)$$

$$\text{Il'in scheme} \quad \bar{\xi} = \coth(\alpha) - 1/\alpha . \quad (3.41b)$$

$$\text{Double asymptotic approximation} \quad \bar{\xi} = \begin{cases} \alpha/3 & -3 \leq \alpha \leq 3 \\ \text{sign}(\alpha) & |\alpha| > 3 . \end{cases} \quad (3.41c)$$

$$\text{Critical approximation} \quad \bar{\xi} = \begin{cases} -1 - 1/\alpha & \alpha \leq -1 \\ 0 & -1 \leq \alpha \leq 1 \\ 1 - 1/\alpha & \alpha \geq 1 . \end{cases} \quad (3.41d)$$

The last choice is such that the amount of artificial diffusion is minimal in order to get a diagonally dominant matrix. In this way it may be considered as a variant of the so-called hybrid method.

This observation motivates us to choose the function  $p$  such that an artificial diffusion of the form (3.40a) is constructed. If we confine ourselves to linear elements, the second derivative of the approximate solution is zero per element and hence SUPG applied to (3.32), (3.33) reduces to

$$\int_{\Omega} \left\{ u \frac{dc_h}{dx} v_h + \varepsilon \frac{dc_h}{dx} \frac{dv_h}{dx} \right\} d\Omega + \sum_{k=1}^{n_e} \int_{\Omega^k} p_h u \frac{dc_h}{dx} d\Omega = 0 , \quad (3.42)$$

where  $c_h$  is defined by (3.8),  $v_h$  represents the discretized classical test function and  $p_h$  the discretization of the extra function  $p$ .

In order to get an artificial diffusion of the shape (3.40a) it is sufficient to choose  $p_h$  equal to

$$p_h = \frac{h\bar{\xi}}{2} \frac{dv_h}{dx} , \quad (3.43)$$

where  $h = \Delta x$ .

With Taylor series expansion it can be shown that if  $\bar{\xi}$  is chosen according to one of the possible values of (3.41a-3.41d) (except the choice  $\bar{\xi} = 1$ ), the accuracy of the scheme is  $O(\Delta x^2) + \varepsilon O(\Delta x)$ , which may be considered to be of  $O(\Delta x^2)$  for small values of  $\varepsilon$ .

If the step size  $\Delta x$  is not a constant,  $h$  in formula (3.43) must be replaced by the step size. For quadratic elements  $h$  equal to half the local element width, has proven to be a good choice.

If we apply the SUPG method based upon formula (3.43) in 2D in each of the directions, a typical cross-wind diffusion arises. That means that the solution perpendicular to the flow direction is smoothed and becomes inaccurate. For that reason the SUPG method must be extended in such a way that the upwinding is applied in the direction of the flow only. Brooks and Hughes (1982) have solved this problem by giving the perturbation parameter  $p$  a tensor character

$$p = \frac{h\bar{\xi}}{2} \frac{\mathbf{u} \cdot \nabla v_h}{\|\mathbf{u}\|}. \quad (3.44)$$

In this formula  $h$  is the local element width, which may depend on the quadrature point. Mizukami (1985) has extended (3.44) for triangles.

Many extensions of the SUPG method have been proposed, all based on different choices of the function  $p$ . These improvements usually have a special function, for example to create monotonous solutions (Rice and Schnipke 1984), discontinuity capturing (Hughes et al 1986), or for time-dependent problems (Shahib 1988).

The SUPG method differs from the classical upwind methods in the sense that not only the advective term is perturbed, but also the right-hand side and the time derivative. This has two important consequences:

- the treatment of source terms is considerably better than for classical upwind techniques.
- the mass matrix is non-symmetric and may not be lumped. Hence explicit methods are as expensive as implicit ones.

Table 3.2 shows the example of Table 3.1 but now with SGA replaced by SUPG. The improvement for small values of  $\varepsilon$  and coarse grids is immediately clear. This table does not clearly show the accuracy of the method in terms of orders  $\Delta x^p$ . Besides the accuracy as-

number of nodes	linear triangles			quadratic triangles		
	$\varepsilon = 1$	$\varepsilon = 10^{-3}$	$\varepsilon = 10^{-6}$	$\varepsilon = 1$	$\varepsilon = 10^{-3}$	$\varepsilon = 10^{-6}$
$6 \times 6$	$6.0_{10^{-4}}$	$5.2_{10^{-3}}$	$5.9_{10^{-3}}$			
$11 \times 11$	$1.6_{10^{-4}}$	$1.6_{10^{-3}}$	$2.0_{10^{-3}}$	$1.6_{10^{-5}}$	$2.8_{10^{-4}}$	$7.6_{10^{-5}}$
$21 \times 21$	$4.0_{10^{-5}}$	$4.2_{10^{-4}}$	$5.5_{10^{-4}}$	$1.1_{10^{-6}}$	$1.3_{10^{-4}}$	$1.3_{10^{-5}}$

Table 3.2: Error in max-norm of convection-diffusion problem (3.22a -3.22c) for various values of  $\varepsilon$  Solution by SUPG. Linear and quadratic triangles.

pects the SPUG method has an another important advantage. The use of upwind makes the matrices to be solved more diagonally dominant. As a consequence iterative matrix solvers

will converge much faster than for SGA. This will be demonstrated in Paragraph 3.6. Finally we show some results of classical benchmark problems to investigate the behavior of various schemes.

### 3.6 Some classical benchmark problems for convection-diffusion solvers

In this section we shall investigate the performance of the standard Galerkin approach as well as the streamline upwind Petrov Galerkin method for some benchmark problems.

First we consider the rotating cone problem introduced in Section 3.4. The solution by SGA is plotted in Figure 3.4. Figure 3.7 shows the lines of equal concentration produced by SUPG and Figure 3.8 the corresponding three-dimensional plot. Exactly the same mesh as for the central scheme is used. These pictures show a large qualitative improvement of the accuracy compared to SGA. Not only the accuracy of the solution is enlarged considerably, also the

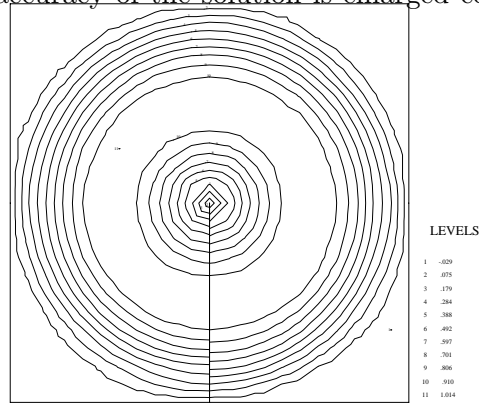


Figure 3.7: Equi-concentration lines for rotating cone problem computed by SUPG.

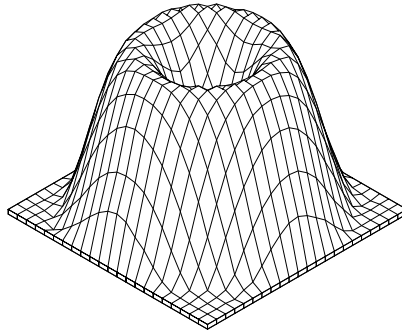


Figure 3.8: 3D plot of concentration for rotating cone problem computed by SUPG

condition with respect to linear solvers. All previous pictures have been created by using a direct linear solver. However, for large problems an iterative solver is much more attractive. Numerical computations show that upwinding has a very important effect on the number of iterations necessary for reaching a certain level of accuracy. Table 3.3 shows the number of iterations required to solve the rotating cone problem for  $\epsilon = 10^{-3}$  and  $\epsilon = 10^{-6}$ . In this table SGA and SUPG are compared for various mesh sizes. From this table it is clear

that for small values of  $\epsilon$  the SUPG method is superior to SGA with respect to iterative solvers.

$\epsilon$	number of nodes	SGA		SUPG	
		accuracy $10^{-3}$	accuracy $10^{-6}$	accuracy $10^{-3}$	accuracy $10^{-6}$
$10^{-3}$	$21 \times 21$	15	19	6	9
	$41 \times 41$	17	21	6	9
	$81 \times 81$	31	38	27	32
$10^{-6}$	$21 \times 21$	-	-	9	12
	$41 \times 41$	-	-	13	17
	$81 \times 81$	-	-	24	32

Table 3.3: Number of iterations by a preconditioned CGS solver for the solution of the rotating cone problem of Section 3.4. SGA and SUPG. A - in the table means that no convergence was possible.

As last benchmark problem we consider is a time-dependent one dimensional convection-diffusion equation given by:

$$\frac{\partial c}{\partial t} + u \frac{\partial c}{\partial x} = D \frac{\partial^2 c}{\partial x^2} \quad 0 \leq x \leq 1, \quad (3.45)$$

$$\begin{aligned} c(x, 0) &= \sin \pi \frac{x-a}{b-a} & a \leq x \leq b, \\ c(x, 0) &= 0 & \text{elsewhere,} \\ c(0, t) &= c(1, t) = 0, \end{aligned} \quad (3.46)$$

$$u = 1, \quad D = 0.002, \quad a = 0.2 \quad \text{and} \quad b = 0.4. \quad (3.47)$$

This benchmark problem has been solved by SGA with and without lumping of the mass matrix and SUPG. Figures 3.9-3.11 shows the results of the various methods. These figures show that lumping drastically decreases the accuracy of the numerical solution. Furthermore for this moderate Peclet number, the standard Galerkin method performs a little bit better than SUPG.

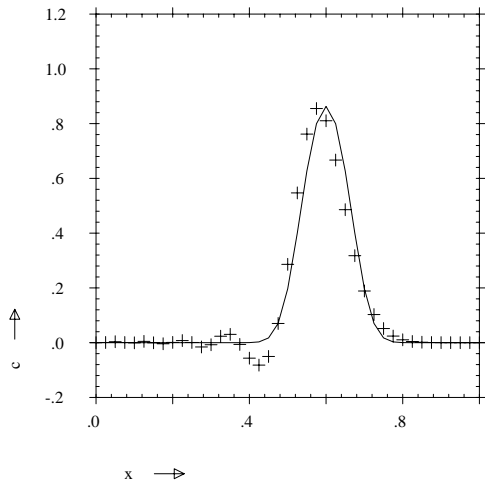


Figure 3.9:  
 SGA applied to (3.40)-(3.42)  
 40 linear elements, lumped mass matrix  
 – exact solution, + numerical solution

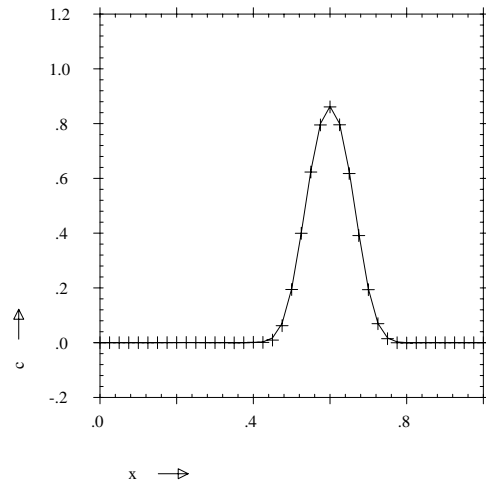


Figure 3.10:  
 SGA applied to (3.40)-(3.42)  
 40 linear elements, consistent mass matrix  
 – exact solution, + numerical solution

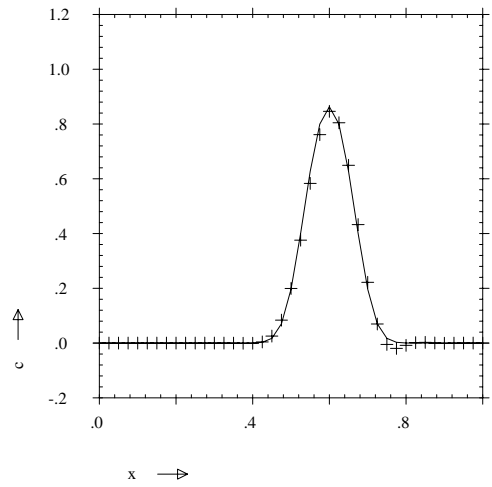


Figure 3.9: SUPG with stationary upwind parameter applied to (3.40)-(3.42), 40 linear elements: - exact solution, + numerical solution

## 4 Discretization of the incompressible Navier-Stokes equations by standard Galerkin

### 4.1 The basic equations of fluid dynamics

In this chapter we shall consider fluids with the following properties:

- The medium is incompressible,
- The medium has a Newtonian character,
- The medium properties are temperature independent and uniform,
- The flow is laminar.

For a three-dimensional flow field the basic equations of fluid flow under the above restrictions, can be written as:

The Continuity equation

$$\operatorname{div} \mathbf{u} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} = 0 . \quad (4.1)$$

The Navier-Stokes equations

$$\rho \left( \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} \right) - \operatorname{div} \boldsymbol{\sigma} = \rho \mathbf{f}, \quad (4.2)$$

in which  $\mathbf{u} = (u_1, u_2, u_3)^T$  denotes the velocity vector,  $\rho$  the density of the fluid,  $\mathbf{f} = (f_1, f_2, f_3)$  the body force per unit of mass, and  $\boldsymbol{\sigma}$  the stress tensor.

Component-wise (4.2) reads:

$$\rho \left( \frac{\partial u_i}{\partial t} + u_1 \frac{\partial u_i}{\partial x_1} + u_2 \frac{\partial u_i}{\partial x_2} + u_3 \frac{\partial u_i}{\partial x_3} \right) - \left( \frac{\partial \sigma_{i1}}{\partial x_1} + \frac{\partial \sigma_{i2}}{\partial x_2} + \frac{\partial \sigma_{i3}}{\partial x_3} \right) = \rho f_i, \quad (i = 1, 2, 3) . \quad (4.3)$$

For an incompressible and isotropic medium the stress terms  $\boldsymbol{\sigma}$  can be written as

$$\boldsymbol{\sigma} = -p \mathbf{I} + \mathbf{d} = -p \mathbf{I} + 2\mu \mathbf{e} , \quad (4.4)$$

where  $p$  denotes the pressure,

$\mathbf{I}$  the unit tensor

$\mathbf{e}$  the rate of strain tensor,

$\mathbf{d}$  the deviatoric stress tensor and

$\mu$  the viscosity of the fluid.

The components  $e_{ij}$  of the tensor  $\mathbf{e}$  are defined by

$$e_{ij} = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right), \quad (4.5)$$

so

$$\sigma_{ij} = -p\delta_{ij} + \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) . \quad (4.6)$$

If  $\mu$  is constant it is possible to simplify the expression (4.2) by substitution of the incompressibility condition (4.1) to

$$\rho \left( \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} \right) - \mu \Delta \mathbf{u} + \nabla p = \rho \mathbf{f}, \quad (4.7)$$

however, we shall prefer expression (4.2) because boundary conditions will be implemented more easily in (4.2) than in (4.7).

Equation (4.2) can be made dimensionless by the introduction of the Reynolds number  $Re$  defined by

$$Re = \frac{\rho U L}{\mu}, \quad (4.8)$$

where  $U$  is some characteristic velocity and  $L$  a characteristic length. Substitution of (4.8) into (4.2), (4.4) gives

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} - \operatorname{div} \boldsymbol{\sigma} = \mathbf{f} \quad (4.9a)$$

$$\boldsymbol{\sigma} = -p \mathbf{I} + \frac{2}{Re} \mathbf{e} \quad (4.9b)$$

provided  $\rho$  does not depend on the space coordinates.

## 4.2 Initial and boundary conditions

In order to solve the equations (4.1), (4.2), it is necessary to prescribe both initial and boundary conditions. Since only first derivatives of time are present in (4.2), it is sufficient to prescribe the initial velocity field at  $t = 0$ . Of course this velocity field must satisfy the incompressibility condition (4.1)

Since (4.2) is a system of second order differential equations in space, it is necessary to prescribe boundary conditions for each velocity component on the complete boundary of the domain. However, at high Reynolds numbers the convective terms dominate the stress tensor and as a consequence the boundary conditions at outflow must be such that they restrict the flow as little as possible.

The continuity equation and the pressure play a very special role in the incompressible Navier-Stokes equations. In fact there is a strong relation between both. It can be shown (Ladyshenskaya, 1969), that for incompressible flows no explicit boundary conditions for the pressure must be given. Usually boundary conditions for the pressure are implicitly given by prescribing the normal stress.

The following types of boundary conditions are commonly used for the two-dimensional incompressible Navier-Stokes equations (the extension to  $\mathbb{R}^3$  is straight forward):

$$1 \quad \mathbf{u} \text{ given (Dirichlet boundary condition),} \quad (4.10a)$$

$$2 \quad u_n \text{ and } \sigma^{nt} \text{ given,} \quad (4.10b)$$

$$3 \quad u_t \text{ and } \sigma^{nn} \text{ given,} \quad (4.10c)$$

$$4 \quad \sigma^{nt} \text{ and } \sigma^{nt} \text{ given,} \quad (4.10d)$$

where  $u_n$  denotes the normal component of the velocity on the boundary and  $u_t$  the tangential component.  $\sigma^{nn}$  ( $\mathbf{n} \cdot \boldsymbol{\sigma} \cdot \mathbf{n}$ ) denotes the normal component of the stress tensor on the boundary and  $\sigma^{nt}$  ( $\mathbf{n} \cdot \boldsymbol{\sigma} \cdot \mathbf{t}$ ) the tangential component.

Typical examples of these boundary conditions are:

- At fixed walls: no-slip condition  $\mathbf{u} = \mathbf{o}$ . This is an example of type (4.10b).
- At inflow the velocity profile given:  $\mathbf{u} = \mathbf{g}$ . This is also an example of type (4.10b). Typical inflow profiles are  $u_t = 0$ ,  $u_n$  parabolic or  $u_t = 0$  and  $u_n$  constant.

At outflow one may prescribe the velocity. However, for convection dominated flows, such a boundary condition may lead to wiggles due to inaccuracies of the boundary conditions. Less restrictive boundary conditions are for example  $u_t = 0$  and  $\sigma^{nn} = 0$  or  $\sigma^{nt} = 0$  and  $\sigma^{nn} = 0$ . The first one ( $u_t = 0$ ,  $\sigma^{nn} = 0$ ) prescribes a parallel outflow with zero normal stress. From (4.6) it can be derived that

$$\sigma^{nn} = -p + \frac{2}{Re} \frac{\partial u_n}{\partial n}, \quad (4.11)$$

$$\text{and } \sigma^{nt} = \frac{1}{Re} \left( \frac{\partial u_n}{\partial t} + \frac{\partial u_t}{\partial n} \right). \quad (4.12)$$

As a consequence for high Reynolds numbers  $\sigma^{nn}$  is approximately equal to  $-p$ . So  $\sigma^{nn} = 0$  implies that implicitly  $p$  is set equal to zero.

The boundary condition  $u_t = 0$ ,  $\sigma^{nn} = 0$  is correct for channel flow. The boundary condition  $\sigma^{nt} = 0$ ,  $\sigma^{nn} = 0$  is in general not correct. For a channel flow, in which case we have a parabolic velocity profile,  $\frac{\partial u_n}{\partial t}$  is linear and hence  $\sigma^{nt} \neq 0$ . However, in practical situations we usually do not have a channel flow and it is very hard to formulate correct boundary conditions at outflow. Vosse (1987) has shown that the boundary condition  $\sigma^{nt} = 0$ ,  $\sigma^{nn} = 0$ , although incorrect, may be a good choice in numerical computations.

He performed some experiments in the flow over a backward facing step. Figure 4.1 shows the streamlines for  $Re = 150$ , and the length of the channel after the step large enough. In this case the flow at the end may be considered as a channel flow and the boundary condition

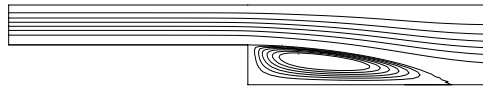


Figure 4.1: Streamlines in backward step. Length of channel is  $44H$ , where  $H$  is the step height. Outflow boundary conditions  $\sigma^{nn} = 0$ ,  $u_t = 0$ . Only the part  $(-6H, 6H)$  is plotted.

$u_t = 0$ ,  $\sigma^{nt} = 0$  is a good approximation. However, if we make the length of the channel such that the outflow boundary intersects the recirculation zone, it is impossible to define correct boundary conditions. Figure 4.2 shows the results of computations with the boundary conditions  $\sigma^{nt} = 0$ ,  $\sigma^{nn} = 0$ . The agreement with the computations in the long channel is remarkably good.

For a free surface we have the condition that there is no flow through the surface and that the tangential stress is equal to zero. In that case we use the boundary condition  $u_n = 0$ ,  $\sigma^{nt} = 0$ .

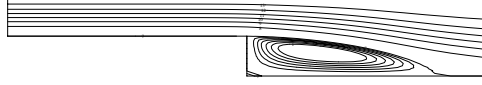


Figure 4.2: Streamlines in backward step. Length of channel is  $12H$ . Outflow boundary conditions  $\sigma^{nn} = 0$ ,  $\sigma^{nt} = 0$ .

One can show that the equations (4.1), (4.2) with a given initial flow field and combinations of boundary conditions of type (4.10b-4.10d) have a unique solution. There is, however, one exception. If we solve the stationary incompressible Navier-Stokes equations with the velocity prescribed on the complete boundary (actually each combination in which the normal velocity component is prescribed), the velocity is unique, but the pressure is fixed up to an additive constant.

### 4.3 Axisymmetric flow

Since in general three-dimensional flow computations are very expensive, one usually tries to reduce the dimension by considering symmetry in the flow or neglect flow in a certain direction. The last possibility results in two-dimensional flow, such as channel flow. If we use cylinder symmetry the flow reduces to so-called axisymmetric flow.

In such a case the Navier-Stokes equations and the velocity vector have to be transformed to a cylindrical co-ordinate system with co-ordinates  $r, \varphi$  and  $z$  and velocity components  $u_r, u_\varphi$  and  $u_z$ . In an axisymmetric flow the variation in  $\varphi$ -direction is zero and all  $\varphi$ -derivatives may be neglected. Whether the  $u_\varphi$  component may be neglected depends on the flow. In a rotating flow  $u_\varphi$  is not equal to zero and we have in that case three velocity unknowns, although we have only two directions.

The incompressible Navier-Stokes equations in cylinder co-ordinates are still given by the expressions (4.1) and (4.2). However, the operators divergence and gradient as well as the stress tensor get a different shape:

$$\nabla v = \left( \frac{\partial v}{\partial r}, \frac{1}{r} \frac{\partial v}{\partial \varphi}, \frac{\partial v}{\partial z} \right)^T, \quad (4.13a)$$

$$\text{div } \mathbf{u} = \frac{1}{r} \left( \frac{\partial r u_r}{\partial r} + \frac{\partial u_\varphi}{\partial \varphi} + \frac{\partial r u_z}{\partial z} \right) = 0, \quad (4.13b)$$

$$\sigma_{rr} = -p + 2\mu \frac{\partial u_r}{\partial r}, \quad \sigma_{\varphi\varphi} = -p + 2\mu \left( \frac{u_r}{r} + \frac{1}{r} \frac{\partial u_\varphi}{\partial \varphi} \right), \quad (4.13c)$$

$$\sigma_{zz} = -p + 2\mu \frac{\partial u_z}{\partial z}, \quad \sigma_{r\varphi} = \sigma_{\varphi r} = \mu \left( r \frac{\partial}{\partial r} \left( \frac{u_\varphi}{r} \right) + \frac{1}{r} \frac{\partial u_r}{\partial \varphi} \right),$$

$$\sigma_{\varphi z} = \sigma_{z\varphi} = \mu \left( \frac{1}{r} \frac{\partial u_z}{\partial \varphi} + \frac{\partial u_\varphi}{\partial z} \right), \quad \sigma_{rz} = \sigma_{zr} = \mu \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right).$$

Note that in these expressions the term  $1/r$  frequently occurs. As a consequence one has to be careful in the numerical computations at  $r = 0$ . At the symmetry axis  $r = 0$ , we need extra boundary conditions, the so-called symmetry conditions. One immediately verifies that these symmetry conditions are given by:

$$u_r = 0 \quad , \quad \frac{\partial u_z}{\partial r} = 0 \quad , \quad u_\varphi = 0 \quad \text{at } r = 0 \quad , \quad (4.14)$$

or translated to stresses:

$$u_r = 0 \quad , \quad u_\varphi = 0 \quad \text{and} \quad \sigma^{nt} = 0 \quad \text{at } r = 0 \quad . \quad (4.15)$$

#### 4.4 The weak formulation

In the next paragraph we shall derive the standard Galerkin equation for the incompressible Navier-Stokes equations. First we shall derive the weak formulation. In order to consider the four boundary conditions (4.10b- 4.10d), we shall assume that the boundary consists of four parts each with one of the boundary conditions (4.10b-4.10d). Furthermore we shall restrict ourselves in this chapter to stationary problems. The instationary case will be treated in Chapter 7.

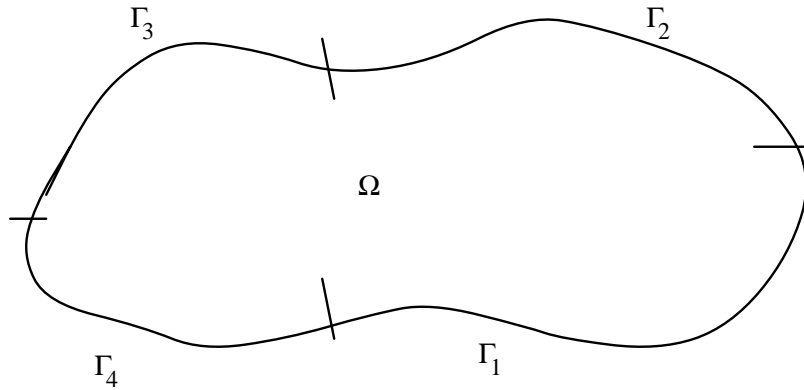


Figure 4.3: Artificial example with region  $\Omega$  and boundaries  $\Gamma_1, \Gamma_2, \Gamma_3$  and  $\Gamma_4$ .

Furthermore we restrict ourselves for the moment to the two-dimensional case. Figure 4.3 shows an artificial example of a region  $\Omega$  with four boundaries  $\Gamma_1$  to  $\Gamma_4$ . On each of these boundaries we have a different type of boundary condition. The formulation of our example is now: For  $\mathbf{x} \in \Omega$  solve  $\mathbf{u}$  satisfying

$$\operatorname{div} \mathbf{u} = 0 \quad , \quad (4.16a)$$

$$-\operatorname{div} \boldsymbol{\sigma} + \rho(\mathbf{u} \cdot \nabla \mathbf{u}) = \rho \mathbf{f} \quad , \quad (4.16b)$$

$$\boldsymbol{\sigma}_{ij} = -p\delta_{ij} + \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \quad , \quad (4.16c)$$

$$\mathbf{u} = \mathbf{g}_1 \quad \text{for } \mathbf{x} \in \Gamma_1 \quad , \quad (4.16d)$$

$$u_n = g_2, \quad \sigma^{nt} = g_3 \quad , \quad \mathbf{x} \in \Gamma_2 \quad , \quad (4.16e)$$

$$u_t = g_4, \quad \sigma^{nn} = g_5 \quad , \quad \mathbf{x} \in \Gamma_3 \quad , \quad (4.16f)$$

$$\sigma^{nt} = g_6 \quad \sigma^{nn} = g_7 \quad , \quad \mathbf{x} \in \Gamma_4 \quad . \quad (4.16g)$$

In order to derive the weak formulation, equation (4.16a-4.16b) must be multiplied by test functions. First equation (4.16a) is multiplied by a test function  $q$  resulting in

$$\int_{\Omega} q \operatorname{div} \mathbf{u} \, d\Omega = 0 \quad (4.17)$$

The momentum equations (4.16b) consist of two equations, which may be each multiplied by separate test functions  $v_1$  and  $v_2$ . If we define  $\mathbf{v} = (v_1, v_2)^T$  these equations can be combined to:

$$\int_{\Omega} (-\operatorname{div} \boldsymbol{\sigma} + \rho(\mathbf{u} \cdot \nabla \mathbf{u})) \cdot \mathbf{v} \, d\Omega = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\Omega \quad (4.18)$$

Choosing  $v_2$  respectively  $v_1$  equal to zero gives us the original weak formulation for each of the equations.

The first term in (4.18) may be further reduced by applying integration by parts (Gauss theorem) to

$$\int_{\Omega} (-\operatorname{div} \boldsymbol{\sigma}) \cdot \mathbf{v} \, d\Omega = \int_{\Omega} \boldsymbol{\sigma} \cdot \nabla \mathbf{v} \, d\Omega - \int_{\Gamma} (v_n \sigma^{nn} + v_t \sigma^{nt}) d\Gamma, \quad (4.19)$$

where  $\Gamma$  denotes the boundary of  $\Omega$ ,  $v_n$  the component of  $\mathbf{v}$  in the normal direction and  $v_t$  in the tangential direction. For a derivation of formula (4.19) we refer to Appendix A.

In order to apply the boundary conditions (4.16d-4.16g), the boundary integral over  $\Gamma$  is split into 4 parts  $\Gamma_1$ , to  $\Gamma_4$ .

On  $\Gamma_1$  we have a prescribed velocity and hence the test function  $\mathbf{v}$  is chosen equal to zero. On boundary  $\Gamma_2$   $u_n$  is prescribed and so  $v_n$  is chosen equal to zero, and on boundary  $\Gamma_3$   $u_t$  is prescribed and  $v_t$  is set equal to zero.

If we furthermore substitute relation (4.4) into (4.19), the first term of (4.18) can be written as:

$$\begin{aligned} \int_{\Omega} -(\operatorname{div} \boldsymbol{\sigma}) \cdot \mathbf{v} \, d\Omega &= \int_{\Omega} 2\mu \mathbf{e} \cdot \nabla \mathbf{v} \, d\Omega - \int_{\Omega} p \operatorname{div} \mathbf{v} \, d\Omega \\ &- \int_{\Gamma_2} g_3 v_t \, d\Gamma - \int_{\Gamma_3} g_5 v_n \, d\Gamma - \int_{\Gamma_4} g_6 v_t + g_7 v_n \, d\Gamma \quad (4.20) \end{aligned}$$

Combinations of all these results leads to the weak formulation of the Navier-Stokes equations (4.16a-4.16g):

Find  $\mathbf{u}$ ,  $p$  with

$$\mathbf{u} = \mathbf{g}_1 \text{ at } \Gamma_1, \quad u_n = g_2 \text{ at } \Gamma_2, \quad u_t = g_4 \text{ at } \Gamma_3,$$

such that

$$\int_{\Omega} q \operatorname{div} \mathbf{u} \, d\Omega = 0, \quad (4.21)$$

$$\begin{aligned} \int_{\Omega} 2\mu \mathbf{e} \cdot \nabla \mathbf{v} \, d\Omega + \int_{\Omega} \rho(\mathbf{u} \cdot \nabla \mathbf{u}) \cdot \mathbf{v} \, d\Omega - \int_{\Omega} p \operatorname{div} \mathbf{v} \, d\Omega = \\ \int_{\Gamma_2} g_3 v_t \, d\Gamma + \int_{\Gamma_3} g_5 v_n \, d\Gamma + \int_{\Gamma_4} g_6 v_t + g_7 v_n \, d\Gamma + \int_{\Omega} \rho \mathbf{f} \cdot \mathbf{v} \, d\Omega, \quad (4.22) \end{aligned}$$

for all  $\mathbf{v}$  such that  $\mathbf{v} = \mathbf{o}$  at  $\Gamma_1$ ,  $v_n = 0$  at  $\Gamma_2$  and  $v_t = 0$  at  $\Gamma_3$ , and  $\mathbf{e}$  given by (4.5).

We see that in the relations (4.21) and (4.22) no derivatives of  $p$  and  $q$  are necessary. Hence it is sufficient that  $p$  and  $q$  are integrable. With respect to  $\mathbf{u}$  and  $\mathbf{v}$ , first derivatives are required and hence not only  $\mathbf{u}$  and  $\mathbf{v}$  but also their first derivatives must be integrable. As a consequence we do not need continuity of  $p$  and  $q$  in the Galerkin formulation, but the functions  $\mathbf{u}$  and  $\mathbf{v}$  must be continuous over the element boundaries.

The weak formulation (4.21-4.22) shows a strong relation between  $\mathbf{u}$  and  $\mathbf{v}$ , as well as between  $p$  and  $q$ . If we for example demand that both  $\mathbf{u}$  and  $\mathbf{v}$  are divergence free, then the first equation (4.21) vanishes and the pressure disappears from (4.22). Indeed in all theoretical investigations with respect to the weak form of the Navier-Stokes equations,  $p$  and  $q$  are taken from the same space and  $\mathbf{u}$  and  $\mathbf{v}$  are taken from the same space. This observation motivates the choice of the basis functions in the standard Galerkin method.

#### 4.5 The standard Galerkin method

In the standard Galerkin method we define two types of basis functions, basis function  $\Psi_i(\mathbf{x})$  corresponding to the pressure and functions  $\varphi_i(\mathbf{x})$  corresponding to the velocity components. We may combine the velocity basis functions into vector form by

$$\boldsymbol{\varphi}_{i1}(\mathbf{x}) = \begin{pmatrix} \varphi_i(\mathbf{x}) \\ 0 \end{pmatrix}, \quad \boldsymbol{\varphi}_{i2}(\mathbf{x}) = \begin{pmatrix} 0 \\ \varphi_i(\mathbf{x}) \end{pmatrix}. \quad (4.23)$$

Now the approximation of  $\mathbf{u}$  and  $p$  will be defined by

$$p_h = \sum_{j=1}^m p_j \Psi_j(\mathbf{x}), \quad (4.24)$$

$$\mathbf{u}_h = \sum_{j=1}^n u_{1j} \boldsymbol{\varphi}_{j1}(\mathbf{x}) + u_{2j} \boldsymbol{\varphi}_{j2}(\mathbf{x}) = \sum_{j=1}^{2n} u_j \boldsymbol{\varphi}_j(\mathbf{x}). \quad (4.25)$$

In (4.25)  $u_j$  is defined by  $u_j = u_{1j}$ , ( $j = 1(1)n$ ),  $u_{j+n} = u_{2j}$ , ( $j = 1(1)n$ ) and  $\boldsymbol{\varphi}_j$  in the same way. For simplicity the summation has been carried out over all degrees of freedom including the prescribed ones at the boundary. The test functions, however, must only be coupled to the free degrees of freedom. Mark that the number of basis functions  $\boldsymbol{\varphi}_i(\mathbf{x})$  and  $\Psi_i(\mathbf{x})$  do not have to be the same, nor that these basis functions must have the same shape. In fact, in most practical applications  $\boldsymbol{\varphi}_i(\mathbf{x})$  and  $\Psi_i(\mathbf{x})$  are chosen differently.

In order to get the standard Galerkin formulation we substitute  $\mathbf{v} = \boldsymbol{\varphi}_i(\mathbf{x})$ ,  $q = \Psi_i(\mathbf{x})$  into the weak formulation (4.21-4.22).

In this way we get:

Find  $p_n$  and  $\mathbf{u}_n$  defined by (4.24),(4.25) such that

$$\int_{\Omega} \Psi_i \operatorname{div} \mathbf{u}_h \, d\Omega = 0, \quad i = 1(1)m, \quad (4.26)$$

and

$$\begin{aligned}
& \int_{\Omega} 2\mu (\mathbf{e}_h \cdot \nabla \boldsymbol{\varphi}_i) d\Omega + \int_{\Omega} \rho(\mathbf{u}_h \cdot \nabla \mathbf{u}_h) \cdot \boldsymbol{\varphi}_i d\Omega \\
& - \int_{\Omega} p_h \operatorname{div} \boldsymbol{\varphi}_i d\Omega = \int_{\Gamma_2} g_3(\boldsymbol{\varphi}_i \cdot \mathbf{t}) d\Gamma + \int_{\Gamma_3} g_5(\boldsymbol{\varphi}_i \cdot \mathbf{n}) d\Gamma \\
& + \int_{\Gamma_4} g_6(\boldsymbol{\varphi}_i \cdot \mathbf{t}) + g_7(\boldsymbol{\varphi}_i \cdot \mathbf{n}) d\Gamma + \int_{\Omega} \rho \mathbf{f} \cdot \boldsymbol{\varphi}_i d\Omega
\end{aligned} \tag{4.27}$$

where  $\mathbf{e}_h$  is given by (4.5);  $u$  replaced by  $\mathbf{u}_h$

$i$  in (4.27) must be taken for all free degrees of freedom  $u_i$ .

Expression (4.27) may be easily evaluated as long as  $\mathbf{n}$  or  $\mathbf{t}$  on the boundaries  $\Gamma_2$ ,  $\Gamma_3$  and  $\Gamma_4$  are in the direction of the co-ordinate axis. If they are not in that direction it is necessary to transform the unknowns on the boundary locally such that they are expanded into normal and tangential direction. The technique of local transformations is described in Zienkiewicz and Taylor (1989).

The finite element method may be used to construct the basis functions  $\boldsymbol{\varphi}_i$  and  $\Psi_i$ , in the same way as for the potential problem in Chapter 1. Once the basis functions are known, the integrals (4.26) and (4.27) may be evaluated element-wise. Finally we arrive at a system of  $m + 2n - n_p$  non-linear equations with  $m + 2n - n_p$  unknowns, where  $n_p$  denotes the number of prescribed boundary values, and  $m$  and  $N$  are defined in (4.24), (4.25).

Formally the system of equations can be written as

$$\mathbf{S}\mathbf{U} + \mathbf{N}(\mathbf{U}) - \mathbf{L}^T \mathbf{P} = \mathbf{F} \tag{4.28a}$$

$$\mathbf{L}\mathbf{U} = \mathbf{o} \tag{4.28b}$$

where  $\mathbf{U}$  denotes the vector of unknowns  $u_{1i}$  and  $u_{2i}$ ,  $\mathbf{P}$  denotes the vector of unknowns  $p_i$ ,  $\mathbf{S}\mathbf{U}$  denotes the discretization of the viscous terms,  $\mathbf{N}(\mathbf{U})$  the discretization of the non-linear convective terms,  $\mathbf{L}\mathbf{U}$  denotes the discretization of the divergence of  $\mathbf{u}$  and  $-\mathbf{L}^T \mathbf{P}$  the discretization of the gradient of  $p$ . The right-hand side  $\mathbf{F}$  contains all contributions of the source term, the boundary integral as well as the contribution of the prescribed boundary conditions.

The solution of the system of equations (4.28a-4.28b) introduces two difficulties. Firstly the equations are non-linear and as a consequence some iterative solution procedure is necessary. Secondly equation (4.28b) does not contain the unknown pressure  $\mathbf{P}$ . The last aspect introduces a number of extra complications which will be treated in Paragraph 4.7. The non-linear iterative procedure will be the subject of Paragraph 4.6.

## 4.6 Treatment of the non-linear terms

In order to solve the system of non-linear equations, an iterative procedure is necessary. In general such a procedure consists of the following steps:

make an initial estimation

while (not converged) do

linearize the non-linear equations based on the previous solution  
 solve the resulting system of linear equations

Examples of such methods are Newton methods, quasi-Newton methods, and Picard type methods.

In order to derive the iterative method one may proceed in two ways. Firstly one can apply the method to equations (4.28a-4.28b), which is the classical approach. An alternative is to linearize the non-linear differential equations first and then to discretize the resulting linear equation. Sometimes both approaches are identical. The last approach is conceptually easier than the first one and will therefore be applied in this paragraph. Since it is the only non-linear term in equation (4.16a-4.16g) we only consider the convective terms.

Suppose we have computed the solution  $\mathbf{u}^k$  at a preceding iteration level  $k$ . We write this solution as  $\mathbf{u}^k$ . First we shall derive the Newton linearization. To that end we define  $f(\mathbf{u}, \nabla \mathbf{u})$  as

$$f(\mathbf{u}, \nabla \mathbf{u}) = \mathbf{u} \cdot \nabla \mathbf{u} . \quad (4.29)$$

Taylor-series expansion of (4.29) gives

$$\begin{aligned} f^{k+1}(\mathbf{u}, \nabla \mathbf{u}) &= f^k(\mathbf{u}^k, \nabla \mathbf{u}^k) + (\mathbf{u}^{k+1} - \mathbf{u}^k) \cdot \frac{\partial f^k}{\partial \mathbf{u}} \\ &\quad + \nabla (\mathbf{u}^{k+1} - \mathbf{u}^k) \cdot \frac{\partial f^k}{\partial \nabla \mathbf{u}} + O(\mathbf{u}^{k+1} - \mathbf{u}^k)^2 . \end{aligned} \quad (4.30)$$

Neglecting the quadratic terms and substitution of (4.29) gives

$$\begin{aligned} \mathbf{u}^{k+1} \cdot \nabla \mathbf{u}^{k+1} &\approx \mathbf{u}^k \cdot \nabla \mathbf{u}^k + (\mathbf{u}^{k+1} - \mathbf{u}^k) \cdot \nabla \mathbf{u}^k + \nabla (\mathbf{u}^{k+1} - \mathbf{u}^k) \cdot \mathbf{u}^k \\ &= \mathbf{u}^{k+1} \cdot \nabla \mathbf{u}^k + \mathbf{u}^k \cdot \nabla \mathbf{u}^{k+1} - \mathbf{u}^k \cdot \nabla \mathbf{u}^k . \end{aligned} \quad (4.31)$$

(4.31) forms the standard Newton linearization. Alternative linearization are constructed by the so-called Picard iteration methods in which one or both terms in (4.29) are taken at the old level.

Hence:

$$(\mathbf{u} \cdot \nabla \mathbf{u})^{k+1} \simeq \mathbf{u}^{k+1} \cdot \nabla \mathbf{u}^k , \quad (4.32)$$

$$(\mathbf{u} \cdot \nabla \mathbf{u})^{k+1} \simeq \mathbf{u}^k \cdot \nabla \mathbf{u}^{k+1} , \quad (4.33)$$

$$(\mathbf{u} \cdot \nabla \mathbf{u})^{k+1} \simeq \mathbf{u}^k \cdot \nabla \mathbf{u}^k . \quad (4.34)$$

Numerical experiments have shown that from these last three possibilities only (4.33) produces a good convergence. (4.31) shows that Newton is in fact a linear combination of (4.32)-(4.34). After linearization of the convective terms the standard Galerkin method may be applied, resulting in a system of linear equations.

An important question with respect to these iterative methods is, how to find a good initial estimate. It is well known that Newton's method converges fast (i.e. quadratically) as soon as the iteration is in the neighborhood of the final solution. However, if the distance between iteration and solution is too large, Newton may converge slowly or even diverge. The Picard iteration seems to have a larger convergence region, which means that this iteration does not

need the same accurate initial estimate, however, this method converges only linearly.

A possible strategy to converge to the final solution is the following:

- start with some initial guess,
- perform one step Picard iteration in order to approach the final solution, sometimes more than one step,
- use Newton iteration in the next steps.

An initial guess may be for example the solution of the Stokes problem, which is formed by the Navier-Stokes equations where the convective terms have been neglected. If the Reynolds number is too high it is possible that the distance between the solution of Stokes and Navier-Stokes is too large. In that case the solution of Navier-Stokes with a smaller Reynolds number might be a good choice. A process in which the Reynolds number is increased gradually is called a continuation method.

In general one may expect that the iteration process no longer converges as soon as the flow becomes instationary or turbulent.

#### 4.7 Necessary conditions for the elements

In Paragraph 4.5 it has been derived that the standard Galerkin method results in a system of non-linear equations of the form (4.28a-4.28b). After linearization this system can be written as

$$\begin{aligned} \mathbf{S}\mathbf{U} + \mathbf{N}(\mathbf{U}^k)\mathbf{U} - \mathbf{L}^T\mathbf{P} &= \mathbf{F} , \\ \mathbf{L}\mathbf{U} &= \mathbf{o} , \end{aligned} \tag{4.35}$$

where  $\mathbf{U}^k$  is the solution of the previous iteration.

In Section 4.5 it has already been pointed out that with respect to the velocity it is necessary that the approximation over the element-sides must be continuous, whereas the pressure approximation may be discontinuous over the element boundaries. However, there is another problem. The continuity equation, discretized as  $\mathbf{L}\mathbf{U} = \mathbf{o}$ , does contain only velocity unknowns. However, the number of rows in this equation is completely determined by the number of pressure unknowns. Suppose that there are more pressure unknowns than velocity unknowns. In that case equation (4.35) contains more rows than unknowns and we have either a dependent or inconsistent system of equations. In both cases the matrix to be solved is singular. So we have to demand that the number of pressure unknowns never exceeds the number of velocity unknowns. Since we want to solve the Navier-Stokes equations by finite element methods for various grid size, this demand should be valid independently of the number of elements. This demand restricts the number of applicable elements considerably. In order to satisfy this criterion, a general accepted rule is that the order of approximation of the pressure must be one lower than the order of approximation of the velocity. So if the velocity is approximated by a linear polynomial, then the pressure is approximated by a constant per element and so on.

Unfortunately this rule is not sufficient to guarantee that the number of pressure unknowns is not larger than the number of velocity unknowns independently of the number of elements. Consider for example the mesh in Figure 4.4a, based upon linear elements for the velocity

and constant elements for the pressure. For convenience the constant has been coupled to the centroid of the element. In this example the mesh contains 8 pressure nodes and 9 velocity

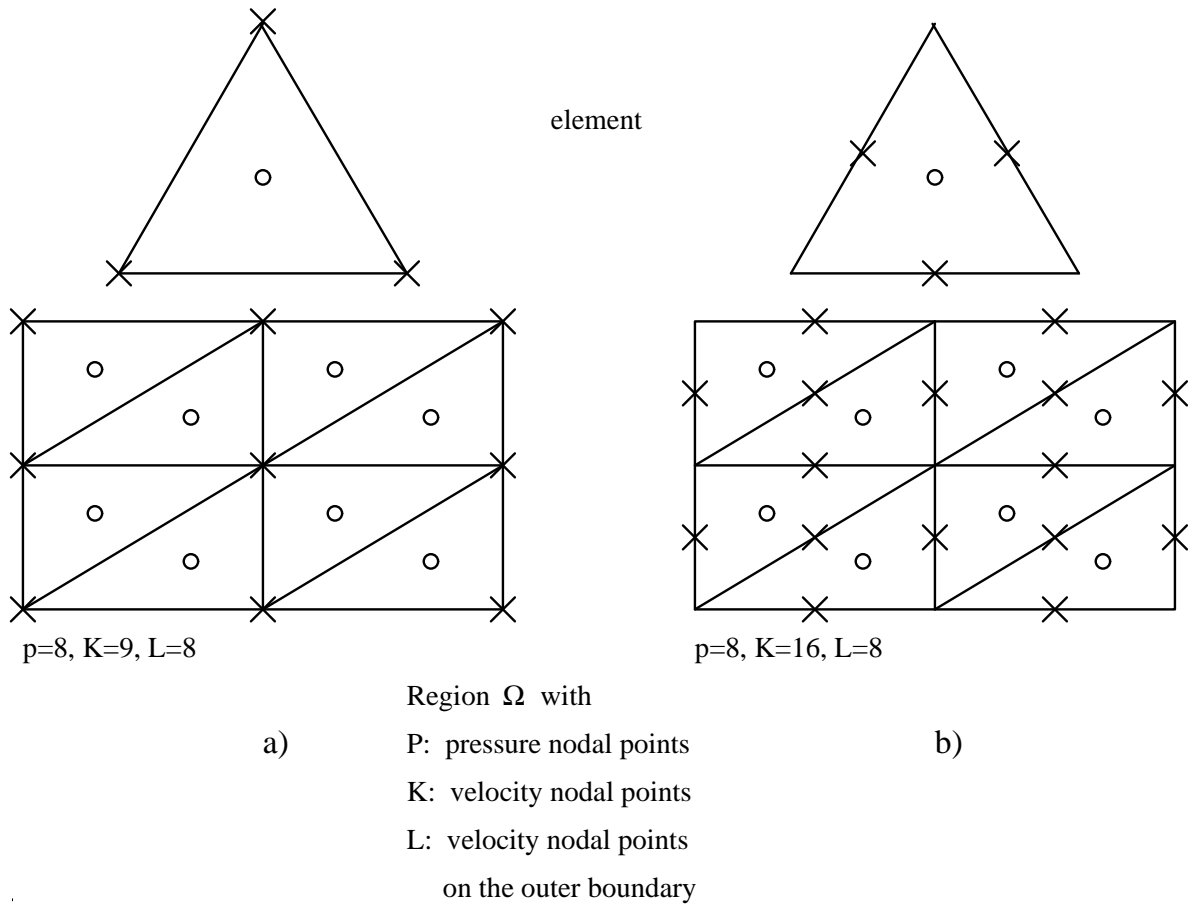


Figure 4.4: Triangular elements with three nodal points for the velocity ( $x$ ) and one nodal point for the pressure ( $0$ ): a) vertices are the velocity nodal points: conforming element, b) mid-points of the sides are the velocity nodal points: non-conforming element.

nodes. Suppose we have Dirichlet boundary conditions for the velocity, which means that all boundary velocities are prescribed. The pressure is unique except for an additive constant. To fix this constant one of the pressure unknowns is given. So finally we have 2 velocity unknowns and 7 pressure unknowns. Hence we have an example of a singular matrix. The corresponding element is not admissible. One might remark that if we add sufficient elements to the mesh eventually the number of velocity unknowns will be larger than the number of pressure unknowns. However, practical computations have shown that in that case still the matrix remains singular.

Figure 4.4b gives an example of an admissible element. The velocity unknowns are not positioned in the vertices of the triangle but in the midside points. The velocity approximation is linear but not continuous over the element boundaries. Such an element is called non-conforming and introduces for that reason extra problems with the approximation. However, with respect to the continuity equation the element satisfies the demand that there must be more velocity unknowns than pressure unknowns. A simple count shows that for the given mesh, the number of velocity unknowns is equal to 16 and the number of pressure unknowns

equal to 7 in the case of Dirichlet boundary conditions.

The derivation of the admissibility condition given above is rather ad-hoc and does not explain why an element is admissible. It just helps to identify non-admissible elements. In the literature, see for example Cuvelier et al (1986), an exact admissibility condition is derived. This condition is known under the name Brezzi-Babuška condition (or BB condition). However, the BB condition is rather abstract and in practice it is very difficult to verify whether the BB condition is satisfied or not. Fortin (1981) has given a simple method to check the BB condition on a number of elements.

The method is based on the following statement:

an element satisfies the BB condition, whenever, given a continuous differentiable vector field  $\mathbf{u}$ , one can explicitly build a discrete vector field  $\tilde{\mathbf{u}}$  such that:

$$\int_{\Omega} \Psi_i \operatorname{div} \tilde{\mathbf{u}} \, d\Omega = \int_{\Omega} \Psi_i \operatorname{div} \mathbf{u} \, d\Omega \quad \text{for all basis functions } \Psi_i . \quad (4.36)$$

In Cuvelier et al (1986) it is demonstrated how (4.36) can be checked for a number of elements.

Fortin (1981) formulates the following engineering statement with respect to the admissibility of elements.

Midside velocity points in two dimensions and centroid velocity points on surfaces in three dimensions make it possible to control the amount of flow through a side (2D) and through a surface (3D) of an element, without altering the amount of flow through other sides or surfaces. Hence such nodal points make it easier to satisfy the continuity equations.

In fact it is sufficient that the normal component of the velocity in these centroid points is available as unknowns.

In the literature frequently elements are used, that do not satisfy the BB condition. Such elements cannot be used with the standard Galerkin method, however the penalty function method (see Chapter 5), permits the use of these elements.

## 4.8 Examples of admissible elements

In this section we shall treat some of the admissible elements for two-dimensional applications. For a more thorough review as well as three-dimensional elements we refer to Cuvelier et al (1986) and Fortin (1981).

With respect to the types of elements that are applied we make a subdivision into two groups: elements with continuous pressure (The Taylor-Hood family) and elements with discontinuous pressure (The Crouzeix-Raviart family). We shall restrict ourselves to quadratic elements, since these elements are the most frequently used.

### The Taylor-Hood family

Taylor-Hood elements (Taylor and Hood 1973) are characterized by the fact that the pressure is continuous in the region  $\Omega$ . A typical example is the quadratic triangle of Figure 4.5. In this element the velocity is approximated by a quadratic polynomial and the pressure by a

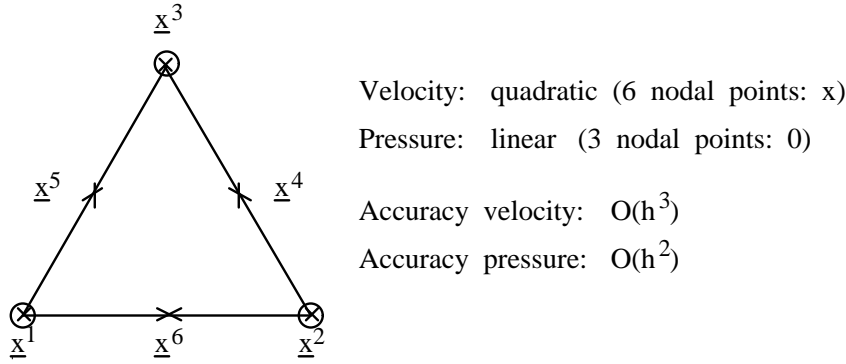


Figure 4.5: Taylor-Hood element ( $P_2 - P_1$ ).

linear polynomial. One can easily verify that both approximations are continuous over the element boundaries. It can be shown, Segal (1979), that this element is admissible if at least 3 elements are used. The quadrilateral counterpart of this triangle is given in Figure 4.6.

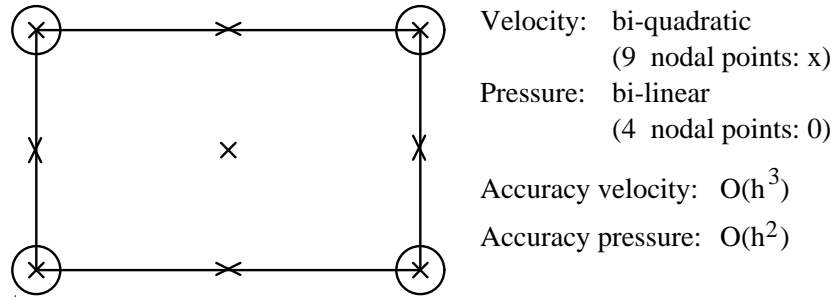


Figure 4.6: Taylor-Hood element ( $Q_2 - Q_1$ )

The Taylor-Hood family is very suitable for the standard Galerkin methods treated in this chapter. However, with respect to the special methods of Chapter 5 and 6, the discontinuous pressure elements are most favorable. For that reason we consider some of these elements.

#### The Crouzeix-Raviart family

These elements are characterized by a discontinuous pressure; discontinuous on element boundaries. For output purposes (printing, plotting etc.) these discontinuous pressures are averaged in vertices for all the adjoining elements, see Figure 4.7. We shall discuss some of the Crouzeix-Raviart elements. The most simple Crouzeix-Raviart element has already been mentioned in Section 4.7. It is the non-conforming linear triangle with constant pressure. Figure 4.8 shows this element again. Although this element has no practical significance, we shall use it to demonstrate how Fortin's translation (4.36) of the BB condition can be checked. To that end we explicitly create a vector  $\tilde{\mathbf{u}}$  such that (4.36) is satisfied, i.e.

$$\int_{\Omega^{e_k}} \operatorname{div} \mathbf{u} \, d\Omega = \int_{\Omega^{e_k}} \operatorname{div} \tilde{\mathbf{u}} \, d\Omega = \int_{\delta\Omega^{e_k}} \tilde{\mathbf{u}} \cdot \mathbf{n} \, d\Gamma \quad (4.37)$$

given the continuous vector field  $\mathbf{u}$ . In (4.37) we have used the fact that the pressure is constant per element, but discontinuously over the element boundary. As a consequence the

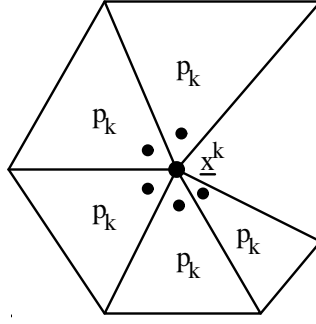
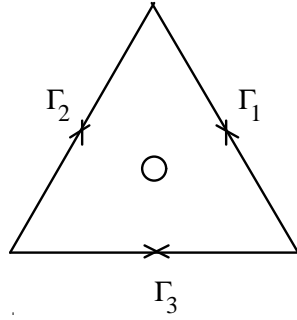


Figure 4.7: Averaging in each nodal point  $\underline{x}_k$  over all elements containing  $\underline{x}_k$  in order to get a continuous pressure for Crouzeix-Raviart elements.



Velocity: linear (3 nodal points:  $\mathbf{x}$ )  
 Pressure: Constant (1 nodal points: 0)  
 Accuracy velocity:  $O(h^2)$   
 Accuracy pressure:  $O(h)$

Figure 4.8: Crouzeix-Raviart element ( $p_1 - p_0$ )

basis functions  $\Psi_i(\mathbf{x})$  are defined by

$$\Psi_i(\mathbf{x}) = \begin{cases} 1 & \text{in element } e_i, \\ 0 & \text{in all other elements.} \end{cases} \quad (4.38)$$

If we define  $\tilde{\mathbf{u}}$  in the midside point of element  $e_i$  by

$$\int_{\Gamma_k} \mathbf{u} \, d\Gamma = \int_{\Gamma_k} \tilde{\mathbf{u}} \, d\Gamma = |\Gamma_k| \mathbf{u}_k, \quad (4.39)$$

with  $\Gamma_k$  the  $k$ -th side of  $e_i$ ,  $|\Gamma_k|$  the length of  $\Gamma_k$  and  $\mathbf{u}_k$  the velocity in the midside point of side  $\Gamma_k$ , we see immediately that (4.37) is satisfied. (4.39) implicitly defines  $\tilde{\mathbf{u}}$ . The definition is unique and does not introduce inconsistencies along adjacent elements since (4.39) is defined along one side of the elements only. The natural extension of the linear-constant element is the quadratic velocity, linear pressure element. The discontinuous linear pressure is defined by three parameters, for example the pressure and the gradient of the pressure in the centroid. Application of the counting mechanism demonstrated in Section 4.7 shows that this element cannot be admissible. In order to make it admissible it is necessary to introduce the velocity vector in the centroid as extra unknowns. In this way we get the so-called extended quadratic triangle of Figure 4.9.

The basis function for this element can be expressed in terms of the linear basis functions

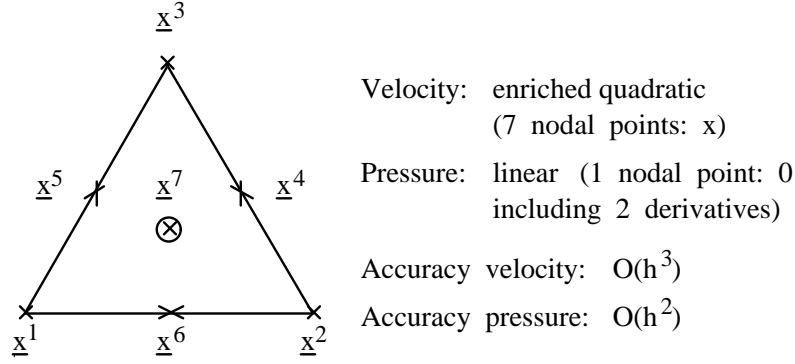


Figure 4.9: Crouzeix-Raviart element ( $P_2^+ - P_1$ )

$\Psi_i(\mathbf{x})$  for triangles, defined in Chapter 2:

$$\tilde{u}_i = \sum_{j=1}^7 u_{ij} \phi_j,$$

with  $\phi_j = \lambda_j(2\lambda_j - 1) + 3\lambda_1\lambda_2\lambda_3$  ,  $j = 1, 2, 3,$

$$\phi_4 = 4\lambda_2\lambda_3 - 12\lambda_1\lambda_2\lambda_3$$
 ,  $\phi_5 = 4\lambda_1\lambda_3 - 12\lambda_1\lambda_2\lambda_3$  ,  $\phi_6 = 4\lambda_1\lambda_2 - 12\lambda_1\lambda_3\lambda_3,$   
 $\phi_7 = 27\lambda_1\lambda_2\lambda_3,$  (4.40)

and  $\tilde{p} = p_7\Psi_1 + \frac{\partial p}{\partial x_1}(\mathbf{x}^7)\Psi_2 + \frac{\partial p}{\partial x_2}(\mathbf{x}^7)\Psi_3,$

with  $\Psi_1 = 1,$  (4.41)

$$\Psi_2 = x_1 - x_1^7,$$

$$\Psi_3 = x_2 - x_2^7.$$

The natural quadrilateral extension of this triangle is given in Figure 4.10.

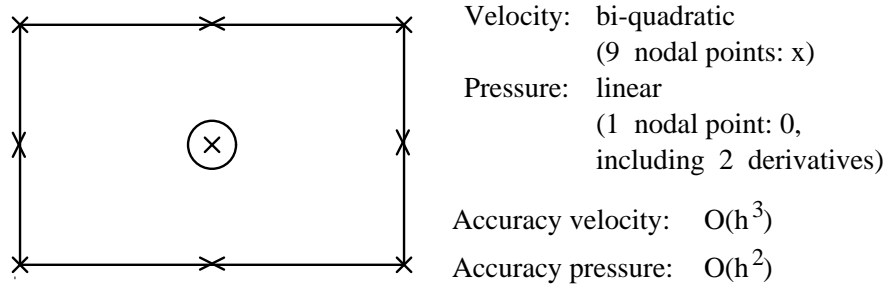


Figure 4.10: Crouzeix-Raviart quadrilateral ( $Q_2 - P_1$ )

### 4.9 Solution of the system of linear equations due to the discretization of Navier-Stokes

In Sections 4.5 and 4.6 the discretization of Navier-Stokes equations has been derived. It has been shown that in each step of the non-linear iteration process it is necessary to solve a system of linear equations of the shape

$$\mathbf{S}\mathbf{u} - \mathbf{L}^t \mathbf{p} = \mathbf{F} , \tag{4.42a}$$

$$Lu = 0. \tag{4.42b}$$

Here  $Su$  denotes the discretization of both the viscous terms and the linearized convective terms. If the unknowns are numbered in the sequence: first all velocity unknowns and then all pressure unknowns it is clear that the system of equations gets the shape as sketched in Figure 4.11 provided an optimal nodal point numbering is applied. Unfortunately this numbering (velocity first, pressure last) is far from optimal. The total profile is still very large.

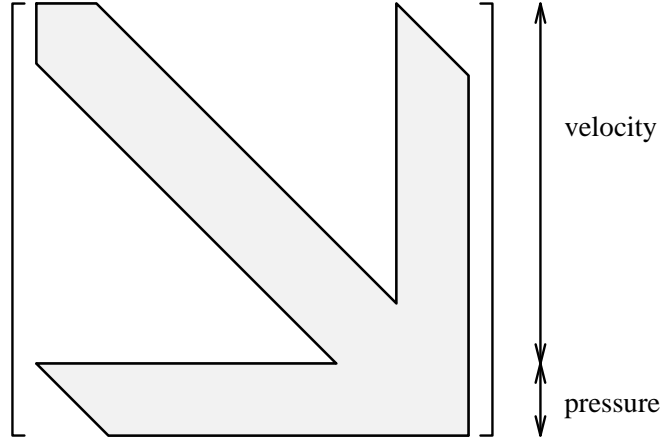


Figure 4.11: Profile of the large matrix.

A much smaller profile may be achieved if pressure and velocity unknowns are intermixed. Figure 4.12 shows a typical example of such a numbering.

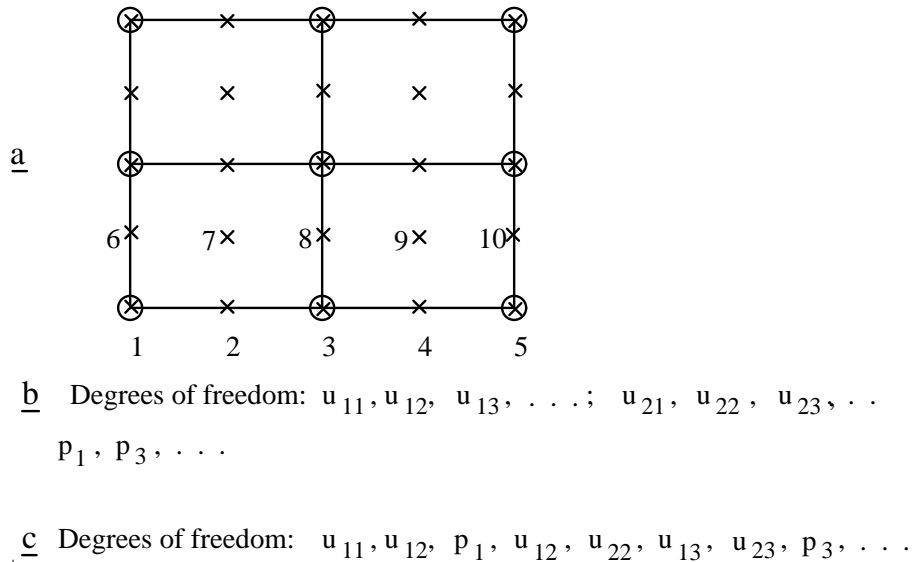


Figure 4.12: Renumbering of unknowns: a the region  $\Omega$ , b sequence of unknowns component-wise, c sequence of unknowns nodal point wise.

The resulting system of equations has a much smaller profile than the one for the original system of equations. Due to renumbering, however, it is possible that the first diagonal elements of the matrix are equal to zero. This is for example the case in Figure 4.12 where,

because of the boundary conditions, the first degrees of freedom are pressures, which do not appear in the continuity equation. In order to prevent zeros on the main diagonal, partial pivoting must be applied. Unfortunately, partial pivoting reorders the sequence of the equations and increases the profile or band width. Therefore a large amount of extra computing time and computer memory is required. However, it still remains cheaper than application of the numbering of Figure 4.12b. It is possible to define a numbering which produces a nearly optimal profile and prevents the appearance of zeros at the start of the main diagonal. Such a numbering, however, goes beyond the scope of this lecture.

Another problem arising from the zeros at the main diagonal is that it is not simple to use iterative methods for the solution of the systems of linear equations.

In Chapters 5 and 6 we shall derive some alternative solution techniques in which the computation of pressure and velocity are segregated and as a consequence partial pivoting is not longer necessary.

## 5 The penalty function method

### 5.1 Introduction

In Chapter 4 the discretization of the Navier-Stokes equations has been derived. It has been shown that the direct solution of the resulting system of linear equations introduces extra complications due to the absence of the pressure in the incompressibility constraint. In this chapter we shall discuss a method which tries to solve this problem by segregating computation of velocity and pressure. For the sake of simplicity we shall restrict ourselves to the stationary Stokes equations, the extension to the instationary and to the non-linear case is straightforward.

Consider the stationary linear Stokes equation in dimensionless form:

$$-\frac{1}{Re}\Delta\mathbf{u} + \nabla p = \mathbf{f} , \quad (5.1a)$$

$$div \mathbf{u} = 0 . \quad (5.1b)$$

(5.1a) follows from (4.9a) by neglecting the time-derivative and the convective terms, substitution of (4.9b) and the incompressibility condition in (4.9a-4.9b). For the sake of the argument we restrict ourselves to homogeneous Dirichlet boundary conditions:

$$\mathbf{u} = \mathbf{0} \quad \mathbf{x} \in \partial\Omega \quad (5.2)$$

The pressure  $p$  is unique up to an additive constant. The idea of the penalty method is to perturb the continuity equation (5.1b) by a small term containing the pressure. An obvious choice is

$$\varepsilon p + div \mathbf{u} = 0 , \quad (5.3)$$

however, in the literature several other possibilities have been proposed.

The pressure  $p$  can be eliminated from (5.3) and substituted into (5.1a) resulting in an equation for the velocity:

$$p = -\frac{1}{\varepsilon} div \mathbf{u} , \quad (5.4)$$

$$-\frac{1}{Re}\Delta\mathbf{u} - \frac{1}{\varepsilon}\nabla(div \mathbf{u}) = \mathbf{f} . \quad (5.5)$$

So one can first solve the velocity from (5.5) and afterwards compute the pressure directly from (5.4). Such an approach will be called segregated approach.

The perturbation (5.3) makes only sense if the solution of (5.4), (5.5) approaches the solution of (5.1a-5.1b) for  $\varepsilon$  approaching zero. It is a simple mathematical exercise to show that this is indeed the case. See for example Cuvelier et al (1986) for the details.

The discretization of the penalty function method may be applied in two ways. One may first discretize the Stokes equations and then apply the penalty function method, or one may discretize the formulation (5.4), (5.5). Both approaches will be treated separately in the Sections 5.2 and 5.3.

Remark: the origin of the penalty method is motivated by the theory of optimization with constraints. One can show (see Chapter 6) that (5.1a-5.1b) is equivalent to the constrained minimization problem:

$$\int_{\Omega} \frac{1}{2} \frac{1}{Re} |\nabla \mathbf{u}|^2 - \mathbf{u} \cdot \mathbf{f} \, d\Omega , \quad (5.6)$$

for all functions  $\mathbf{u}$  satisfying  $div \mathbf{u} = 0$ .

## 5.2 The discrete penalty functions approach

In the discrete penalty function method, the (Navier-)stokes equations are discretized before applying the penalty function method. So we start with the formulation (4.42a-4.42b):

$$\mathbf{S}\mathbf{u} - \mathbf{L}^T \mathbf{p} = \mathbf{F} , \quad (5.7a)$$

$$\mathbf{L}\mathbf{u} = \mathbf{0} . \quad (5.7b)$$

The continuity equation is perturbed by a term  $\varepsilon \mathbf{M}_p \mathbf{p}$ , where  $\mathbf{M}_p$  is the so-called pressure mass matrix, defined by

$$\mathbf{M}_p(i, j) = \int_{\Omega} \psi_i \psi_j \, d\Omega , \quad (5.8)$$

Hence

$$\varepsilon \mathbf{M}_p \mathbf{p} + \mathbf{L}\mathbf{u} = \mathbf{0} , \quad (5.9)$$

or

$$\mathbf{p} = -\frac{1}{\varepsilon} \mathbf{M}_p^{-1} \mathbf{L}\mathbf{u} . \quad (5.10)$$

Substitution of (5.10) in (5.7a) gives

$$\left( \mathbf{S} + \frac{1}{\varepsilon} \mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L} \right) \mathbf{u} = \mathbf{F} . \quad (5.11)$$

So  $\mathbf{u}$  is computed from (5.11) and afterwards  $\mathbf{p}$  is computed from (5.10). In exactly the same way as for the continuous equation, it can be shown that the solution of (5.10), (5.11) approaches the solution of (5.7a-5.7b).

If we want to solve (5.10), (5.11), it is necessary that the matrix  $\mathbf{M}_p^{-1}$  can be computed easily. This is for example the case if  $\mathbf{M}_p$  is a lumped mass matrix. In the discontinuous pressure elements,  $\mathbf{M}_p$  is in a block diagonal matrix, i.e. a diagonal matrix consisting of small matrices as diagonal elements. One can easily verify that these small matrices have the size of the number of pressure unknowns per element, since  $\mathbf{M}_p(i, j) = 0$  if  $\psi_i$  and  $\psi_j$  correspond to different elements. So for the Taylor-Hood family the matrix  $\mathbf{M}_p$  is lumped, in the Crouzeix-Raviart family inversion of  $\mathbf{M}_p$  is quite simple.

Another practical aspect is that the building of the matrix  $\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}$  must be easy. Moreover, it would be very nice if this matrix could be build per element by element matrices. In that case the structures of  $\mathbf{S}$  and  $\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}$  are identical and the solution of (5.11) is as simple as the solution of  $\mathbf{S}\mathbf{u} = \mathbf{F}$ . One can immediately verify that for the Taylor-Hood elements this is not the case. Consider for example the simple triangular mesh in Figure 5.1. For simplicity

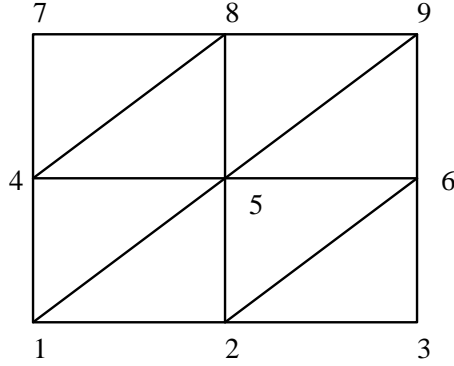


Figure 5.1: Triangular mesh with quadratic Taylor-Hood triangles.

only the vertices of the triangles have been numbered. The midside points are present but are not shown. From Chapter 2 it is clear that in the momentum equation for unknowns in point 5 only the vertex unknowns in the points 1, 2, 4, 5, 6, 8 and 9 are present not those of the points 3 and 7. If we furthermore simplify the matrix  $\mathbf{M}_p$  to a unity matrix, we can compute the elements of  $\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}$  relatively simple by:

$$\mathbf{L}^T \mathbf{L}(i, j) = \sum_k \mathbf{L}^T(i, k) \mathbf{L}(k, j) = \sum_k \mathbf{L}(k, i) \mathbf{L}(k, j) . \quad (5.12)$$

Let us for the sake of the argument identify the matrix elements with the vertex numbers. In fact each matrix element is in that case a  $2 \times 2$  matrix itself.

From chapter 2 it is clear that  $\mathbf{S}(5, 7)$  is equal to zero, since node 5 and node 7 do not belong to the same element. However,  $\mathbf{L}^T \mathbf{L}(5, 7) = \sum_k \mathbf{L}(k, 5) \mathbf{L}(k, 7)$  is in general unequal to zero, since for example  $\mathbf{L}(4, 5)$  and  $\mathbf{L}(4, 7)$  are unequal to zero. So  $\mathbf{L}^T \mathbf{L}$  has a larger bandwidth or profile than  $\mathbf{S}$ .

In the case of a Crouzeix-Raviart element  $L(i, j)$  is only non-zero as long as point  $i$  and point  $j$  belong to the same element, because of the discontinuity of the pressure approximation. As a consequence the matrix  $\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}$  may be split into a sum over element matrices and  $\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}$  may be evaluated at element level. This makes the implementation of the penalty function method relatively easy.

Before we consider some practical remarks concerning the penalty method in Section 5.4, we shall first analyze the so-called continuous penalty function method.

### 5.3 The continuous penalty function method

The penalty function method as introduced in Section 5.2 will be called discrete penalty function method, since first the equations are discretized and then the pressure is eliminated. Conceptually it is much easier to start with the penalty function formulation (5.4), (5.5), and then to discretize the equations.

For the sake of the argument we shall restrict ourselves to Dirichlet boundary conditions for the velocity. Application of the standard Galerkin method to equation (5.5) gives component-

wise:

$$\int_{\Omega} \left\{ \frac{1}{Re} \nabla(\mathbf{u}_h)_k \cdot \nabla \varphi_i + \frac{1}{\varepsilon} \operatorname{div} \mathbf{u}_h \frac{\partial \varphi_i}{\partial x_k} \right\} d\Omega = \int_{\Omega} f_k \varphi_i d\Omega, \quad (5.13)$$

$$i = 1, 2, \dots, n; \quad k = 1, 2, 3,$$

where  $(\mathbf{u}_h)_k$  denotes the  $k$ -th component of  $\mathbf{u}_h$  and  $f_k$  the  $k$ -th component of  $\mathbf{f}$ . In matrix vector notation (5.13) can be written as

$$\mathbf{S}\mathbf{u} + \frac{1}{\varepsilon} \mathbf{A}\mathbf{u} = \mathbf{F}. \quad (5.14)$$

A clear advantage of the formulation (5.14) is that it is no longer necessary to compute the matrix  $\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}$  and as a consequence Taylor-Hood elements are as simple as Crouzeix-Raviart elements. However, a closer examination of (5.13) shows that (5.14) has a disadvantage which is not present in (5.11).

If in (5.11) we let  $\varepsilon$  approach zero, it is immediately clear that

$$\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}\mathbf{u} \rightarrow 0, \quad (5.15)$$

and it is easy to show that also

$$\mathbf{L}\mathbf{u} \rightarrow \mathbf{0}. \quad (5.16)$$

From (5.14) it follows that  $\varepsilon \rightarrow 0$  implies

$$\mathbf{A}\mathbf{u} \rightarrow 0, \quad (5.17)$$

or

$$\int_{\Omega} \operatorname{div} \mathbf{u}_n \operatorname{div} \varphi_i d\Omega = 0, \quad (5.18)$$

for all basis functions  $\varphi_i$ .

(5.18) is equivalent to (4.26), where  $\operatorname{div} \varphi_i$  plays the role of the basis function  $\psi_i$ .

Now consider the quadratic Taylor-Hood element. In equation (5.18)  $\operatorname{div} \varphi_i$  is a linear discontinuous polynomial. So relation (5.18) is comparable to the discretization of the continuity equation for a quadratic Crouzeix-Raviart element with linear pressure. From Section 4.8, however, we know that such an element is not admissible.

Although the penalty function formulation does not give rise to singular systems of equations, still one can expect some troubles with elements which in the limit approach non-admissible elements. Indeed, computations with this approach, show that the velocity behaves rather good, but that the pressure produce unrealistic wiggles (Sani et al 1981). These wiggles are generally known as spurious modes or checkerboard modes for the pressure.

In the literature non-admissible elements are frequently used. To suppress the wiggles one either uses some filtering (smoothing) of the computed pressure, or the penalty matrix is computed with a so-called reduced integration technique (Malhus et al 1978). The filtering technique may produce nice results, however, this technique is not so easy at non-rectangular grids. With the reduced integration technique, actually the term  $\int_{\Omega} \operatorname{div} \mathbf{u}_n \operatorname{div} \varphi_i d\Omega$  is approximated by an inaccurate quadrature rule. This is comparable to approximating  $\operatorname{div} \varphi_i$  by

a lower degree polynomial. As a consequence the actual pressure approximation is reduced, leading to an admissible but less accurate element.

From the discussion given above it is clear that the discrete penalty function approach is superior above the continuous penalty method. In the remainder of this lecture we shall restrict ourselves to this discrete form.

## 5.4 Practical aspects of the penalty function method

In the previous sections the continuous and discrete penalty function method have been derived. It has been shown that the discrete penalty function method, applied to admissible elements is the most recommendable. What remains is the choice of the parameter  $\varepsilon$ . It is clear that  $\varepsilon$  must be so small that the computed velocity and pressure approximates the actual solution accurately. However, there is one draw back with respect to the penalty function formulation. In fact we add the matrix  $\frac{1}{\varepsilon}\mathbf{L}^T\mathbf{M}_p^{-1}\mathbf{L}$  to the matrix  $\mathbf{S}$  in (5.11). The matrix  $\mathbf{S}$  corresponds to the discretization of a vector Laplacian equation (in the case of Stokes flow) or a convection-diffusion type vector equation (Navier-Stokes flow). It is well known that this matrix is good conditioned and has nice properties for many kinds of solvers.

The matrix  $\mathbf{L}$  is a  $m \times 2n$  matrix, where in general  $m \ll n$ . The maximal *rank* of  $\mathbf{L}$  is  $m$  or  $m - 1$ , depending on the type of boundary conditions. As a consequence the *rank* of the  $2n \times 2n$  matrix  $\mathbf{L}^T\mathbf{L}$  can also not exceed  $m$  or  $m - 1$ . The same is true for the matrix  $\mathbf{L}^T\mathbf{M}_p^{-1}\mathbf{L}$ . As a consequence the penalty matrix is a singular matrix with a large number of dependent rows. This penalty matrix is multiplied by a large number  $1/\varepsilon$  and added to a non-singular matrix. It is very natural to assume that the resulting matrix has a condition number which is proportional to  $1/\varepsilon$ . Indeed practical computations show such a behavior. As a consequence  $\varepsilon$  may not be chosen too small since otherwise the condition of the resulting matrix is so bad that an accurate numerical solution is not longer possible. As a rule of the thumb one may choose  $\varepsilon$  such that

$$\|\varepsilon p\| \approx k\|u\|, \quad (5.19)$$

where  $k$  is some value between  $O(10^{-3})$  and  $O(10^{-9})$ . This statement is based on a 64 bits accuracy for the computations, i.e. double precision arithmetic on a 32 - bits computer. Especially for very viscous flow, which for example appear in non-newtonian fluids, a good choice for  $\varepsilon$  may be hard to find.

The fact that we have to choose  $\varepsilon$  carefully is a clear disadvantage. The relative large condition number has also another disadvantage. It is nearly impossible to solve the matrix with standard iterative techniques. Only if we enlarge the value of  $\varepsilon$  and use some outer iterative procedure, it is possible to use penalty function type methods in combination with iterative linear solvers. A well known outer iterative procedure is the so-called Uzawa scheme (Cuvelier et al 1986), which is however, beyond the scope of this lecture.

Despite the clear disadvantages of the penalty function method, still this method is very popular. The reason is that it is a rather simple and fast method, provided the number of unknowns are not too large. The segregation of pressure and velocity gives a large reduction in computing time compared to the direct solution of the original equations. Only for large three-dimensional problems, direct linear solvers become so expensive that it is practically

nearly impossible to apply this method.

In the next chapter we shall derive an alternative segregated method to solve the incompressible Navier-Stokes equations, the so-called solenoidal approach.

## 6 Divergence-free elements

### 6.1 Introduction

In Chapter 4 we have treated the standard Galerkin approach. It has been shown that this method may be applied, provided an admissible element is used. A clear disadvantage of the Galerkin method is the unavailability of the pressure in the continuity equation, and as a consequence the presence of zeros on the main diagonal of the equations (4.42a-4.42b). As a consequence the solution of the equations introduces extra difficulties. In the penalty function method we have solved this problem by segregating pressure and velocity, thus reducing the number of unknowns as well as avoiding the zeros at the main diagonal. The only problem with this method is that it is sometimes difficult to get a good choice of the small parameter  $\varepsilon$  and the bad condition of the remaining system of equations. Especially for very viscous (non-newtonian fluids) this may be a problem. In this chapter we shall derive an alternative segregated approach in which it is not necessary to choose some parameter, and which does not lead to ill conditioned systems of equations.

To that end we consider the weak formulation (4.21-4.22). If for the sake of simplicity we neglect both the convective terms and all boundary integrals, substitute the continuity equation in the stress tensor and use the dimensionless form, (4.21-4.22) can be written as:

$$\int_{\Omega} q \operatorname{div} \mathbf{u} d\Omega = 0 , \quad (6.1a)$$

$$\int_{\Omega} \frac{1}{Re} \nabla \mathbf{u} \cdot \nabla \mathbf{v} d\Omega - \int_{\Omega} p \operatorname{div} \mathbf{v} d\Omega = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\Omega . \quad (6.1b)$$

Except with respect to the essential boundary conditions,  $\mathbf{u}$  and  $\mathbf{v}$  are chosen in the same space. So if we restrict this space to all divergence-free vector fields, then it is immediately clear that (6.1a) is satisfied automatically and, moreover,  $\int_{\Omega} p \operatorname{div} \mathbf{v} d\Omega$  vanishes. In other words an equation in the velocity alone remains. Unfortunately it is very hard to find functions which are completely divergence-free. However, formulation (6.1a-6.1b) shows that it is not necessary to demand  $\operatorname{div} \mathbf{u} = 0$ , but that it is sufficient to weaken this statement to

$$\int_{\Omega} q \operatorname{div} \mathbf{v} d\Omega = 0 \quad \text{for all } q . \quad (6.2)$$

If both our test functions and the solution  $\mathbf{u}$  satisfy (6.2), (6.1a) is satisfied and (6.1b) reduced to

$$\int_{\Omega} \frac{1}{Re} \nabla \mathbf{u} \cdot \nabla \mathbf{v} d\Omega = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\Omega , \quad \text{for all } \mathbf{v} , \quad (6.3)$$

which is again an equation for the velocity alone.

If we construct a basis  $\varphi_i$  in the space of approximately divergence-free vector fields satisfying (6.2), (6.3) can be written as:

$$\mathbf{u}_h = \sum_{j=1}^n u_j \varphi_j(\mathbf{x}) , \quad (6.4a)$$

$$\sum_{j=1}^n u_j \int_{\Omega} \frac{1}{Re} \nabla \varphi_j \cdot \nabla \varphi_i d\Omega = \int_{\Omega} \mathbf{f} \cdot \varphi_i d\Omega, \quad i = 1(1)n. \quad (6.4b)$$

The extension to the non-linear Navier-Stokes equations, the general form of stress tensor and non-vanishing boundary integrals is trivial.

The system of equations (6.4b) is of double Laplacian type (in  $\mathbb{R}^2$ ), and may be solved quite easily. The only problem is of course, how to construct basis functions that are divergence-free in the sense of (6.2). In Section 6.2 we shall show the construction of such basis functions for one specific element.

## 6.2 The construction of divergence-free basis functions for 2D elements

In this section we shall construct divergence-free basis functions in the sense of (6.2) for two-dimensional elements. The extension to  $\mathbb{R}^3$  is quite complicated and introduces a number of extra problems. We refer to Cuvelier et al (1986) for a derivation. The construction of divergence-free basis functions is relatively simple for elements of the Crouzeix-Raviart type, how to construct such functions for Taylor-Hood elements is not known at this moment. For simplicity we shall restrict ourselves to triangular elements; the extension to quadrilaterals is straightforward. In fact we shall derive these basis functions for the extended quadratic triangle of Figure 4.9, with the basis functions given in (4.40). But in order to get some insight in the problems associated with this derivation we shall first consider the non-conforming linear triangle with constant pressure given in Figure 4.8.

In general (6.2) can be written as

$$\int_{\Omega} \Psi_i \operatorname{div} \mathbf{u}_h d\Omega = 0 \quad \text{for all pressure basis functions } \Psi_i. \quad (6.5)$$

In other words we have to construct basis functions  $\varphi_i$  such that

$$\int_{\Omega} \Psi_j \operatorname{div} \varphi_i d\Omega = 0, \quad \text{for all } \Psi_j. \quad (6.6)$$

One may expect that the basis functions  $\varphi_i(\mathbf{x})$  satisfying (6.6), have vector components, which are both nonzero, so these will be linear combinations of the classical basis functions defined by (4.23).

In order to find these linear combinations we recall that for the non-conforming element, (6.6) reduces to

$$\int_{\Omega^{e_j}} \operatorname{div} \varphi_i d\Omega = 0, \quad (6.7)$$

for all elements  $e_j$ , since  $\Psi_j$  is one in element  $e_j$  and zero outside the element.

Application of the Gauss-divergence theorem to (6.7) gives

$$\int_{\Gamma^{e_j}} \varphi_i \cdot \mathbf{n} d\Gamma = 0, \quad (6.8)$$

where  $\Gamma^{e_j}$  is the boundary of the triangle  $e_j$ . In each triangle we have 6 unknown velocities corresponding to the three midside points. In all previous examples (compare with 4.25) these

velocity components were in fact the Cartesian components. However, for our purposes it is better to decompose the velocity in a tangential and a normal component along the boundary of the triangle. See Figure 6.1 for a definition. In this element the velocity  $\mathbf{u}$  is approximated by

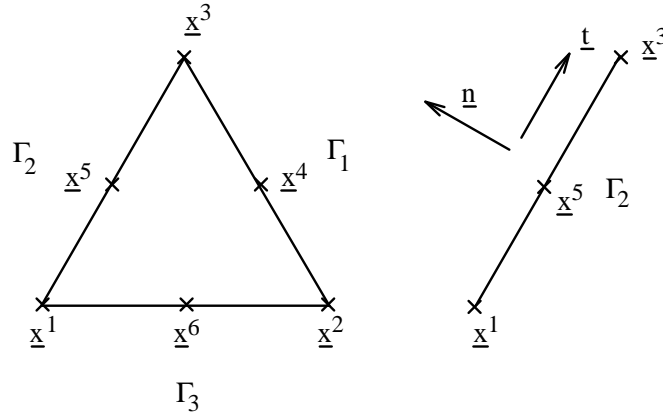


Figure 6.1: Element  $e$ , with boundary  $\Gamma_2$  separately plotted. Normal and tangential unity vectors on  $\Gamma_2$  are indicated.

$$\mathbf{u}_h = \sum_{k=2,4,6} u_{1k} \varphi_{k1}(\mathbf{x}) + u_{2k} \varphi_{k2}(\mathbf{x}), \quad (6.9)$$

which can be written as

$$\mathbf{u}_h = \sum_{k=2,4,6} u_{nk} \varphi_{kn}(\mathbf{x}) + u_{tk} \varphi_{kt}(\mathbf{x}), \quad (6.10)$$

where  $u_{nk}$  respectively  $u_{tk}$  denote the normal and tangential component of  $\mathbf{u}$  in node  $k$ .  $\varphi_{kn}(\mathbf{x})$  and  $\varphi_{kt}(\mathbf{x})$  are defined by

$$\varphi_{kn}(\mathbf{x}) = \varphi_k(\mathbf{x}) \mathbf{n}_k, \quad (6.11a)$$

$$\varphi_{kt}(\mathbf{x}) = \varphi_k(\mathbf{x}) \mathbf{t}_k. \quad (6.11b)$$

Here  $\varphi_k(x)$  denotes the scalar basis function corresponding to point  $k$  and  $\mathbf{n}_k$  and  $\mathbf{t}_k$  the normal respectively tangential vector corresponding to the edge on which node  $k$  is positioned.

The basis functions  $\varphi_{kt}(\mathbf{x})$  satisfy (6.8) exactly, since  $\varphi_k(\mathbf{x}) = 1$  on the edge containing node  $k$  and linear from -1 to 1 at the other sides. On the edge containing node  $k$  we have  $\mathbf{n}_k \cdot \mathbf{t}_k = 0$ , on the other edges  $\mathbf{n}$  is constant and the integral over  $\varphi_k(\mathbf{x})$  vanishes because of the linearity. So one set of divergence-free basis functions is formed by the set of basis functions corresponding to the tangential components.

The other set of basis functions must be constructed such that (6.8) is satisfied. Now

$$\int_{\Gamma_k} \mathbf{u}_h \cdot \mathbf{n} d\Gamma \quad (6.12)$$

defines the amount of flow through side  $\Gamma_k$ . For an incompressible flow one can define a stream function  $\Psi$  by

$$\mathbf{u} = \left( \frac{\partial \Psi}{\partial y}, -\frac{\partial \Psi}{\partial x} \right). \quad (6.13)$$

An important property of the stream function is that the difference between the values of the stream function in two points defines the amount of flow between these two points. See Figure 6.2 for an explanation. So it is quite natural to define a discrete stream function in

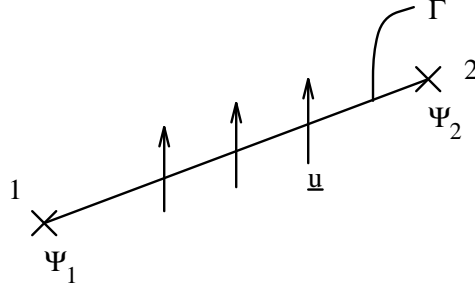


Figure 6.2: Amount of flow between points 1 and 2 is given by  $\int_{\Gamma} \underline{u} \cdot \underline{n} d\Gamma = \Psi_2 - \Psi_1$

the vertices by

$$\Psi_{k+1} - \Psi_k = \int_{\Gamma_{k+2}} \mathbf{u}_h \cdot \mathbf{n} d\Gamma, \quad k = 1, 2, 3, \quad (6.14)$$

where a cyclic permutation with the numbers 1, 2 and 3 is used. It is clear that definition (6.14) does not introduce any contradictions. Moreover, this stream function  $\Psi$  is constructed such that (6.8) is satisfied exactly.

Definition (6.14) is unique for the complete mesh, since in (6.14) only values on one element side are used, so the definition in contiguous elements is the same. Furthermore, given a divergence-free vector field in the sense of (6.2) or (6.5), the stream function  $\Psi$  can be computed in each vertex, provided it is fixed in an arbitrary vertex.

From (6.14) we can express the normal component on the mid side points into the values of the stream function on the vertices. Using the fact that the basis function  $\varphi_k(\mathbf{x})$  is equal to 1 along the edge corresponding to node  $k$ , it follows immediately that

$$u_{n2} = \frac{\Psi_2 - \Psi_1}{L_1}, \quad u_{n4} = \frac{\Psi_3 - \Psi_2}{L_2}, \quad u_{n6} = \frac{\Psi_1 - \Psi_3}{L_3}, \quad (6.15)$$

where  $L_k$  denotes the length of side  $\Gamma_k$ .

Substitution of (6.15) into (6.10) gives

$$\begin{aligned} \mathbf{u}_h &= \sum_{k=2,4,6} u_{tk} \varphi_{kt}(\mathbf{x}) + \left( \frac{1}{L_2} \varphi_{6n} - \frac{1}{L_3} \varphi_{2n} \right) \Psi_1 \\ &+ \left( \frac{1}{L_3} \varphi_{2n} - \frac{1}{L_1} \varphi_{4n} \right) \Psi_2 + \left( \frac{1}{L_1} \varphi_{4n} - \frac{1}{L_2} \varphi_{6n} \right) \Psi_3. \end{aligned} \quad (6.16)$$

In other words, the second set of basis functions, denoted by  $\varphi_{k\Psi}$  ( $k = 1, 2, 3$ ) is given by

$$\varphi_{k\Psi} = \frac{1}{L_{k+1}} \varphi_{2(k+2)n} - \frac{1}{L_{k+2}} \varphi_{2k} \quad (\text{cyclic}). \quad (6.17)$$

One easily verifies that these functions satisfy (6.8).

This completes the construction of the 6 basis functions that are divergence-free. Substitution of these basis functions into the weak formulation (6.4b) gives a system of linear equations in the unknowns  $u_{tk}$  and  $\Psi_k$ .

The basis functions  $\varphi_{kt}$  and  $\varphi_{k\Psi}$  are characterized by the following properties:

- i The components of  $\varphi_{kt}$  and  $\varphi_{k\Psi}$  are linear in  $x$  and  $y$  per element.
- ii  $\varphi_{tj} = 0$  at the midside nodes not equal to  $j$ .  $\varphi_{tj}$  is equal to the unit tangential vector in midside node  $j$ .
- iii  $\varphi_{\Psi j} = 0$  at the mid-side node opposite to vertex  $j$ .  $\varphi_{\Psi j}$  is equal to plus or minus the unit normal vector divided by the length of the side, in the two mid-side points of the sides containing the vertex  $j$ . The sign is opposite for these two points.

It must be remarked that in order to get a unique definition of normal and tangential components, it is necessary to define the normal and tangential vector in the same way in adjacent elements. A possible unique definition is to choose the tangential vector from smallest node number to highest node number and defining the corresponding normal vector in the clockwise direction. See Figure 6.3.

In conclusion, the procedure to construct divergence-free basis functions consists of the following steps:

- define basis functions corresponding to normal components and tangential components at mid side points
- the first set of basis functions is formed by the basis functions corresponding to the tangential components
- Introduce stream function unknowns at the vertices and eliminate the normal components of the velocity at mid-side points by expressing them in the stream function unknowns. The basis functions corresponding to the stream function unknowns from the second set of basis functions.

The computation of the pressure is postponed to Section 6.5. Now we have seen how the

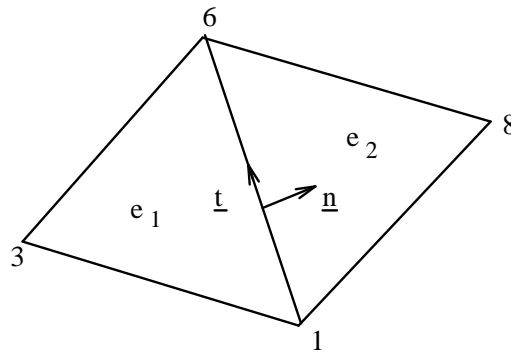


Figure 6.3: Definition of normal and tangential vector at side 1-6 for elements  $e_1$  and  $e_2$ . The global node numbering is plotted.

divergence-free basis functions may be derived for the non-conforming triangle, we shall apply the same procedure for the more complex extended quadratic triangle with linear pressure.

The first step in the construction of the divergence-free basis functions is the elimination of the velocity in the centroid points of the elements and as a consequence the gradient of the pressure in these points. This process is also known as static condensation.

Substitution of the pressure basis functions given in (4.40) into (6.5) gives

$$\int_e \operatorname{div} \mathbf{u}_h d\Omega = 0, \quad (6.18a)$$

$$\int_e (\mathbf{x} - \mathbf{x}_7) \operatorname{div} \mathbf{u}_h d\Omega = 0, \quad (6.18b)$$

where the centroid is denoted by  $\mathbf{x}_7$ .

From (6.18b)  $\mathbf{u}_h$  in the centroid can be expressed in terms of the velocity components in the remaining points of the element. It can be easily shown that (6.18b) is a regular system of equations for  $\mathbf{u}_7$  (see for example Cuvelier et al 1986). Substitution of the basis functions  $\varphi_{17}(\mathbf{x})$  and  $\varphi_{27}(\mathbf{x})$  per element into (6.1b) gives an expression for the gradient of the pressure in terms of the velocity unknowns at the boundary. Also in Cuvelier et al (1986) it is shown that this expression is never singular. So in fact both the pressure gradient and the velocity in the centroid have been eliminated. The practical procedure will be treated in Section 6.3.

Once the velocity components in the centroid have been eliminated, twelve velocity components remain, six in the vertices and six in the mid side points.

We write the approximation  $\mathbf{u}_h$  in the following form per triangle

$$\mathbf{u}_h = \sum_{i=1}^3 \{u_{1i} \Phi_{1i} + u_{2i} \Phi_{2i}\} + \sum_{i=4}^6 \{u_{ni} \Phi_{ni} + u_{ti} \Phi_{ti}\}. \quad (6.19)$$

The functions  $\Phi_{1i}$  and  $\Phi_{2i}$  are adapted basis functions because of the elimination of the centroid degrees of freedom. The velocity components in the midside points are split in normal and tangential components in exactly the same way as for the non-conforming element.

For this element again, the stream function in the vertices is introduced as new unknowns and the velocity components in the midside points are eliminated using relation (6.14). It can be shown that the thus constructed basis functions are divergence-free.

In the next section we shall treat how the element matrices and vectors corresponding to the (approximate) divergence-free basis functions may be computed, without actually creating these basis functions.

### 6.3 The construction of element matrices and vectors for (approximate) divergence-free basis functions

The Galerkin equations for the Stokes equations using divergence-free basis functions are given in (6.4b). The corresponding element matrices and vectors may be constructed by explicit substitution of the divergence-free basis functions constructed in Section 6.2.

However, an alternative possibility is to start with the original set of equations (4.26), (4.27) using the classical basis functions and then to perform the elimination process. Let us demonstrate this process for the extended quadratic triangle. We shall execute the algorithm in two steps. In the first step the centroid velocity components and the gradient of the pressure are eliminated. In the second step the normal components of the velocity at mid-side points are eliminated.

To perform step 1 we start with the system of linear equations (4.42a-4.42b):

$$\mathbf{S}\mathbf{u} - \mathbf{L}^T\mathbf{p} = \mathbf{F} , \quad (6.20a)$$

$$\mathbf{L}\mathbf{u} = \mathbf{0} . \quad (6.20b)$$

The velocity  $\mathbf{u}$  will be split into a part corresponding to the centroid ( $\mathbf{u}_z$ ) and the rest of the velocities ( $\hat{\mathbf{u}}$ ). In the same way the pressure  $\mathbf{p}$  will be split into a part  $p_\nabla$  corresponding to the gradient of the pressure in the centroid and a part  $\hat{\mathbf{p}}$  corresponding to the values of  $p$  in the centroids. Hence we define

$$\mathbf{u} = \begin{bmatrix} \hat{\mathbf{u}} \\ \mathbf{u}_z \end{bmatrix} , \quad \mathbf{p} = \begin{bmatrix} \hat{\mathbf{p}} \\ \mathbf{p}_\nabla \end{bmatrix} . \quad (6.21)$$

If we split equations (6.20a-6.20b) according to (6.21) we get

$$\mathbf{S}_1\hat{\mathbf{u}} + \mathbf{S}_2\mathbf{u}_z - \mathbf{L}_2^T\hat{\mathbf{p}}_\nabla = \mathbf{F} , \quad (6.22a)$$

$$\mathbf{L}_1\mathbf{u} = \mathbf{L}_{11}\hat{\mathbf{u}} + \mathbf{L}_{12}\mathbf{u}_z = \mathbf{0} , \quad (6.22b)$$

$$\mathbf{L}_2\mathbf{u} = \mathbf{L}_{21}\hat{\mathbf{u}} + \mathbf{L}_{22}\mathbf{u}_z = \mathbf{0} . \quad (6.22c)$$

The elimination of the velocity components in the centroid follows from (6.22c):

$$\mathbf{u}_z = -\mathbf{L}_{22}^{-1}\mathbf{L}_{21}\hat{\mathbf{u}} , \quad (6.23)$$

in other words

$$\mathbf{u} = \mathbf{R}_z\hat{\mathbf{u}} , \quad (6.24)$$

with

$$\mathbf{R}_z = \begin{bmatrix} \mathbf{I} \\ \mathbf{R}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \\ -\mathbf{L}_{22}^{-1} & \mathbf{L}_{21} \end{bmatrix} . \quad (6.25)$$

From (6.25) we get

$$\mathbf{L}_2\mathbf{R}_z = \mathbf{L}_{21}\mathbf{I} + \mathbf{L}_{22}\mathbf{R}_0 = \mathbf{L}_{21} - \mathbf{L}_{22}\mathbf{L}_{22}^{-1}\mathbf{L}_{21} = \mathbf{0} , \quad (6.26)$$

and hence also

$$\mathbf{R}_z^T\mathbf{L}_2 = \mathbf{0} . \quad (6.27)$$

Substitution of (6.24) into (6.22a) gives:

$$\mathbf{S}_1\hat{\mathbf{u}} + \mathbf{S}_2\mathbf{R}_0\hat{\mathbf{u}} - \mathbf{L}_1^T\hat{\mathbf{p}} - \mathbf{L}_2^T\mathbf{p}_\nabla = \mathbf{0} . \quad (6.28)$$

If we premultiply (6.28) by  $\mathbf{R}_z^T$  and use (6.27) we get:

$$(\mathbf{R}_z^T\mathbf{S}_1 + \mathbf{R}_z^T\mathbf{S}_2\mathbf{R}_0)\hat{\mathbf{u}} - \mathbf{R}_z^T\mathbf{L}_1^T\hat{\mathbf{p}} = \mathbf{R}_z^T\mathbf{F} , \quad (6.29)$$

and (6.22b) can be written as

$$\mathbf{L}_1 \mathbf{u} = \mathbf{L}_1 \mathbf{R}_z \hat{\mathbf{u}} = \mathbf{0} . \quad (6.30)$$

In other words the result of the elimination process is

$$\hat{\mathbf{S}} \hat{\mathbf{u}} - \hat{\mathbf{L}}^T \hat{\mathbf{p}} = \hat{\mathbf{F}} , \quad (6.31a)$$

$$\hat{\mathbf{L}} \hat{\mathbf{u}} = \mathbf{0} , \quad (6.31b)$$

with

$$\hat{\mathbf{S}} = \mathbf{R}_z^T \mathbf{S}_1 + \mathbf{R}_z^T \mathbf{S}_2 \mathbf{R}_0 , \quad (6.32a)$$

$$\hat{\mathbf{L}} = \mathbf{L}_1 \mathbf{R}_z , \quad (6.32b)$$

$$\hat{\mathbf{F}} = \mathbf{R}_z^T \mathbf{F} . \quad (6.32c)$$

Due to the discontinuous character of the pressure, the matrices  $\hat{\mathbf{S}}$  and  $\hat{\mathbf{L}}$  and the right-hand side vector  $\hat{\mathbf{F}}$  may be computed at element level.

The next step is the elimination of the normal components in the midside points in favor of the stream function at the vertices. This elimination process can be expressed by a matrix  $\mathbf{R}_d$  according to

$$\hat{\mathbf{u}} = \mathbf{R}_d \mathbf{u}_d , \quad (6.33)$$

where  $\mathbf{u}_d$  is the vector of new unknowns.

The transformation is such that the continuity equation is satisfied exactly in other words

$$\hat{\mathbf{L}} \mathbf{R}_d = \mathbf{0} . \quad (6.34)$$

Substitution of (6.34) in (6.31a-6.31b) and premultiplication by  $\mathbf{R}_d^T$  gives

$$\mathbf{R}_d^T \hat{\mathbf{S}} \mathbf{R}_d \mathbf{u}_d = \mathbf{R}_d^T \hat{\mathbf{F}} . \quad (6.35)$$

Again the matrix and vector can be constructed at elements level.

So we have shown that it is not necessary to construct the divergence-free basis functions explicitly. It is sufficient to construct the transformation matrices  $\mathbf{R}_z$  and  $\mathbf{R}_d$  per element and to compute the final element matrix and element right-hand side by matrix-matrix respectively matrix-vector multiplications.

We have constructed a new set of equations with new unknowns. The question that remains is of course: is this new system of equations uniquely solvable? Furthermore, which type of boundary conditions must be prescribed to the new unknowns. These questions will be the subject of Section 6.4.

## 6.4 Boundary conditions with respect to the divergence-free elements

In the construction of the divergence-free elements treated in the preceding sections, it was necessary to introduce new unknowns. First of all the velocity has been decomposed into normal and tangential part. With respect to the boundary conditions this does not introduce extra problems, since in general boundary conditions are formulated in tangential and normal direction and not in Cartesian directions. Next the stream function has been introduced as

new unknown. This introduces two extra problems. The first one is that the stream function is never unique but fixed up to an additive constant. As a consequence, it is necessary to prescribe the stream function in at least one point. The second one is that, if the normal velocity is unknown at a part of the boundary, it is not automatically possible to compute the stream function along that part. This is best demonstrated with the configuration of Figure 6.4. In this Figure we have one inflow with prescribed velocity field (boundary i), three fixed walls with no-slip boundary condition (boundaries ii, iv and vi) and two outflow boundaries where for example the normal stress is prescribed (boundaries iii and v). If the stream function at the common point of sides i and vi is set equal to zero then  $\Psi$  along sides i, ii and vi may be computed from the definition (6.14). The stream function value at sides iii and v does not have to be prescribed, since the normal component at these sides is not prescribed. However, at side iv we have a no-slip condition, implying  $u_n = 0$ . Hence  $\Psi$  is constant at side iv but the value is unknown. So for such boundaries it is necessary to prescribe the boundary condition  $\Psi$  is unknown constant. For a practical implementation of such a boundary condition the reader is referred to Cuvelier et al (1986).

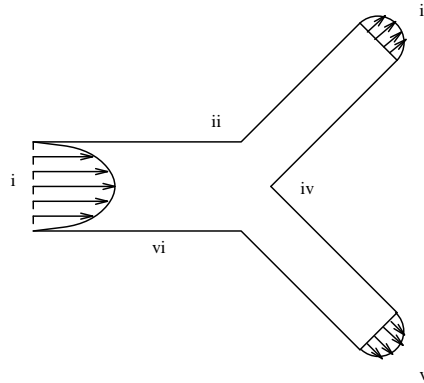


Figure 6.4: Example of a region with two outflow parts. On the boundary iv we have the boundary condition  $\Psi$  equals unknown constant

## 6.5 Computation of the pressure

Once the velocity is known, the pressure must be computed. Let us first restrict ourselves to the non-conforming triangle. We return to the weak formulation (6.1b), and substitute non-divergence-free basis functions. An obvious choice is to use the basis functions  $\varphi_{nh}$  corresponding to the normal components of the velocity. Substitution of these basis functions in (6.1b) gives

$$\int_{\Omega} \frac{1}{Re} \nabla \mathbf{u}_h \cdot \nabla \varphi_{nk} d\Omega - \int_{\Omega} p_h \operatorname{div} \varphi_{nk} = \int_{\Omega} \mathbf{f} \cdot \varphi_{nk} d\Omega, \quad k = 1, 2, \dots \quad (6.36)$$

Since  $\mathbf{u}_h$  is known, this is an equation in the unknowns  $p$  alone. Since  $\varphi_{nk} = 0$  outside the two elements containing node  $k$  it is sufficient to consider two adjacent elements  $e_1$  and  $e_2$  as indicated in Figure 6.5. (6.36) reduces to

$$\int_{e_1 \cup e_2} p_h \operatorname{div} \varphi_n d\Omega = \int_{e_1 \cup e_2} \left\{ \frac{1}{Re} \nabla \mathbf{u}_h \cdot \nabla \varphi_n - \mathbf{f} \cdot \varphi_n \right\} d\Omega, \quad (6.37)$$

where  $\varphi_n$  is the abbreviated notation for  $\varphi_{nk}$  in the common midside point. Application of

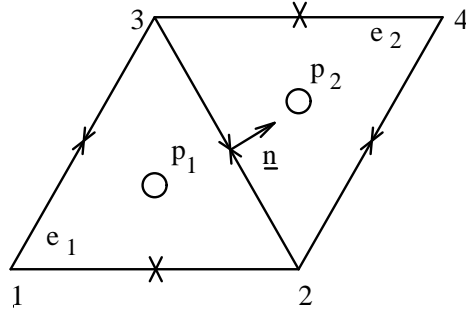


Figure 6.5: Two adjacent elements  $e_1$  and  $e_2$ . The pressure nodal points have been indicated.

the Gauss divergence theorem to the left-hand side of (6.37) gives

$$\int_{e_1 \cup e_2} p_h \operatorname{div} \varphi_n d\Omega = \int_{\partial e_1} p_h \varphi_h \cdot \mathbf{n} d\Gamma + \int_{\partial e_2} p_h \varphi_n \cdot \mathbf{n} d\Gamma = (p_1 - p_2)L_{23}, \quad (6.38)$$

where  $L_{23}$  is the length of side 23.

Hence given  $p_1, p_2$  can be computed immediately. This procedure may be repeated for all adjacent elements. So starting by prescribing the pressure in one element, the pressure can be computed in all elements by finding neighboring elements and applying (6.38).

The computation of the pressure in the case of the extended quadratic Crouzeix-Raviart triangle is again performed in two steps. In step 1 the pressure in the centroids is computed using the method described for the non-conforming triangle.

In the second step the gradient of the pressure in the centroid is computed. To that end equation (6.22a) is applied at element level. This is possible since each row of the divergence matrix  $L$  has only non-zero contributions for one element at a time, due to the discontinuous pressure definition.

Hence

$$-\mathbf{L}_2^T \mathbf{p}_\nabla = \mathbf{F} - \mathbf{S}_1 \hat{\mathbf{u}} - \mathbf{S}_2 \mathbf{u}_z + \mathbf{L}_1^T \hat{\mathbf{p}}, \quad (6.39)$$

where  $\mathbf{u}_z = -\mathbf{L}_{22}^{-1} \mathbf{L}_{21} \hat{\mathbf{u}}$ .

Per element  $\mathbf{L}_2^T$  reduces to a  $(2 \times 2)$  matrix, hence (6.39) immediately defines  $\mathbf{p}_\nabla$  per element.

## 6.6 Practical aspects of the divergence-free elements

In this chapter we have shown how divergence-free elements may be constructed in 2D. The derivation has been restricted to discontinuous pressure elements. The extension to three-dimensional element is quite complicated, and seems rather impractical. The construction of element matrices and vectors may be performed by the introduction of transformation matrices. Complicating factor may be the definition of boundary conditions as shown in Section 6.4. Once the velocity is computed a post-processing step is necessary to compute the pressure. A clear advantage of the use of divergence-free elements is that velocity and pressure are segregated, without the introduction of an extra parameter to be chosen. As a consequence, this method allows the solution of the resulting systems of equations by iterative techniques. So for two-dimensional problems the method based on divergence-free elements seems very promising.

## 7 The instationary Navier-Stokes equations

### 7.1 Introduction

Until now we have restricted ourselves to stationary Navier-Stokes equations only. However, in some practical applications one is also interested in the time-dependent behavior of the solution. Or, alternatively, sometimes it is hard to find the solution of a stationary problem, because the iteration process does not converge well enough. In that case, considering the stationary solution as limit of a time-dependent solution may help to get a convergent solution. For that reason we shall consider some methods to solve the time-dependent Navier-Stokes equations.

In first instance we shall use the method of lines as derived in Chapter 3. With respect to the discretization of the continuity equation all three methods derived in the previous chapters may be applied. This will be the subject of Section 7.2.

In Section 7.3 an alternative approach will be treated, which is especially developed for time-dependent incompressible flows. This method, the so-called pressure correction method, consists of two steps per time-step. In first instance the velocity is computed using the pressure at the old time-level and neglecting the continuity equation. In general, the computed velocity is not divergence-free. In the next step the velocity is projected onto the space of divergence-free vectors. This step also introduces a Laplacian type equation for the pressure.

### 7.2 Solution of the instationary Navier-Stokes equations by the method of lines

The instationary incompressible Navier-Stokes equations read (compare with 4.3):

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho \mathbf{u} \cdot \nabla \mathbf{u} - \operatorname{div} \boldsymbol{\sigma} = \rho \mathbf{f} , \quad (7.1a)$$

$$\operatorname{div} \mathbf{u} = 0 . \quad (7.1b)$$

In order to get a finite element discretization, the weak form of (7.1a) - (7.1b) is derived. To that end equation (7.1a) is multiplied by a time-independent test function  $v$  and (7.1b) by a test function  $q$ . If we neglect the boundary integrals and furthermore apply (4.4) in the same way as for the stationary case, the weak form of (7.1a) may be written as

$$\int_{\Omega} \rho \frac{\partial \mathbf{u}}{\partial t} \cdot v d\Omega + \int_{\Omega} 2\mu \mathbf{e} \cdot \nabla v d\Omega + \int_{\Omega} \rho (\mathbf{u} \cdot \nabla \mathbf{u}) v d\Omega - \int_{\Omega} p \operatorname{div} v d\Omega = \int_{\Omega} \mathbf{f} \cdot v d\Omega , \quad (7.2a)$$

$$\int_{\Omega} q \operatorname{div} \mathbf{u} d\Omega = 0 . \quad (7.2b)$$

We see that (7.2b) does not contain a time-derivative, and this will be an extra complicating factor. If we approximate  $\mathbf{u}$  and  $p$  in the same way as in (4.24), (4.25), however, with time-dependent coefficients, and if we substitute the time-independent basis functions  $\varphi_i(\mathbf{x})$

respectively  $\Psi_i(\mathbf{x})$ , the Galerkin method reduces to the solution of a system of nonlinear ordinary differential equations of the form

$$\mathbf{M}\dot{\mathbf{u}} + \mathbf{N}(\mathbf{u}) - \mathbf{L}^T \mathbf{p} = \mathbf{F} , \quad (7.3a)$$

$$\mathbf{L}\mathbf{u} = \mathbf{0} , \quad (7.3b)$$

where  $\mathbf{N}$ ,  $\mathbf{L}$ ,  $\mathbf{f}$ ,  $\mathbf{u}$  and  $\mathbf{p}$  are defined as in (4.28a-4.28b) and  $\mathbf{M}$  denotes the velocity mass matrix, which can be written as:

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & 0 \\ 0 & \mathbf{M}_1 \end{pmatrix} , \quad (7.4)$$

$$\mathbf{M}_1(i, j) = \int_{\Omega} \varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\Omega .$$

The absence of a time-derivative in (7.3b) has as consequence that (7.3b) must be satisfied in every stage of the time integration. An important consequence is that if the equations (7.3a-7.3b) are solved in a coupled way, explicit methods do not make sense. With respect to the time-integration, all the classical methods, such as for example the ones mentioned in Chapter 3 may be used. Very popular are the  $\theta$  methods, especially  $\theta = 1/2$  and  $\theta = 1$ , and the two-step Adams-Bashfort discretization. In this last formulation one usually uses an implicit formulation with respect to the viscous terms and an explicit formulation for the convective terms. The reason for this splitting is that the matrix due to the convective terms changes in each time-step, whereas the other matrices remain constant in time.

Of course the solution of the coupled equations (7.3a-7.3b), introduces exactly the same problems as for the stationary case. As a consequence the same type of solution procedures will be used. Hence it is quite usual to apply a segregated formulation in order to solve (7.3a-7.3b). Both the penalty function approach, as the method with divergence-free basis functions may be applied.

The penalty function approach, applied to (7.3a-7.3b) reads (compare with 5.12):

$$\mathbf{M}\dot{\mathbf{u}} + \mathbf{N}(\mathbf{u}) + \frac{1}{\varepsilon}\mathbf{L}^T \mathbf{M}_p^{-1} \mathbf{L}\mathbf{u} = \mathbf{F} , \quad (7.5a)$$

$$\mathbf{p} = -\frac{1}{\varepsilon}\mathbf{M}_p^{-1} \mathbf{L}\mathbf{u} . \quad (7.5b)$$

It is clear that the pressure has only to be computed, if at a certain moment the pressure is required. In the time-stepping algorithm it is sufficient to solve (7.5a).

With respect to the non-linear terms, it is necessary to perform a kind of linearization. In general exactly the same type of linearization as for the stationary case are used. However, in contrast to the stationary case, no iteration per time-step is applied. In general, the non-linear terms are linearized with respect to the solution at the preceding time-level. If the linearization is not accurate enough, a smaller time-step must be used.

The Picard type linearization of Section 4.6, all produce an  $0(\Delta t)$  error, whereas the Newton type linearization gives an  $0(\Delta t^2)$ . Hence if the Crank-Nicolson scheme is applied, it is more or less necessary to combine this scheme with a Newton linearization.

We have seen in the stationary case that the matrix corresponding to the penalty function

method has a large condition number. As a consequence it was not possible to use iterative methods for the solution of the linear systems of equations. With respect to the instationary case there is another drawback. One can show that the solution of (7.5a) by an explicit time integrator, requires time-steps which are proportional to  $\varepsilon$ , due to stability requirements. Hence, in practice, only implicit methods are used to solve (7.5a). See Cuvelier et al (1986) for the details. The Crank Nicolson scheme has the property that it does not damp high frequencies. Due to the penalty term such frequencies may always be present in the equations. As a consequence extra damping is necessary if the solution is non-smooth in time. For example in the case of a transient one usually starts with one time-step Euler-implicit in order to damp high frequencies, and than one resumes with Crank-Nicolson in order to get a good accuracy.

An alternative for the penalty function method is of course to use divergence-free elements. Although it looks as if there is no reason to use implicit time-methods for this approach, it must be remarked that due to the coupled character of the basis functions, the mass matrix  $\mathbf{M}$  can never be put into diagonal form. As a consequence, even an explicit method, requires the solution of system of linear equations per time-step.

Finally in the next section we shall treat an alternative approach for the incompressible time-independent Navier-Stokes equations, the so-called pressure-correction formulation.

### 7.3 The pressure-correction method

The pressure-correction method has, in first instance, been developed for finite difference methods. It is a special method for incompressible flows. In fact the pressure-correction method consists of two steps. In the first step the momentum equation is solved with the pressure at the preceding time-level. In this step the continuity equation is not taken into account. The resulting velocity field may be considered as an intermediate field. In the next step this intermediate field is projected onto the space of divergence-free vector fields. This step implicitly introduces a Poisson-type equation for the pressure. The pressure-correction method is strongly coupled with the type of time-discretization. We shall demonstrate it for the general  $\theta$ -method. First we shall derive pressure-correction in the case that the space discretization has not yet been applied. Next we shall apply the space discretization first and then derive the pressure-correction method.

#### Continuous approach

Consider the incompressible Navier-Stokes equations in dimensionless form:

$$\frac{\partial \mathbf{u}}{\partial t} - \frac{1}{Re} \Delta \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p = \mathbf{f} , \quad (7.6a)$$

$$div \mathbf{u} = 0 . \quad (7.6b)$$

The  $\theta$  method applied to (7.6a-7.6b) reads

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \theta \left( -\frac{1}{Re} \Delta \mathbf{u}^{n+1} + \mathbf{u}^{n+1} \cdot \nabla \mathbf{u}^{n+1} + \nabla p^{n+1} \right) \\ + (1 - \theta) \left( -\frac{1}{Re} \Delta \mathbf{u}^n + \mathbf{u}^n \cdot \nabla \mathbf{u}^n + \nabla p^n \right) = \theta \mathbf{f}^{n+1} + (1 - \theta) \mathbf{f}^n , \end{aligned} \quad (7.7a)$$

$$\operatorname{div} \mathbf{u}^{n+1} = 0 . \quad (7.7b)$$

Here,  $n$  denotes the old time level and  $n + 1$  the new time level. In the first step of the algorithm, the momentum equation is solved using  $p$  at the old time level. This yields an intermediate velocity field  $\mathbf{u}^*$  satisfying

$$\begin{aligned} \frac{\mathbf{u}^* - \mathbf{u}^n}{\Delta t} + \theta \left( -\frac{1}{Re} \Delta \mathbf{u}^* + \mathbf{u}^* \cdot \nabla \mathbf{u}^* \right) + (1 - \theta) \left( -\frac{1}{Re} \Delta \mathbf{u}^n + \mathbf{u}^n \cdot \nabla \mathbf{u}^n \right) + \nabla p^n \\ = \theta \mathbf{f}^{n+1} + (1 - \theta) \mathbf{f}^n . \end{aligned} \quad (7.8)$$

$\mathbf{u}^*$  is provided with the boundary conditions at level  $n + 1$ . In order to solve (7.8) of course the term  $\mathbf{u}^* \cdot \Delta \mathbf{u}^*$  must be linearized with respect to the old solution  $\mathbf{u}^n$ .

Subtraction of (7.8) from (7.7a) gives

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^*}{\Delta t} + \theta \left( -\frac{1}{Re} \Delta \mathbf{u}^{n+1} + \mathbf{u}^{n+1} \cdot \nabla \mathbf{u}^{n+1} + \frac{1}{Re} \Delta \mathbf{u}^* - \mathbf{u}^* \cdot \nabla \mathbf{u}^* \right) \\ + \theta \nabla (p^{n+1} - p^n) = 0 . \end{aligned} \quad (7.9)$$

It can be shown that the second term of (7.9) is of the same order as the truncation error of the method and hence may be neglected. As a consequence (7.9) reduces to

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^*}{\Delta t} + \theta \nabla (p^{n+1} - p^n) = 0 . \quad (7.10)$$

In the second step  $\mathbf{u}^*$  is projected onto the space of divergence-free vector fields by applying the divergence operator to (7.10):

$$\frac{\operatorname{div} \mathbf{u}^{n+1} - \operatorname{div} \mathbf{u}^*}{\Delta t} + \theta \operatorname{div} \nabla (p^{n+1} - p^n) = 0 . \quad (7.11)$$

Since  $\operatorname{div} \mathbf{u}^{n+1} = 0$ , (7.11) can be considered as an equation for the pressure difference  $p^{n+1} - p^n$ :

$$\Delta (p^{n+1} - p^n) = \frac{\operatorname{div} \mathbf{u}^*}{\theta \Delta t} . \quad (7.12)$$

We have implicitly assumed that  $\theta \neq 0$ ;  $\theta = 0$  requires a slight modification.

Equation (7.12) may be solved by a standard Galerkin method, provided boundary conditions for the pressure are defined along the complete boundary.

Once the pressure correction  $p^{n+1} - p^n$  has been computed,  $p^{n+1}$  follows immediately. Finally  $\mathbf{u}^{n+1}$  may be computed from (7.10).

The pressure-correction method requires the solution of two partial differential equations: (7.8) and (7.12). For both equations the standard Galerkin method may be used. Since no special parameter is introduced, it is possible to solve the resulting systems of linear equations by iterative methods. A clear disadvantage of the continuous pressure-correction method is that it is necessary to define boundary conditions for the pressure. This may be difficult for some types of boundaries and is not natural since the originating Navier-Stokes equations do not require any pressure boundary conditions at all. This problem does not appear in the so-called discrete pressure-correction method, in which first the space discretization is applied and afterwards the pressure correction.

## Discrete approach

In the discrete approach we start with the discrete equations (7.3a-7.3b). For simplicity we consider only the Stokes equations. The extension to Navier-Stokes is straight-forward. So we start with:

$$M\dot{\mathbf{u}} + \mathbf{S}\mathbf{u} - \mathbf{L}^T\mathbf{p} = \mathbf{F} , \quad (7.13a)$$

$$\mathbf{L}\mathbf{u} = \mathbf{0} . \quad (7.13b)$$

Application of the  $\theta$  method to (7.13a) gives

$$M\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \theta(\mathbf{S}\mathbf{u}^{n+1} - \mathbf{L}^T\mathbf{p}^{n+1}) + (1-\theta)(\mathbf{S}\mathbf{u}^n - \mathbf{L}^T\mathbf{p}^n) = \theta\mathbf{F}^{n+1} + (1-\theta)\mathbf{F}^n , \quad (7.14a)$$

$$\mathbf{L}\mathbf{u}^{n+1} = \mathbf{0} . \quad (7.14b)$$

Now the momentum equation (7.14a) is solved with the pressure at the old level. Hence

$$M\frac{\mathbf{u}^* - \mathbf{u}}{\Delta t} + \theta\mathbf{S}\mathbf{u}^* + (1-\theta)\mathbf{S}\mathbf{u}^n - \mathbf{L}^T\mathbf{p}^n = \theta\mathbf{F}^{n+1} + (1-\theta)\mathbf{F}^n \quad (7.15)$$

Subtraction of (7.15) from (7.14a), and neglecting the difference of the viscous terms gives:

$$M\frac{\mathbf{u}^{n+1} - \mathbf{u}^*}{\Delta t} - \theta\mathbf{L}^T(\mathbf{p}^{n+1} - \mathbf{p}^n) = \mathbf{0} . \quad (7.16)$$

In order to apply the continuity equation (7.14b), it is necessary to premultiply (7.16) by  $M^{-1}$ . Then the pressure-correction step becomes:

$$\theta\mathbf{L}M^{-1}\mathbf{L}^T(\mathbf{p}^{n+1} - \mathbf{p}^n) = -\frac{\mathbf{L}\mathbf{u}^*}{\Delta t} . \quad (7.17)$$

In order to solve this equation in a simple way it is necessary that the matrix  $M$  is a diagonal matrix and furthermore that the matrix  $\mathbf{L}\mathbf{L}^T$  can be constructed in an easy way. In practice this is a problem, since the structure of the matrix  $\mathbf{L}\mathbf{L}^T$  is in general different from the structure of a standard Laplacian matrix. Only for the discontinuous pressure elements, both structures are the same. Efficient solution of (7.17) is still a research subject.

Once (7.17) has been solved  $\mathbf{p}^{n+1}$  can be computed, and finally  $\mathbf{u}^{n+1}$  from (7.16).



## A Derivation of the integration by parts for the momentum equations

In this appendix we shall prove the relation

$$\int_{\Omega} (-div \boldsymbol{\sigma} \cdot \mathbf{v}) d\Omega = \int_{\Omega} \boldsymbol{\sigma} \cdot \nabla \mathbf{v} - \int_{\Gamma} \sigma^{nn} v_n + \sigma^{nt} v_t d\Gamma . \quad (\text{A.1})$$

Proof:

If we substitute  $\mathbf{w} = v\mathbf{u}$  in the Gauss divergence theorem

$$\int_{\Omega} div \mathbf{w} d\Omega = \int_{\Gamma} \mathbf{w} \cdot \mathbf{n} d\Gamma , \quad (\text{A.2})$$

and use the relation

$$div \mathbf{w} = v div \mathbf{u} + \mathbf{u} \cdot \nabla v , \quad (\text{A.3})$$

we get

$$- \int_{\Omega} v div \mathbf{u} d\Omega = \int_{\Omega} \mathbf{u} \cdot \nabla v d\Omega - \int_{\Gamma} v \mathbf{u} \cdot \mathbf{n} d\Gamma . \quad (\text{A.4})$$

Writing  $div \boldsymbol{\sigma} \cdot \mathbf{v}$  in components (we restrict ourselves to 2D), we get by applying (A.4) to the left-hand side of (A.1)

$$\begin{aligned} & \int_{\Omega} - \left( div \begin{pmatrix} \sigma_{11} \\ \sigma_{12} \end{pmatrix} v_1 + div \begin{pmatrix} \sigma_{21} \\ \sigma_{22} \end{pmatrix} v_2 \right) d\Omega = \\ & \int_{\Omega} \boldsymbol{\sigma} \cdot \nabla \mathbf{v} d\Omega - \int_{\Gamma} v_1 \begin{pmatrix} \sigma_{11} \\ \sigma_{12} \end{pmatrix} \mathbf{n} + v_2 \begin{pmatrix} \sigma_{21} \\ \sigma_{22} \end{pmatrix} \mathbf{n} d\Gamma . \end{aligned} \quad (\text{A.5})$$

So it remains to prove that the boundary integrals in (A.1) and (A.5) are equal.

The integrand in the boundary integral in (A.5) can be written as

$$\begin{aligned} \boldsymbol{\sigma} \cdot \mathbf{v} \cdot \mathbf{n} &= \mathbf{n} \cdot \boldsymbol{\sigma} \cdot \mathbf{v} = \mathbf{n} \cdot \boldsymbol{\sigma} (v_n \mathbf{n} + v_t \mathbf{t}) \\ &= \mathbf{n} \cdot \boldsymbol{\sigma} \cdot \mathbf{n} v_n + \mathbf{n} \cdot \boldsymbol{\sigma} \cdot \mathbf{t} v_t = \sigma^{nn} v_n + \sigma^{nt} v_t \end{aligned} \quad (\text{A.6})$$

## References

1. A.N. Brooks and T.J.R. Hughes. Stream-line upwind/Petrov Galerkin formulation for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equation. *Comp. Meth. Appl. Mech. Eng.*, 32:199–259, 1982.
2. C. Cuvelier, A. Segal, and A.A. van Steenhoven. *Finite Element Methods and Navier-Stokes Equations*. Reidel Publishing Company, Dordrecht, Holland, 1986.
3. M. Fortin. Old and new finite elements for incompressible flows. *Int. J. Num. Meth. in Fluids*, 1:347–364, 1981.
4. A. George and J.W.H. Liu. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Engelwood Cliffs, New Jersey, (USA), 1981.
5. T.J.R. Hughes. *The Finite Element Method, Linear Static and Dynamic Finite Element Analysis*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1987.
6. T.J.R. Hughes, M. Mallet, and M. Mizukami. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Comp. Meth. Appl. Mech. Eng.*, 54:341–355, 1986.
7. D.S. Malkus and T.J.R. Hughes. Mixed finite element methods-reduced and selective integration techniques: a unification of concepts. *Comp. Meth. Appl. Mech. Eng.*, 15:63–81, 1978.
8. A. Mizukami. An implementation of the stream-line upwind, Petrov-Galerkin method for linear triangular elements. *Comp. Meth. Appl. Mech. Eng.*, 49:357–364, 1985.
9. J.G. Rice and R.J. Schnipke. A monotone streamline upwind finite element method for convection-dominated flows. *Comp. Meth. Appl. Mech. Eng.*, 48:313–327, 1984.
10. R.L. Sani, P.M. Gresho, R.L. Lee, and D.F. Griffiths. The cause and cure (?) of the spurious pressure generated by certain FEM solutions of the incompressible Navier-Stokes equations. *Int. J. Num. Meth. in Fluids. Part I and Part II*, 1 and 1:17–43 and 171–204, 1981.
11. A. Segal. On the numerical solution of the Stokes equations using the finite element method. *Comp. Meth. Appl. Mech. Eng.*, 19:165–185, 1979.
12. F. Shabib. *Finite element analysis of the compressible Euler and Navier-Stokes equations*. PhD thesis, Dept. of Mech. Engng., Stanford University, USA, Stanford, California, USA, 1988.
13. G. Strang and G.J. Fix. *An Analysis of the Finite Element Method*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1973.
14. C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Comput. Fluids*, 1:73–100, 1973.
15. F.N. Van de Vosse. *Numerical analysis of carotid artery flow*. PhD thesis, Eindhoven University of Technology, The Netherlands, 1987.

16. J.J.I.M. van Kan. A second-order accurate pressure correction method for viscous incompressible flow. *SIAM J. Sci. Stat. Comp.*, 7:870–891, 1986.