UNIVERSITAT POLITÈCNICA DE CATALUNYA

A FINITE ELEMENT MODEL FOR INCOMPRESSIBLE FLOW PROBLEMS

by

RAMON CODINA I ROVIRA

UNIVERSITAT POLITÈCNICA DE CATALUNYA

A FINITE ELEMENT MODEL FOR INCOMPRESSIBLE FLOW PROBLEMS

by

RAMON CODINA I ROVIRA

DOCTORAL THESIS

Barcelona, June 1992

Als meus pares
i a la meva germana

CONTENTS

FOREWORD

PART I: THE CONVECTION–DIFFUSION EQUATION

CHAPTER 1

THE STREAMLINE DIFFUSION METHOD FOR THE STEADY–STATE PROBLEM

1.1 INTRODUCTION AND MOTIVATION 1.1

1.1.1 The standard Galerkin method and a first approach to its instability problems ... 1.2

1.1.2 Artificial diffusion and the former Petrov-Galerkin methods 1.3

1.1.3 Multidimensional case: The Streamline Diffusion method 1.6

1.1.4 The Galerkin/least-squares method 1.9

1.2 CONVERGENCE ANALYSIS 1.10

1.2.1 Introduction and variational problem 1.10

1.2.2 Interpolation estimates 1.11

1.2.3 Error analysis 1.11

1.3 THE OPTIMAL UPWIND FUNCTIONS FOR ONE-DIMENSIONAL QUADRATIC ELEMENTS 1.14

1.3.1 General considerations 1.14

1.3.2 Standard formulation of the SD method 1.16

1.3.3 Hierarchic formulation of the SD method 1.20

1.3.4 Standard formulation of the GLS method 1.22

1.3.5 Introduction of source terms 1.24

1.4 NUMERICAL IMPLEMENTATION 1.26

1.4.1 The characteristic length 1.26

1.4.2 Assignment of upwind functions 1.27

1.5 NUMERICAL EXAMPLES 1.29

1.6 SUMMARY AND CONCLUSIONS 1.33

REFERENCES 1.36

CHAPTER 2

TRANSIENT ALGORITHMS—STABILITY ANALYSIS OF AN EXPLICIT SCHEME

2.1 INTRODUCTION 2.1

2.2 THE GENERALIZED TRAPEZOIDAL RULE 2.3

2.2.1 The continuous problem 2.3

2.2.2 Discretization in space and time 2.4

2.3 STABILITY ANALYSIS OF THE FORWARD EULER SCHEME 2.7

2.3.1 General considerations 2.7

2.3.2 Linear elements 2.10

Stability and accuracy 2.10

Remarks on the Taylor-Galerkin method 2.13

Algorithmic damping ration (ADR) and frequency ratio (AFR) 2.13

2.3.3 Quadratic elements I: canonical basis 2.16

Stability and accuracy 2.17

Algorithmic damping ration (ADR) and frequency ratio (AFR) 2.21

2.3.4 Quadratic elements II: hierarchic approach 2.23

Diagonalization of $\mathbf{M}^e$ 2.23

Stability 2.24

2.3.5 Extension to multidimensional problems 2.28

2.4 NUMERICAL EXAMPLES 2.29

2.5 SUMMARY AND CONCLUSIONS 2.36

REFERENCES 2.39

PART II: THE INCOMPRESSIBLE NAVIER-STOKES EQUATIONS

CHAPTER 3

PENALTY FINITE ELEMENT METHODS FOR THE STATIONARY NAVIER–STOKES EQUATIONS

3.1 INTRODUCTION 3.1

3.2 STATEMENT OF THE PROBLEM AND PENALTY METHODS 3.3

3.2.1 The continuous problem 3.3

3.2.2 Penalty methods for the Stokes problem 3.5

3.2.3 Finite element discretization 3.7

The discrete problem 3.7

Some finite element spaces 3.8

Strong penalization and RIP methods 3.9

Matrix formulation 3.11

Some bibliographical notes on the BB condition 3.13

3.3 AN EXAMPLE OF STRONG PENALIZATION: THE BIQUADRATIC ELEMENT FOR TWO–DIMENSIONAL INCOMPRESSIBLE FLOWS 3.14

3.3.1 Gauss-Legendre quadrature for the volumetric term 3.14

3.3.2 About the (im)possibility of emulating the Q_2/P_1 element 3.15

A three-point quadrature rule 3.16

Inconsistency of $b_*(\cdot,\cdot)$ 3.18

3.3.3 Pressure calculation 3.20

Pressure filter and Least-squares smoothing (LSS) 3.20

Pressure Poisson equation 3.21

Numerical performance 3.23

3.3.4 Numerical procedures for the Navier-Stokes equations 3.24

Linearization technique 3.24

Continuation method 3.24

Petrov-Galerkin weighting 3.24

Final algorithm 3.25

Numerical experiments 3.26

3.4 WEAK PENALIZATION: ANALYSIS OF AN ITERATIVE PENALTY METHOD 3.27

3.4.1 Iterative penalization for the Stokes problem 3.29

Motivation and statement of the algorithm 3.29

Some remarks on related methods 3.30

Convergence of the algorithm 3.30

3.4.2 A Picard-based iterative algorithm for the Navier-Stokes equations 3.33

3.4.3 A Newton-Raphson-based iterative algorithm for the Navier-Stokes equations ... 3.37

3.4.4 *Uncoupling of the iterative penalization* 3.41

3.5 NUMERICAL EXAMPLES 3.43

3.6 SUMMARY AND CONCLUSIONS 3.48

REFERENCES 3.52

CHAPTER 4

TRANSIENT NAVIER-STOKES EQUATIONS: FULLY DISCRETE ALGORITHM AND COMPUTATIONAL ASPECTS

4.1 INTRODUCTION 4.1

4.2 THE CONTINUOUS PROBLEM 4.2

4.3 DISCRETIZATION IN TIME 4.5

4.4 SPACE DISCRETIZATION AND STREAMLINE DIFFUSION OPERATOR 4.8

4.4.1 Galerkin approach and finite element spaces 4.8

The semidiscrete problem 4.8

Finite element spaces 4.8

Fully discrete problem 4.12

Convergence results 4.13

4.4.2 Streamline Diffusion operator 4.15

Semidiscrete problem 4.15

Fully discrete problem 4.16

Definition of the intrinsic time 4.17

Some remarks about least-squares techniques 4.18

4.5 LINEARIZED EQUATIONS AND PENALTY METHODS 4.19

4.5.1 Linearization of the convective term and the Streamline Diffusion operator 4.19

4.5.2 Penalty methods 4.20

4.5.3 Fully discrete and linearized algorithm 4.21

4.6 MATRIX FORMULATION 4.22

4.7 COMPUTING SECONDARY VARIABLES 4.26

4.7.1 Least-squares smoothing 4.26

4.7.2 Nodal quadrature rules 4.27

4.7.3 Pressure, vorticity and physical properties smoothing 4.27

4.7.4 An algorithm for the calculation of the streamfunction 4.30

4.8 NUMERICAL RESULTS 4.33

4.9 SUMMARY AND CONCLUSIONS 4.56

REFERENCES 4.61

PART III: APPLICATIONS

CHAPTER 5

THERMALLY COUPLED FLOWS AND NONLINEAR MATERIALS

5.1 INTRODUCTION 5.1

5.2 THE BOUSSINESQ MODEL 5.2

5.2.1 The continuous problem 5.2

5.2.2 Discretization in space and time 5.5

5.2.3 Block iterative algorithm 5.8

5.3 CREEPING FLOWS OF NONLINEAR MATERIALS 5.13

5.3.1 Generalized Newtonian fluids 5.13

5.3.2 Stationary problem and finite element discretization 5.16

5.3.3 Iterative techniques 5.18

5.4 GENERAL PROBLEM—ITERATIVE PROCEDURE 5.20

5.4.1 Motivation 5.20

5.4.2 Time discretization 5.21

5.4.3 Fully discrete and linearized problem 5.22

5.5 SOME APPLICATIONS OF THE NUMERICAL METHOD 5.24

5.5.1 Thermoconvective instability of plane Poiseuille flow 5.24

5.5.2 Transient natural convection of low-Prandtl-number fluids 5.37

5.5.3 The 4:1 plane extrussion of a power-law fluid 5.47

5.6 SUMMARY AND CONCLUSIONS 5.56

REFERENCES 5.57

CHAPTER 6

MOULD FILLING SIMULATION

6.1 INTRODUCTION 6.1

6.2 THE PSEUDO-CONCENTRATION METHOD 6.2

6.2.1 Basic formulation 6.2

6.2.2 Numerical solution of the pseudo-concentration problem 6.3

6.3 SOME NUMERICAL TECHNIQUES 6.4

6.3.1 General considerations 6.4

6.3.2 Smoothing of the pseudo-concentration surface 6.5

6.3.3 Air release—Introduction of holes 6.8

6.4 THE NAVIER-STOKES EQUATIONS WITH A MOVING FREE SURFACE 6.10

6.4.1 Statement of the problem 6.10

6.4.2 Numerical treatment 6.12

6.5 APPLICATION TO SOME PRACTICAL PROBLEMS 6.14

6.5.1 Mould filling by gravity 6.15

6.5.2 Injection mould filling 6.20

6.5.3 Hot rolling of a rectangular slab 6.30

6.6 SUMMARY AND CONCLUSIONS 6.40

REFERENCES 6.41

AFTERWORD

FOREWORD

INTRODUCTION AND OBJECTIVES

Although the numerical simulation of flow problems began in the sixties using finite difference or panel methods, it wasn't until the early seventies that the Finite Element Method (FEM) entered the field of computational fluid dynamics (CFD). Since then, a lot of progress has been made, both in the understanding of the difficulties lying on the application of the general finite element ideas and in the development of numerical strategies to overcome them.

This work deals with FEMs to solve viscous incompressible flow problems, an important branch of CFD whose applications are widespread in many areas of engineering and science. Numerical techniques are likely to become a serious competitor to experimentation because of their reliability and their reduced cost, and also because in some areas experiments are very difficult to make. Industries are aware of this fact and the evolution of their numerical budget reflects this interest. Nevertheless, the numerical simulation of complex real-world flow problems for many practical engineering applications lies still far in the future, not only because of the present knowledge on numerical methods but also because of today's computer facilities and capabilities. Flow problems are extremely demanding in what concerns numerical computations and the present computer technology cannot supply all the computational power that would be needed to solve many real flow problems.

All the terms of the incompressible Navier-Stokes involve a more or less important numerical difficulty and a lot of questions are still open. Temporal derivatives are usually dealt with using finite differences, in spite of the fact that the solution may develop high and quick variations in time and instability problems or lack of accuracy may be encountered. The incompressibility constraint, closely related to the presence of pressure forces, is another of the most important problems. Doubtless, the nonlinear convective term is the reason why the Navier-Stokes equations are so difficult to solve numerically and to analyse mathematically. Maybe the only 'nice' term is viscous one, which gives a parabolic character to the transient equations.

The purpose of this work is twofold. First, new numerical methods are developed to treat two of the problems mentioned above, namely, the incompressibility constraint and the instability problems found when the standard Galerkin approach is applied to convection dominated flows. For this last case, though, a laminar behavior is considered, that is, without the introduction of turbulence models. The second objective is to develop a general purpose finite element code, implementing the new techniques presented here and incorporating several computational features also original from this work. This general finite element model includes the numerical solution of the incompressible Navier-Stokes equations together with the energy balance equation and the tracking of free surfaces. Applications to thermally coupled flows and the flow of nonlinear materials are provided.

OVERVIEW

The first part of this work deals with the numerical solution of the convection-diffusion equation as a model to study convection-dominated flows. The finite element formulation employed here is the Streamline Diffusion method. A thorough description of this method is presented in Chapter 1, where several extensions are introduced, such as the use of quadratic elements, the computation of the algorithmic parameters of this formulation and a particular version of its convergence analysis. Chapter 2 is concerned with the application of the generalized trapezoidal rule to advance in time for the transient equation. A complete stability and accuracy analysis is performed for the explicit Euler scheme, both using linear and quadratic finite elements.

Part II deals with the incompressible Navier-Stokes equations. The stationary equation is considered in Chapter 3, where the treatment of the incompressibility constraint is studied in detail. Standard penalty methods are described and a new iterative method is introduced and fully analysed. In Chapter 4 the emphasis moves from the mathematically-oriented exposition to computational aspects. The transient Navier-Stokes equations are studied there, combining the ideas of the previous chapters and also introducing several numerical facilities.

The complete numerical model, including thermal coupling and free surface tracking, is presented in Part III in connexion with several physical and engineering applications. Thermally driven flows and flows of nonlinear materials, with application to metal forming processes, are the subject of Chapter 5, and the numerical simulation of mould filling is presented in Chapter 6.

The motivation for the specific subjects treated here, as well as a brief state-of-the-art, are given in the introduction to each chapter. Also, the subjects covered and the conclusions drawn from the work developed are presented at the end of each chapter.

NOTATION

The notation employed in this work is fairly standard in the mathematical literature, although perhaps not vey common in engineering circles. As far as possible, the matrix version of the abstract formulation of the problems studied is given, in particular for presenting the basic flow chart of transient and iterative algorithms.

Apart from a few exceptions, matrices and vectors are denoted by boldface characters and scalars by lightface italic characters. Cartesian notation is used when referring to a particular coordinate system, denoting by (x_1, x_2, x_3) or (x, y, z) the Cartesian coordinates for the three-dimensional case.

The space of square integrable functions over a domain ω of the Euclidian space has been denoted by $L^2(\omega)$, and the inner product in this space by $(\cdot, \cdot)_\omega$. The norm associated to this inner product has been indicated by

$$\|\cdot\|_\omega \equiv \|\cdot\|_{L^2(\omega)} \equiv \|\cdot\|_{0,\omega}$$

The subscript ω is often dropped if this is the domain where the problem is to be solved, always denoted by Ω.

The Sobolev space of functions whose (distributional) derivatives of order up to m belong to $L^2(\omega)$ has been denoted by $H^m(\omega)$. The space $H_0^1(\omega)$ consists of functions of $H^1(\omega)$ with zero trace on the boundary $\partial\omega$. Any of the symbols

$$\|\cdot\|_{m,\omega} \equiv \|\cdot\|_{H^m(\omega)}$$

has been employed to denote the norm of these spaces, although no subscript at all has been used when no confusion is possible. In general, the norm of a space V has been denoted by $\|\cdot\|_V$ and the Euclidian norm of a vector by $|\cdot|$.

The classical gradient, divergence, curl and Laplacian operators have been denoted by

$$\nabla(\cdot), \quad \nabla\cdot(\cdot), \quad \nabla\times(\cdot) \quad \text{and} \quad \Delta(\cdot),$$

respectively. The symbol Δt has been used for the time step size, not for the Laplacian of t. For the temporal derivative and for the partial derivative with respect to a Cartesian coordinate x_i any of the symbols

$$\partial_t \equiv \frac{\partial}{\partial t}, \qquad \partial_i \equiv \frac{\partial}{\partial x_i}$$

has been used.

Various integers have been employed in the text. Some of them are

N_{sd} :	number of space dimensions
N_{el} :	number of elements of the discretization
N_{no} :	number of nodes per element
N_{gp} :	number of integration points per element
N_{tp} :	total number of nodes of the finite element mesh
N_{fp} :	number of free points (without Dirichlet conditions)
N_{nv} :	number of nodes per element with velocity unknowns
N_{qp} :	number of nodes per element with pressure unknowns
N_{vu} :	number of velocity unknowns ($= N_{fp} \times N_{sd}$)
N_{pu} :	number of pressure unknowns ($= N_{el} \times N_{qp}$)
N :	number of time steps

A generic shape function has been denoted by N (not to be confused with the number of time steps), perhaps with a superscript to indicate the node to which it is associated.

The symbol $\{\Omega^e\}$, $e = 1, ..., N_{el}$, has been used to denote a finite element partition of the domain Ω. It is understood that the subdomains Ω^e are open, nonoverlapping and the union of their closures is the closure of Ω. A function beloging to the finite element space is recognized by the subscript h, the diameter of $\{\Omega^e\}$. Vectors of nodal unknowns are denoted by the boldface capital letter corresponding to the lower case variable.

The rest of the notation is explained in the text.

ACKNOWLEDGEMENTS

I would like to take the chance to express my appreciation to many people whose support and stimulation have made this work possible. In particular, I am indebted to my thesis director, Prof. Eugenio Oñate, for encouraging my research in this field. I would like also to thank some of my colleagues and friends for their help: Dr. Miguel Cervera, who collaborated with me in preparing the basic structure of the code written for this thesis; Uwe Schäfer, who had an important participation in the research presented in the last chapter; Dr. Gabriel Bugeda, who provided the codes for mesh generation; Javier Castro and Prof. Juan Miquel, for fruitful discussions; Knut Eckstein and Orlando Soto, for their assistence in preparing some numerical results; and Mario Galindo, for his continuous help in using computer facilities.

This research has benefited from financial support from the CIRIT (Generalitat de Catalunya), the Universitat Politècnica de Catalunya and a grant FPI from the Spanish Government. Partial support from Project BRITE, Ref. BREU-CT-91-0443, and Programme HERMES, Ref. 08W01581, are also gratefully acknowledged.

Finally, I want to express my deepest gratitude to the lovely people to whom this work is dedicated.

PART I:

THE CONVECTION–DIFFUSION EQUATION

CHAPTER 1

THE STREAMLINE DIFFUSION METHOD FOR THE STEADY–STATE PROBLEM

1.1 Introduction and motivation

Besides the interest of the convection-diffusion equation as a mathematical model for several physical phenomena, it also represents a good model for the development of numerical methods for the approximate solution of more complicated transport equations. When the convective terms of these equations become important the standard Galerkin formulation fails and numerical oscillations occur. These oscillations can only be avoided after a drastic refinement of the finite element mesh. The lack of stability that the Galerkin formulation shows in those cases is the common explanation for the nonphysical behavior of the numerical solution, although we will see that an examination of the analytical solution of the discrete equations obtained for the one-dimensional convection-diffusion equation shows the same problem.

Several numerical methods have been introduced in order to overcome this misbehavior. The purpose of this chapter is to present one of them, introduced by Hughes & Brooks [BH], [HB1], [HB2] under the acronym SUPG: Streamline Upwind/Petrov-Galerkin. Almost simultaneously, the mathematical analysis of the method was undertaken by Johnson & Nävert [Jo1], [Na], [JNP], who preferred the name 'Streamline diffusion' (just SD, from now onwards). Nowadays, the use of this method has become widespread and the name SD seems to prevail over SUPG, especially in mathematical circles. For this reason, the former option will also be used in this work (see [Hu2] for further discussion).

As it happens for any numerical method, it is not fair to give all the credit to the authors mentioned above. This first section tries to draw a schematic evolution to the SD method and also to mention some other existing methods. The starting point will be looking at the Galerkin solution for the 1D steady-state equation using linear elements. What happens in this very simple case gives the clue for the development of any numerical remedy, both the simplest and the most elegant. Extensions to other problems (multidimensional, transient, other transport equations) are particular of each approach. We will only concentrate on the SD method for the steady-state problem. The transient equation will be adressed in the next chapter. The problem of removing the localized oscillations that still remain using the SD method will not be treated in this work.

1.1.1 The standard Galerkin method and a first approach to its instability problems

The motivation of the finite element formulation that will be used throughout this work can be found in the one-dimensional, stationary and homogeneous convection-diffusion problem with Dirichlet boundary conditions: Find a function $\phi = \phi(x)$ such that

$$u\frac{d\phi}{dx} - k\frac{d^2\phi}{dx^2} = 0, \quad 0 < x < \ell \tag{1.1}$$

$$\phi(0) = \phi_0, \quad \phi(\ell) = \phi_\ell \tag{1.2}$$

where $k > 0$ and u are constants (having the physical meaning of diffusion and velocity, respectively) and ϕ_0, ϕ_ℓ are the prescribed values of ϕ on the boundary. The fact that we consider the boundary conditions (1.2) is not any restriction for what follows.

Let $0 = x_0 < x_1 < ... < x_N = \ell$ be a uniform partition of the interval $[0, \ell]$, with $x_{m+1} - x_m = h$, $m = 0, ..., N-1$. Let us call $\gamma := uh/2k$ the *element Péclet number* of (1.1) for this partition. This dimensionless number gives an idea of the relative importance of convection and diffusion. Convection will be dominant when $|\gamma|$ is large, whereas diffusive effects will predominate for small values of $|\gamma|$. In the former case, an examination of the analytical solution of problem (1.1)–(1.2) reveals that boundary layers develop near $x = \ell$ if $\gamma > 0$ and near $x = 0$ if $\gamma < 0$, that is, according to the sign of the velocity u. The function ϕ will be very steep in these zones and numerical problems can be anticipated if one tries to approximate it with a few discretization points.

If problem (1.1)–(1.2) is solved numerically using linear finite elements and the standard Galerkin method, the following difference equations are found:

$$(1-\gamma)\phi_{m+1} - 2\phi_m + (1+\gamma)\phi_{m-1} = 0, \quad m = 1, ..., N-1 \tag{1.3}$$

where ϕ_m is the nodal unknown at the point m and ϕ_0, ϕ_N are given by the boundary conditions (1.2). The same system of equations is found if instead of using the finite element method (FEM, for short) one uses finite differences with a centered approximation for both the first and the second derivatives:

$$\begin{aligned} \left.\frac{d\phi}{dx}\right|_m &\approx \frac{\phi_{m+1} - \phi_{m-1}}{2h} \\ \left.\frac{d^2\phi}{dx^2}\right|_m &\approx \frac{\phi_{m+1} - 2\phi_m + \phi_{m-1}}{h^2} \end{aligned} \tag{1.4}$$

If the exact solution of problem (1.1)–(1.2) is introduced in equations (1.3) and is expanded in Taylor series, one finds that (cf. [Co]):

$$\left(u\frac{d\phi}{dx} - k\frac{d^2\phi}{dx^2}\right)\Big|_m + k^* \left.\frac{d^2\phi}{dx^2}\right|_m = 0 \tag{1.5}$$

where

$$k^* = -\frac{k}{2\gamma}\left[\frac{1}{\gamma}(\cosh(2\gamma) - 1) - \sinh(2\gamma)\right] \tag{1.6}$$

that is, the truncation error of the scheme (1.3) is

$$E_\tau = k^* \left.\frac{d^2\phi}{dx^2}\right|_m \tag{1.7}$$

It is easy to prove (cf. [Co]) that $k^* \to 0$ when $\gamma \to 0$ and that $\text{sgn}(k^*) = \text{sgn}(k)$. From Eqn. (1.5) we see that the scheme (1.3) gives nodally exact solutions at the nodes for the modified equation

$$u\frac{d\phi}{dx} - (k - k^*)\frac{d^2\phi}{dx^2} = 0 \tag{1.8}$$

Since $k - k^* < k$, we have a first explanation for the failure of the Galerkin method: *it solves exactly an underdiffusive equation, or equivalently, it introduces an artificial negative diffusion.* Thus, spurious oscillations can be expected when $\gamma \to \infty$, since it can be shown from the expression (1.6) of k^* that in this case $k^* \to \infty$.

A different aproach is to solve exactly the difference equations (1.3) (see, e.g., Reference [IK] for background). The characteristic equation of (1.3) is $(1-\gamma)\lambda^2 - 2\lambda + (1+\gamma) = 0$. Since the roots of this equation are $\lambda = 1$ and $\lambda = (1+\gamma)(1-\gamma)^{-1}$, the exact solution of Eqns. (1.3) is given by

$$\phi_m = C_1 + C_2\left(\frac{1+\gamma}{1-\gamma}\right)^m \tag{1.9}$$

where C_1 and C_2 are constants fixed by the boundary conditions. From Eqn. (1.9) it is clear that *oscillations will be found whenever* $|\gamma| > 1$.

A third method to justify the wrong behavior of the Galerkin method is to compute the eigenvalues of the system (1.3) [Ba]. One can prove (cf. [Pi]) that for $\gamma \neq 0$ there exists a multiple eigenvalue given by $\lambda = 2\gamma^{-1}$. Thus, the matrix associated to the system (1.3), say A, is nearly singular when $\gamma \to \infty$ and so *this system is unstable when γ is very large.* This point of view is related to the analysis of the transient problem [GP]. The almost singularity of the matrix A indicates that the semidiscrete problem, that will have the form $\dot{x} + Ax = 0$ (the dot denoting the temporal derivative), will be *structurally unstable* in convection dominated problems.

1.1.2 Artificial diffusion and the former Petrov-Galerkin methods

From the previous discussion it is clear that *any method* whose goal be the elimination of the instability problems of the Galerkin formulation must introduce, in one way or another, an artificial dissipation. The crudest approach is just to add a diffusion in the original continuous equation and then to use the Galerkin approach for this modified equation. Although this idea is old and goes back to the early finite difference methods, a pioneering work oriented to the justification of this method in the context of finite element methods was done by Kikuchi [Ki], who considered the introduction of artificial diffusion as a way to satisfy the discrete maximum principle.

In the finite difference literature, the idea of adding numerical dissipation was first introduced (cf. [Ro]) by von Neumann and Richtmyer [NRi] and it was early recognized that this dissipation could be introduced by means of a non-centered difference approximation for the first derivatives taking into account the direction of the flow, i.e., the sign of u in Eqn. (1.1). This fact motivated the name *upwind methods* for the numerical formulations based on a modification of centered schemes according to the flow direction. We will see how this can be done in an accurate manner. For the analysis of this method, the reader is referred to the classical book of Richtmyer and Morton [RM].

The introduction of artificial diffusion designed in order to obtain an accurate solution will be the seed of the SD method and will allow us to introduce the concept

of upwind function. Let k' be a numerical dissipation of the form

$$k' = \alpha \frac{uh}{2} \tag{1.10}$$

where α is a function of the Péclet number γ to be determined and that will be called *upwind function.* Now, if in Eqn. (1.1) the diffusion k' is added to the 'real' diffusion k and, as before, the standard Galerkin method is applied (or the finite difference approximations (1.4) are used) the following equations are found instead of Eqns. (1.3):

$$[1+\gamma(\alpha-1)]\phi_{m+1} - 2(1+\alpha\gamma)\phi_m + [1+\gamma(\alpha+1)]\phi_{m-1} = 0 \tag{1.11}$$

and the resulting truncation error is

$$E_\tau = -\frac{k}{2\gamma}\left[(\frac{1}{\gamma}+\alpha)(\cosh(2\gamma)-1) - \sinh(2\gamma)\right]\left.\frac{d^2\phi}{dx^2}\right|_i$$

If one imposes $E_\tau = 0$ the following expression for the function α is found:

$$\alpha = \coth\gamma - \frac{1}{\gamma} \tag{1.12}$$

Since the truncation error is zero for this choice of α, the numerical solution will be nodally exact. The error of the scheme will be exactly the error of the canonical projection of the analytical solution onto the discrete finite element space. For this reason, the function (1.12) is called *optimal.*

Now we come to the main point of this discussion. In the finite difference method, scheme (1.11) is obtained if the second derivatives are approximated by a centered scheme *and the first derivatives by*

$$\left.\frac{d\phi}{dx}\right|_m \approx \frac{(1-\alpha)\phi_{m+1} + 2\alpha\phi_m - (1+\alpha)\phi_{m-1}}{2h} \tag{1.13}$$

If finite elements are used, the weak form of problem (1.1)–(1.2) has to be introduced. Multiplying Eqn. (1.1) by a suitable test function ψ (with $\psi(0) = \psi(\ell) = 0$) and after integration by parts one gets

$$\begin{aligned} 0 &= \int_0^\ell \psi u\frac{d\phi}{dx}dx + \int_0^\ell \left(k + \alpha\frac{uh}{2}\right)\frac{d\psi}{dx}\frac{d\phi}{dx}dx \\ &= \int_0^\ell \left(\psi + \alpha\frac{h}{2}\frac{d\psi}{dx}\right)u\frac{d\phi}{dx}dx + \int_0^\ell k\frac{d\psi}{dx}\frac{d\phi}{dx}dx \end{aligned} \tag{1.14}$$

from where we see that *scheme (1.11) is obtained if the weighting function*

$$\bar{\psi} := \psi + \alpha\frac{h}{2}\frac{d\psi}{dx} \tag{1.15}$$

is applied only for the convective term. It will be shown later how this inconsistency can be removed. When the space of tests functions is different from the space of trial solutions, as it will happen in this case, the resulting formulation is said to belong to the class of *Petrov-Galerkin methods.*

Remarks 1.1

(1) It is easy to see that the function α given by (1.12) is skew-symmetric and that it verifies $\alpha \to 1$ when $\gamma \to \infty$ and $\alpha = \frac{1}{3}\gamma + O(\gamma^3)$ as $\gamma \to 0$. Hence, a good asymptotic approximation, often used, is

$$\alpha_a(\gamma) = \begin{cases} \frac{\gamma}{3} & \text{if } 0 \le |\gamma| \le 3 \\ \text{sgn}(\gamma) & \text{if } |\gamma| > 3 \end{cases} \tag{1.16}$$

(2) Expression (1.12) was obtained by Christie *et al.* [CGM] using a weighting function different from the one given by (1.15) but that leads to the same scheme (1.11).

(3) If $\alpha = 1$, we observe from (1.13) that the first derivatives are approximated by the backward differences

$$\left.\frac{d\phi}{dx}\right|_m \approx \frac{\phi_m - \phi_{m-1}}{h}$$

and if $\alpha = -1$ by the forward differences

$$\left.\frac{d\phi}{dx}\right|_m \approx \frac{\phi_{m+1} - \phi_m}{h}$$

These approximations for the first derivatives were the starting point for the so called *upwind techniques* in finite differences. It should be remarked that they are only first order approximations. □

The interpretation we have given to scheme (1.11) as a modification of the weighting function for the convective term to the one given by (1.15) is not the only one possible. Christie *et al.* [CGM] interpreted (1.11) through the use of a continuous third order polynomial weighting function modified in order to give more weight upstream of the flow. Hughes [Hu1] obtained (1.11) by using a one-point integration rule for the convective term. An expression similar to (1.15) was first used by Wahlbin [Wa], who considered test functions of the form $\bar{\psi} := \psi + h d\psi/dx$ for a semilinear hyperbolic problem in one dimension. For a good review of early upwind methods, see Reference [HZ2].

The upwind function α given by (1.12) has been obtained using a very stringent requirement: the numerical solution should be nodally exact. One can also try to achieve the more modest goal of avoiding numerical oscillations. For that, consider the analytical solution of (1.11), that is found to be:

$$\phi_m = C_1 + C_2 \left(\frac{1 + \gamma(1+\alpha)}{1 - \gamma(1-\alpha)}\right)^m \tag{1.17}$$

Of course, if this expression is compared with the exact solution $\phi(x)$ of problem (1.1)-(1.2) one finds that $\phi_m = \phi(x_m)$ if, and only if, α is chosen as (1.12) indicates. On the other hand, if one only wishes to preclude the oscillations of the Galerkin method, from (1.17) it is seen that $|\alpha|$ must exceed the critical value:

$$\alpha_c = 1 - \frac{1}{|\gamma|} \tag{1.18}$$

This expression will be found again from a different approach in the next chapter. It can be shown that it is also the value below wich the numerical scheme does not satify

the discrete maximum principle [Ki]. Therefore, there are several reasons for taking $|\alpha| \geq \alpha_c$.

1.1.3 Multidimensional case: the Streamline Diffusion method

Consider first the continuous steady-state convection-diffusion problem. Let Ω be an open bounded domain of $\mathbb{R}^{N_{sd}}$ ($N_{sd} = 2$ or 3) and $\Gamma = \partial\Omega = \overline{\Gamma_D \cup \Gamma_N}$, with $\Gamma_D \cap \Gamma_N = \emptyset$, the empty set. The problem to be solved consists in finding a function $\phi = \phi(\mathbf{x})$ such that

$$\mathbf{u} \cdot \nabla\phi - \nabla \cdot (\mathbf{k} \cdot \nabla\phi) = f, \qquad \text{in } \Omega \tag{1.19}$$

$$\phi = g, \qquad \text{on } \Gamma_D \tag{1.20}$$

$$\mathbf{n} \cdot \mathbf{k} \cdot \nabla\phi = r, \qquad \text{on } \Gamma_N \tag{1.21}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{x})$ is the velocity field, $\mathbf{k} = \mathbf{k}(\mathbf{x})$ is the diffusion tensor, that we assume is symmetric and positive-definite, $f = f(\mathbf{x})$ is the source term, $g = g(\mathbf{x})$ is a prescribed value of ϕ defined on the part of the boundary where Dirichlet conditions are fixed, $\mathbf{n}$ is the unit outward normal to Γ and $r = r(\mathbf{x})$ is a prescribed diffusive flux.

In order to write the weak form of problem (1.19)–(1.21), let us introduce the spaces of test functions Ψ and of trial solutions Φ:

$$\Psi := \{\psi \in H^1(\Omega) \mid \psi = 0 \text{ on } \Gamma_D\} \tag{1.22}$$

$$\Phi := \{\phi \in H^1(\Omega) \mid \phi = g \text{ on } \Gamma_D\} \tag{1.23}$$

Having introduced this notation, the weak form of the problem we consider can be written as follows: Find $\phi \in \Phi$ such that

$$a(\phi, \psi) = l(\psi) \qquad \forall \psi \in \Psi \tag{1.24}$$

where the bilinear form a and the linear form l are

$$a(\phi, \psi) := \int_\Omega (\psi \mathbf{u} \cdot \nabla\phi + \nabla\psi \cdot \mathbf{k} \cdot \nabla\phi)\, d\Omega \tag{1.25}$$

$$l(\psi) := \int_\Omega \psi f \, d\Omega + \int_{\Gamma_N} \psi r \, d\Gamma \tag{1.26}$$

Construct now a finite element discretization $\{\Omega^e\}$ of Ω, with index e ranging from 1 to the number of elements N_{el} and let us consider the discrete finite element spaces

$$\Psi_h := \{\psi \in \Psi \mid \psi|_{\Omega^e} \in P_m(\Omega^e)\} \subset \Psi \tag{1.27}$$

$$\Phi_h := \{\phi \in \Phi \mid \phi|_{\Omega^e} \in P_m(\Omega^e)\} \subset \Phi \tag{1.28}$$

where $P_m(\Omega^e)$ denotes the set of complete polynomials of degree m in Ω^e. The Galerkin method applied to problem (1.19)–(1.21) reads as follows: Find a function $\phi_h \in \Phi_h$ such that

$$a(\phi_h, \psi_h) = l(\psi_h) \qquad \forall \psi_h \in \Psi_h \tag{1.29}$$

Let us define the *element Péclet number* associated to an element e of the partition $\{\Omega^e\}$ by

$$\gamma^e := \frac{|\mathbf{u}^e| h^e}{2k^e} \tag{1.30}$$

where $|\mathbf{u}^e|$ is the Euclidian norm of a characteristic velocity of the element, h^e is a characteristic length and k^e is a characteristic diffusion. The election of these quantities will be discussed later.

The instability problems that method (1.29) has are the same as those encountered for the one-dimensional problem. Therefore, using the same ideas as before, a possible remedy would be to introduce a numerical dissipation in the original (continuous) equation. If this dissipation is isotropic, the numerical results happen to be overdiffusive. In fact, the first attempts to extend some upwind methods that had proved to be successful in 1D problems showed also an excessive crosswind diffusion in multidimensional situations [Hul], [HHZ]. The main idea underlying the SD method is to introduce numerical dissipation *only along the streamlines.* The reason for this is clear if we write Eqn. (1.19) for the 2D problem in orthogonal coordinates (σ, ν), σ being the arc parameter along the streamlines. If, for simplicity, we assume that the real diffusion tensor $\mathbf{k}$ is isotropic, with $\mathbf{k} = k\mathbf{I}$, we will have that:

$$|\mathbf{u}|\frac{\partial \phi}{\partial \sigma} - \frac{\partial}{\partial \sigma}\left(k\frac{\partial \phi}{\partial \sigma}\right) - \frac{\partial}{\partial \nu}\left(k\frac{\partial \phi}{\partial \nu}\right) = f \tag{1.31}$$

from where it follows that *only the diffusion in the $\sigma-$direction has to be balanced with the convection.* This very important idea was introduced almost simultaneously by Kelly *et al.* [KNZ] and by Hughes & Brooks [HB1]. However, it should be remarked that this reasoning is also valid if instead of the velocity field $\mathbf{u}$ we consider another field $\mathbf{v}$ such that

$$\mathbf{u} \cdot \nabla \phi = \mathbf{v} \cdot \nabla \phi \tag{1.32}$$

Assume now that an artificial directional dissipation of magnitude d is added to the real diffusion along the lines tangent to a vector field $\mathbf{v}$ satisfying condition (1.32). This artificial diffusion will be given by

$$\mathbf{k}' = \frac{d}{|\mathbf{v}|^2}\mathbf{v} \otimes \mathbf{v} \tag{1.33}$$

and problem (1.29) will be replaced by: Find a function $\phi_h \in \Phi_h$ such that

$$a_d(\phi_h, \psi_h) = l(\psi_h) \qquad \forall \psi_h \in \Psi_h \tag{1.34}$$

where the bilinear form a_d is

$$\begin{aligned} a_d(\phi_h, \psi_h) :&= \int_\Omega \left[\psi_h \mathbf{u} \cdot \nabla \phi_h + \nabla \psi_h \cdot \mathbf{k} \cdot \nabla \phi_h + \nabla \psi_h \cdot \frac{d}{|\mathbf{v}|^2}(\mathbf{v} \otimes \mathbf{v}) \cdot \nabla \phi_h\right] d\Omega \\ &= \int_\Omega \left[\psi_h \mathbf{u} \cdot \nabla \phi_h + \nabla \psi_h \cdot \mathbf{k} \cdot \nabla \phi_h + \frac{d}{|\mathbf{v}|^2}(\mathbf{v} \cdot \nabla \psi_h)(\mathbf{u} \cdot \nabla \phi_h)\right] d\Omega \\ &= \int_\Omega \left[\left(\psi_h + \frac{d}{|\mathbf{v}|^2}\mathbf{v} \cdot \nabla \psi_h\right)\mathbf{u} \cdot \nabla \phi_h + \nabla \psi_h \cdot \mathbf{k} \cdot \nabla \phi_h\right] d\Omega \end{aligned} \tag{1.35}$$

The basic idea of the SD method now follows easily. From (1.35) we see that the convective term has been weighted by $\psi_h + \zeta_h$, where

$$\zeta_h := \frac{d}{|\mathbf{v}|^2}\mathbf{v} \cdot \nabla \psi_h$$

Consider this expression within each element and take $\mathbf{v} = \mathbf{u}^e$ and $d = \frac{1}{2}\alpha^e h^e |\mathbf{u}^e|$ as in the 1D case, where α^e is a function of the element Péclet number γ^e to be determined. If we define

$$\tau^e := \frac{\alpha^e h^e}{2|\mathbf{u}^e|} \tag{1.36}$$

we will have that

$$\zeta_h = \tau^e \mathbf{u}^e \cdot \nabla \psi_h \tag{1.37}$$

for each element. The parameter τ^e has dimensions of time. It will be called *intrinsic time*. Although the function ζ_h in (1.35) only affects the convective term, the straightforward way to obtain a consistent weighted residual method is to make it affect *all the terms* of the equation [BH], [HB2]. The only problem to be faced is the definition of $\int_\Omega \zeta_h \nabla \cdot (\mathbf{k} \cdot \nabla \phi_h) d\Omega$, since it doesn't make sense for typical C^0 finite elements ($\nabla \phi_h$ will be discontinuous across interelement boundaries). The way to overcome this problem is to consider that ζ_h only affects the element interiors. These ideas lead to the final version of the SD method: Find a function $\phi_h \in \Phi_h$ such that

$$a_{sd}(\phi_h, \psi_h) = l_{sd}(\psi_h) \qquad \forall \psi_h \in \Psi_h \tag{1.38}$$

Here, the bilinear form a_{sd} and the linear form l_{sd} are

$$a_{sd}(\phi_h, \psi_h) := a(\phi_h, \psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} (\tau^e \mathbf{u}^e \cdot \nabla \psi_h) \left[\mathbf{u} \cdot \nabla \phi_h - \nabla \cdot (\mathbf{k} \cdot \nabla \phi_h)\right] d\Omega \tag{1.39}$$

$$l_{sd}(\psi_h) := l(\psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} (\tau^e \mathbf{u}^e \cdot \nabla \psi_h)\, f \, d\Omega \tag{1.40}$$

Remarks 1.2

(1) The Euler-Lagrange equations for the variational formulation (1.38) are precisely (1.19) and the boundary conditions are (1.20) and (1.21) (essential and natural, respectively), together with the additional condition of diffusive flux continuity across interelement boundaries [BH], [HB2].

(2) For rectangular bilinear elements in 2D or trilinear elements in 3D, with $k_{ij} = k\, \delta_{ij}$, k being a positive constant and δ_{ij} the Kronecker delta, we have that $\nabla \cdot (\mathbf{k} \cdot \nabla \phi_h) = k \Delta \phi_h \equiv 0$ within each element. This is always the case with linear triangles or tetrahedra. However, this term cannot be neglected if higher-order elements are used.

(3) When linear elements (Lagrangian or simplicial) are used, the common choice for the upwind function α^e is to compute it through the expression (1.12) or its asymptotic approximation (1.16), replacing γ by the element Péclet number γ^e given by (1.30). See Reference [HMM]. □

We have now a complete description of the SD method and the ideas underlying it. The purpose of this chapter is to give a precise definition of the intrinsic time τ^e. In Section 1.2, we shall analyse the convergence properties of the method for a simplified problem in order to get insight on the role played by the upwind functions. Section 1.3 contains the derivation of an expression for this functions when quadratic finite elements are employed and it is based in part on Reference [CO1], although several results are new. Some computational aspects will be considered in Section 1.4.

1.1.4 The Galerkin/least-squares method

From Eqn. (1.37) it is seen that the perturbation ζ_h of the test function $\psi_h \in \Psi_h$ that defines the SD method is nothing but the convective operator applied to this function multiplied by τ^e. Let us write the original continuous equation (1.19) as $\mathcal{L}\phi = f$, where $\mathcal{L}$ is the linear operator defined by

$$\mathcal{L}\phi := \mathbf{u} \cdot \nabla\phi - \nabla \cdot (\mathbf{k} \cdot \nabla\phi) \tag{1.41}$$

A natural variant of the SD method is to consider the following perturbation $\tilde{\zeta}_h$ of ψ_h:

$$\tilde{\zeta}_h := \tau^e \mathcal{L}\psi_h \tag{1.42}$$

that is, *the whole differential operator* is applied to ψ_h. The resulting formulation is known as the Galerkin/least-squares method (GLS), introduced by Hughes *et al.* [HFH] first in the context of the Stokes problem [HFB], [HF], [FH] but that has been successfully applied to a variety of other variational problems with constraints in structural mechanics (see [FH] and references therein).

The variational formulation we are led to using the GLS method reads as follows: Find $\phi_h \in \Phi_h$ such that

$$a_{gls}(\phi_h, \psi_h) = l_{gls}(\psi_h) \qquad \forall \psi_h \in \Psi_h \tag{1.43}$$

where the bilinear form a_{gls} and the linear form l_{gls} are

$$a_{gls}(\phi_h, \psi_h) := a(\phi_h, \psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} \tau^e \mathcal{L}\psi_h \; \mathcal{L}\phi_h \; d\Omega \tag{1.44}$$

$$l_{gls}(\psi_h) := l(\psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} \tau^e \mathcal{L}\psi_h \; f \; d\Omega \tag{1.45}$$

The forms a and l are those given by (1.25) and (1.26), respectively.

The GLS method doesn't seem to offer any improvement over the SD method for the convection-diffusion equation. However, it can be proved [HFB] that it allows to circumvent the Babuška-Brezzi stability condition (see Chapter 3) for the Stokes problem. In this case, the formulation depends on an algorithmic parameter whose physical meaning and optimal values are not known yet. In Section 1.3, the upwind functions for the GLS method will be derived. These functions will be different from the optimal correspondig to the SD formulation. If the zero convection limit is considered, the perturbation of the test function (1.37) for the SD method vanishes, but the perturbation (1.42) for the GLS method *does not.* This gives a natural way for computing the above mentioned parameter.

In what follows, no reference will be made to other existing methods except in Chapter 2, where the relation between the SD and the Taylor-Galerkin method coined by Donea [Do] will be discussed. Some of them, such as the Least-squares method used in References [CJ], [NRe], the characteristic Galerkin method (see, e.g., [DR], [LPZ], [VF]) or the use of weighted L^2-inner products in the forms that define the variational problem [Ax], [GH], are closely related to the SD method. For a review of different upwind methods placed in the same mathematical framework, see Reference [BBF].

1.2 Convergence analysis

1.2.1 Introduction and variational problem

The simplified problem we will consider in this section is the following: Find a function $\phi = \phi(\mathbf{x})$ such that

$$-k\Delta\phi + \mathbf{u}\cdot\nabla\phi + \sigma\phi = f, \qquad \text{in } \Omega \tag{1.46}$$

$$\phi = 0, \qquad \text{on } \Gamma \tag{1.47}$$

where k and σ are positive constants. For simplicity, we will also assume that $\nabla\cdot\mathbf{u} = 0$, with $|\mathbf{u}| \neq 0$. When this condition does not hold and σ is variable, the following condition is needed in order to ensure coercivity of the bilinear form associated to (1.46)–(1.47):

$$\sigma - \frac{1}{2}\nabla\cdot\mathbf{u} \geq \delta \geq 0 \tag{1.48}$$

where δ is a constant. If $\delta = 0$, a change of variables in the above problem can be done in such a way that the new value of δ be positive. For example, let $\beta(\mathbf{x})$ be a smooth function and define $\hat{\phi} = \phi\exp\beta$, $\hat{k} = k\exp\beta$, $\hat{\mathbf{u}} = \mathbf{u}\exp\beta - k(\exp\beta)\nabla\beta$ and $\hat{\sigma} = (\exp\beta)\mathbf{u}\cdot\nabla\beta - k(\exp\beta)\Delta\beta - k(\exp\beta)|\nabla\beta|^2$. Then, ϕ satisfies (1.46) iff $\hat{\phi}$ satisfies

$$-\hat{k}\Delta\hat{\phi} + \hat{\mathbf{u}}\cdot\nabla\hat{\phi} + \hat{\sigma}\hat{\phi} = f$$

Now, condition (1.48) applied to $\hat{\mathbf{u}}$ and $\hat{\sigma}$ with $\hat{\delta} > 0$ gives the following condition on β:

$$\mathbf{u}\cdot\nabla\beta - k\Delta\beta - k|\nabla\beta|^2 > 0$$

Explicit functions β satisfying this inequality can be constructed.

Remark 1.3

When $\delta = 0$, Nävert [Na] multiplies Eqn. (1.46) by a function χ, called $\delta-$compensating, such that $-\mathbf{u}\cdot\nabla\chi \geq \varrho\chi$, with $\varrho > 0$, and proves that such a function does exist (building it explicitly). Introducing the weighted inner product $(\phi_1,\phi_2)_\chi := \int_\Omega \chi\phi_1\phi_2\, d\Omega$ and using the properties of χ, the norm associated to $(\cdot,\cdot)_\chi$ happens to be equivalent to the standard L^2 norm. The $\delta-$compensating functions are also needed in order to obtain local error estimates. See [Na], [JNP] for details. □

Assume then that $\nabla\cdot\mathbf{u} = 0$ and that $\sigma > 0$, constant. If $V = H_0^1(\Omega)$, the weak form of problem (1.46)–(1.47) consists in finding $\phi \in V$ such that

$$a(\phi,\psi) = l(\phi) \qquad \forall\psi \in V \tag{1.49}$$

where

$$a(\phi,\psi) := k(\nabla\phi,\nabla\psi) + (\mathbf{u}\cdot\nabla\phi,\psi) + \sigma(\phi,\psi) \tag{1.50}$$

$$l(\psi) := (f,\psi) \tag{1.51}$$

If $V_h \subset V$ is a finite element subspace of V constructed with elements of degree m, the SD method will read: Find $\phi \in V_h$ such that

$$a_{sd}(\phi,\psi) = l_{sd}(\psi) \qquad \forall\psi \in V_h \tag{1.52}$$

where

$$a_{sd}(\phi,\psi) := a(\phi,\psi) + \sum_{e=1}^{N_{el}} (-k\Delta\phi_h + \mathbf{u}\cdot\nabla\phi_h + \sigma\phi_h, \tau^e \mathbf{u}\cdot\nabla\psi_h)_{\Omega^e} \tag{1.53}$$

$$l_{sd}(\psi) := l(\psi) + \sum_{e=1}^{N_{el}} (f, \tau^e \mathbf{u}\cdot\nabla\psi_h)_{\Omega^e} \tag{1.54}$$

where the expression of the intrinsic time τ^e is given by Eqn. (1.36). Again for simplicity, we will assume that $\tau^e = \tau$, constant for all the elements.

Our objective is to see which is the behavior of the upwind function $\alpha = \alpha(\gamma)$ dictated by the convergence analysis. This analysis will be basically the one that can be found in [Jo2], [JNP], [Na]. However, in these References τ is set to zero when γ is small and taken as a constant when γ is large. In this sense, our approach will be closer to the one used by Hughes *et al.* in [HFH] for the GLS method.

1.2.2 Interpolation estimates

We will need some standard error estimates from interpolation theory [Ci]. Let $\tilde{\psi}_h \in V_h$ be the finite element interpolant of a function $\psi \in V$ and h the diameter of the partition $\{\Omega^e\}$, that is assumed to satisfy the usual regularity requirements. We will use the following interpolation error and inverse estimates:

$$h^2 \left(\sum_{e=1}^{N_{el}} \|\psi - \tilde{\psi}_h\|^2_{2,\Omega^e} \right)^{\frac{1}{2}} + h\|\psi - \tilde{\psi}_h\|_{1,\Omega} + \|\psi - \tilde{\psi}_h\|_{\Omega} \leq \kappa_1 h^{m+1} \tag{1.55}$$

$$\|\psi\|_{s,\Omega^e} \leq \kappa_2 h^{-1} \|\psi\|_{s-1,\Omega^e}, \qquad s = 1,2 \tag{1.56}$$

We shall make use of the abbreviation

$$\|\cdot\|_e := \sum_{e=1}^{N_{el}} \|\cdot\|_{\Omega^e} \tag{1.57}$$

A generic constant will be denoted by C or C', possibly different at different occurrences. The constants κ_1 and κ_2 will be those appearing in (1.55) and (1.56). Observe that κ_1 contains the seminorm $|\psi|_{m+1}$, that will be bounded for regular enough functions ψ. Finally, recall that m is the degree of the polynomials used in the finite element discretization.

1.2.3 Error analysis

We will see in what follows that in order to ensure stability for the method (1.52) the function $\alpha(\gamma)$ must be of order γ as $\gamma \to 0$. The behavior when $\gamma \to \infty$ will be dictated by the asymptotic order of convergence.

We first establish stability.

Lemma 1.1 *If the upwind function $\alpha(\gamma)$ satisfies $\alpha(\gamma) \leq 4\gamma\kappa_2^{-2}$, then the bilinear form a_{sd} given by (1.53) is coercive in the norm $|||\cdot|||$ defined by:*

$$|||\psi_h|||^2 := k\|\nabla\psi_h\|^2 + \tau\|\mathbf{u}\cdot\nabla\psi_h\|^2 + 2\sigma\|\psi_h\|^2, \qquad \psi_h \in V_h \tag{1.58}$$

that is, $a_{sd}(\psi_h, \psi_h) \geq C|||\psi_h|||^2$ *for all* $\psi_h \in V_h$. *In particular, one can take* $C = \frac{1}{2}$.

Proof: Observe first that, using the inverse estimate (1.56):

$$\sum_{e=1}^{N_{el}} (-k\Delta\psi_h, \tau\mathbf{u}\cdot\nabla\psi_h)_{\Omega^e} \geq -k\|\Delta\psi_h\|_e\|\tau\mathbf{u}\cdot\nabla\psi_h\|$$

$$\geq -\frac{1}{2}k\left(\|\nabla\psi_h\|^2 + \kappa_2^2 h^{-2}\tau^2\|\tau\mathbf{u}\cdot\nabla\psi_h\|^2\right) \qquad (1.59)$$

and therefore

$$a_{sd}(\psi_h, \psi_h) \geq k\|\nabla\psi_h\|^2 + \sigma\|\psi_h\|^2 - \frac{1}{2}k\|\nabla\psi_h\|^2$$

$$+ \tau\left(1 - \frac{1}{2}k\kappa_2^2 h^{-2}\tau\right)\|\mathbf{u}\cdot\nabla\psi_h\|^2 \qquad (1.60)$$

The rest of the terms vanish, since $(\mathbf{u}\cdot\nabla\psi_h, \psi_h) = 0$. From the hypothesis on α we have that

$$1 - \frac{1}{2}k\kappa_2^2 h^{-2}\tau = 1 - \frac{1}{2}k\kappa_2^2 h^{-2}\frac{\alpha h}{2|\mathbf{u}|} \geq \frac{1}{2}$$

and the Lemma follows from (1.60). □

Consistency of problem (1.52) is a trivial consequence of the fact that SD is a residual method:

Lemma 1.2 *Let* ϕ *be the solution of the continuous problem and* ϕ_h *the solution of (1.52). Then*

$$a_{sd}(\phi - \phi_h, \psi_h) = 0 \qquad (1.61)$$

for all $\psi_h \in V_h$. □

Convergence is next established. We will use the fact that

$$(\mathbf{u}\cdot\nabla\psi_1, \psi_2) = -(\psi_1, \mathbf{u}\cdot\nabla\psi_2)$$

for $\psi_1, \psi_2 \in V$. This follows from the assumption that $\nabla\cdot\mathbf{u} = 0$.

Theorem 1.1 *Assume that the hypothesis of Lemma 1.1 hold. If* $\alpha(\gamma) = O(1)$ *as* $\gamma \to \infty$, *then there is a constant* C *such that*

$$|||\phi - \phi_h||| \leq Ch^{m+\frac{1}{2}} \qquad (1.62)$$

for γ *large enough.*

Proof: Let us split the error $e := \phi - \phi_h$ as

$$e = (\phi - \tilde{\phi}_h) + (\tilde{\phi}_h - \phi_h) =: \eta + e_h$$

where η is the interpolation error and $e_h \in V_h$. To find an estimate for $|||e_h|||$, Lemmas 1.1 and 1.2 and the definition of the bilinear form a_{sd} will be used:

$$\frac{1}{2}|||e_h|||^2 \leq a_{sd}(e_h, e_h) = a_{sd}(e - \eta, e_h) = -a_{sd}(\eta, e_h)$$

$$
\begin{aligned}
&\leq k\|\nabla\eta\|\,\|\nabla e_h\| + \|\eta\|\,\|\mathbf{u}\cdot\nabla e_h\| + \sigma\|\eta\|\,\|e_h\| \\
&+ k\tau\|\Delta\eta\|_e\|\mathbf{u}\cdot\nabla e_h\| + \tau\|\mathbf{u}\cdot\nabla\eta\|\,\|\mathbf{u}\cdot\nabla e_h\| + \sigma\tau\|\mathbf{u}\cdot\nabla\eta\|\,\|e_h\| \\
&\leq \left(\frac{1}{4}k\|\nabla e_h\|^2 + k\|\nabla\eta\|^2\right) + \left(\frac{1}{16}\tau\|\mathbf{u}\cdot\nabla e_h\|^2 + 4\tau^{-1}\|\eta\|^2\right) \\
&+ \left(\frac{1}{4}\sigma\|e_h\|^2 + \sigma\|\eta\|^2\right) + \left(\frac{1}{16}\tau\|\mathbf{u}\cdot\nabla e_h\|^2 + 4k^2\tau\|\Delta\eta\|_e^2\right) \\
&+ \left(\frac{1}{8}\tau\|\mathbf{u}\cdot\nabla e_h\|^2 + 2\tau\|\mathbf{u}\cdot\nabla\eta\|^2\right) + \left(\frac{1}{4}\sigma\|e_h\|^2 + \sigma\tau^2\|\mathbf{u}\cdot\nabla\eta\|_e^2\right) \\
&= \frac{1}{4}|||e_h|||^2 + E(\eta) \qquad (1.63)
\end{aligned}
$$

where

$$
\begin{aligned}
E(\eta) :&= k\|\nabla\eta\|^2 + 4\tau^{-1}\|\eta\|^2 + \sigma\|\eta\|^2 + 4k^2\tau\|\Delta\eta\|_e^2 \\
&+ 2\tau\|\mathbf{u}\cdot\nabla\eta\|^2 + \sigma\tau^2\|\mathbf{u}\cdot\nabla\eta\|^2
\end{aligned}
$$

Using the bound for α stated in Lemma 1.1 and the inverse estimate (1.56) we obtain:

$$
4k^2\tau\|\Delta\eta\|_e^2 \leq 4k^2\frac{\alpha h}{2|\mathbf{u}|}\kappa_2^2 h^{-2}\|\nabla\eta\|^2 = k\frac{\alpha}{\gamma}\kappa_2^2\|\nabla\eta\|^2 \leq 4k\|\nabla\eta\|^2
$$

and using this inequality and the interpolation estimate (1.55) we get

$$
\begin{aligned}
E(\eta) &\leq 5k\|\nabla\eta\|^2 + 4\tau^{-1}\|\eta\|^2 + \sigma\|\eta\|^2 + 2\tau\|\mathbf{u}\cdot\nabla\eta\|^2 + \sigma\tau^2\|\mathbf{u}\cdot\nabla\eta\|^2 \\
&\leq \kappa_1\left(5kh^{2m} + \frac{8|\mathbf{u}|}{\alpha h}h^{2m+2} + \sigma h^{2m+2} + \frac{\alpha h}{|\mathbf{u}|}|\mathbf{u}|^2h^{2m} + \sigma\frac{\alpha^2h^2}{4|\mathbf{u}|^2}|\mathbf{u}|^2h^{2m}\right) \\
&\leq C|\mathbf{u}|\left(\frac{1}{\gamma}h^{2m+1} + \frac{8}{\alpha}h^{2m+1} + \alpha h^{2m+1}\right) + C'\alpha^2\sigma h^{2m+2} \qquad (1.64)
\end{aligned}
$$

Using inequality (1.63) we see that $|||e_h|||^2 \leq 4E(\eta)$ and from this last bound (1.64) we get $|||e_h|||^2 \leq Ch^{2m+1}$. The Theorem follows from the fact that $|||\eta|||^2 \leq Ch^{2m+1}$ (obtained using (1.55)) and applying the triangle inequality. □

Remarks 1.4

(1) From the bound (1.64), it is seen that when γ is small, say $\gamma \leq h$, the dominant term will be of order h^{2m}. But in this case, the norm $|||\cdot|||$ is equivalent to $\|\cdot\|_1$ and through the classical duality argument of Aubin-Nitsche an optimal L^2 estimate can be obtained.

(2) The term αh^{2m+1} is the reason why α must be bounded by a constant when $\gamma \to \infty$ if an error of order h^{2m+1} is sought.

(3) If Neumann boundary conditions are prescribed on a part of $\partial\Omega$, the following interpolation estimate has to be used [Ci]:

$$
\|\psi - \tilde{\psi}_h\|_{L^2(\partial\Omega)} \leq Ch^{m+\frac{1}{2}}|\psi|_{H^{m+1}(\Omega)}
$$

See References [Na], [HFH].

(4) The fact that $\sigma > 0$ implies that $|||\cdot|||$ will be equivalent to the L^2 norm. Since the error of the SD formulation is of order $h^{m+1/2}$ in this norm and the interpolation error is of order h^{m+1}, the SD method is said to have a 'gap' 1/2 from being

optimal. For the standard Galerkin method, only an error estimate of order h^m can be obtained when convection is dominant [Na] and therefore the gap is 1. □

The main conclusion of this analysis is the following: *the function $\alpha(\gamma)$ must be of the form*

$$\alpha(\gamma) = \begin{cases} C_1\gamma & \text{as } \gamma \to 0 \\ C_2 & \text{as } \gamma \to \infty \end{cases} \tag{1.65}$$

in order to have stability and optimal rate of convergence for the SD method.

A possible extension of the SD formulation to convection-diffusion systems of equations was introduced by Hughes & Mallet in Reference [HM] and the analysis was carried out in Reference [HFM]. A review of *a posteriori* error estimates and adaptive finite elements for the SD method obtained by Johnson and collaborators can be found in Reference [Jo3].

1.3 The optimal upwind functions for one–dimensional quadratic elements

1.3.1 General considerations

In this section we will consider again the one-dimensional steady-state problem (1.1)–(1.2), but now with a source term $f = f(x)$. As for linear elements, this will provide us a way to calculate the upwind functions. These functions, together with the definition of the characteristic length, velocity and diffusion of each element, are the necessary ingredients to compute the intrinsic time τ^e defined by Eqn. (1.36). Up to now, the only thing we know about the upwind functions is that they have to behave as Eqn. (1.65) dictates. Nevertheless, numerical experiments indicate that a proper evaluation of α^e greatly influences the accuracy (not the stability) of the results. Overdiffusive answers are found if this function is overestimated, whereas oscillations may occur if a too small estimate is employed. The use of expression (1.12) with γ replaced by the element Péclet number given by (1.30) has proved to be very effective for linear elements. Approaches other than the SD formulation using quadratic elements have been studied [CM], [DBS], [He], [HZ1]. However, for this one it seems that an 'optimal' upwind function is missing, although using one half of the optimal for linear elements has been proposed [Sh]. This choice will be justified in what follows.

The purpose of this section is to obtain an expression similar to (1.12) for quadratic elements in different cases (results will be recapitulated at the end of this chapter). For that, we will consider the problem: Find $\phi = \phi(x)$ such that

$$u\frac{d\phi}{dx} - k\frac{d^2\phi}{dx^2} = f(x), \quad 0 < x < \ell \tag{1.66}$$

$$\phi(0) = \phi_0, \quad \phi(\ell) = \phi_\ell \tag{1.67}$$

where u and k will be considered positive constants, $\ell > 0$ and ϕ_0 and ϕ_ℓ are given boundary values of the function ϕ. First, we shall assume $f(x) \equiv 0$ and in Subsection 1.3.5 the introduction of source terms will be addressed.

From here onwards, N will denote a generic shape function of a quadratic element e and W a weighting function. According to (1.15), this weighting function will be

expressed as

$$W(x) = N(x) + \tau^e \, u \, \frac{dN}{dx} \tag{1.68}$$

and the intrinsic time of (1.36) as

$$\tau^e = \frac{\alpha^e \, h^e}{2|u|} \tag{1.69}$$

Throughout this section we assume that $[0, \ell]$ is discretized using a uniform finite element partition with elements of length h. Thus, the Péclet number $\gamma = |u|h/2k$ and the function α will be the same for all the elements. From (1.68) and (1.69) we have

$$W(x) = N(x) + \frac{\alpha \, h}{2} \mathrm{sgn}(u) \, \frac{dN}{dx} \tag{1.70}$$

where α will depend on the Péclet number γ. The sign of u will be considered included in γ and therefore in α. For the reasons explained is Subsection 1.1.2, this function will be considered *optimal* if the finite element solution obtained with the weighting functions given by (1.70) is nodally exact, i.e., both the analytical and the finite element solution of (1.66)–(1.67) take the same values at the nodes of the finite element mesh.

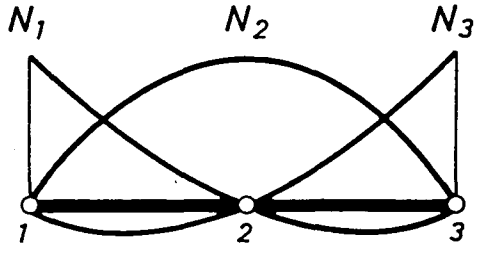

Figure 1.1 Three noded quadratic element and shape functions

We have already seen that for linear elements optimality is attained if the expression (1.12) is used for $\alpha(\gamma)$. Our aim is to derive the expressions of the upwind functions using quadratic elements. First, we observe that applying the Galerkin method (i.e, $W = N$) to (1.66)–(1.67) with $f(x) \equiv 0$ the following difference equations are found:

$$[1+\gamma]\phi_{m-1} - [8+4\gamma]\phi_{m-\frac{1}{2}} + 14\phi_m + [-8+4\gamma]\phi_{m+\frac{1}{2}} + [1-\gamma]\phi_{m+1} = 0 \tag{1.71}$$

for the 'extreme' nodes (nodes 1 and 3 in Figure 1.1) and

$$[-4-2\gamma]\phi_m + 8\phi_{m+\frac{1}{2}} + [-4+2\gamma]\phi_{m+1} = 0 \tag{1.72}$$

for the 'central' nodes (node 2 in Figure 1.1) The indexes in these equations are used according to Figure 1.2

Since different equations hold for the extreme and the central nodes, it can be anticipated that no single optimal upwind function will exist for quadratic elements. Instead, we will consider

$$W_i(x) = N_i(x) + \frac{\alpha h}{2} \frac{dN_i}{dx} \quad \text{for } i = 1, 3 \tag{1.73}$$

$$W_2(x) = N_2(x) + \frac{\beta h}{2} \frac{dN_2}{dx} \tag{1.74}$$

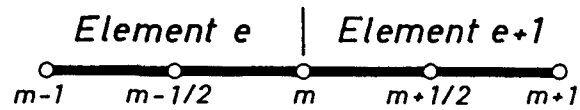

Figure 1.2 Indexes referring the nodes of two adjacent elements

Our purpose is to find analytical expressions for α and β.

1.3.2 Standard formulation of the SD method

We now consider the case in which the bases of the discrete finite element spaces are constructed using the classical shape functions depicted in Figure 1.1, that is, the standard or canonical bases are chosen. The upwind functions α and β appearing in (1.73) and (1.74) will be determined following the same criteria as for linear elements, i.e., by solving analytically the resulting difference equations obtained applying the SD method to (1.66) and (1.67) and by subsequently imposing that the numerical solution be nodally exact.

If the weighting functions (1.73) and (1.74) are used, the new difference equations (instead of (1.71) and (1.72)) are

$$\begin{aligned}&[1-6\alpha+\gamma(1+\alpha)]\phi_{m-1}-[8-12\alpha+\gamma(4+8\alpha)]\phi_{m-\frac{1}{2}}\\&+[14+14\alpha\gamma]\phi_m+[-8-12\alpha+\gamma(4-8\alpha)]\phi_{m+\frac{1}{2}}\\&+[1+6\alpha+\gamma(-1+\alpha)]\phi_{m+1}=0\end{aligned}\tag{1.75}$$

for the extreme nodes and

$$[-4-\gamma(2+4\beta)]\phi_m+[8+8\gamma\beta]\phi_{m+\frac{1}{2}}+[-4+\gamma(2-4\beta)]\phi_{m+1}=0\tag{1.76}$$

for the central nodes. Obtaining $\phi_{m+\frac{1}{2}}$ in terms of ϕ_m and ϕ_{m+1} from (1.76) and the analogous expression of $\phi_{m-\frac{1}{2}}$ in terms of ϕ_{m-1} and ϕ_m and inserting both expressions in (1.75) the following equation is found

$$a_1\phi_{m-1}+a_2\phi_m+a_3\phi_{m+1}=0\tag{1.77}$$

where we have introduced the notation

$$\begin{aligned}a_1&:=3+3\gamma+\gamma^2+3\gamma\beta+\gamma^2\beta+2\gamma^2\alpha+3\gamma^2\alpha\beta\\a_2&:=-(6+2\gamma^2+6\gamma\beta+6\gamma^2\alpha\beta)\\a_3&:=3-3\gamma+\gamma^2+3\gamma\beta-\gamma^2\beta-2\gamma^2\alpha+3\gamma^2\alpha\beta\end{aligned}\tag{1.78}$$

Since $\lambda=1$ and $\lambda=a_1/a_3$ are the roots of the characteristic polynomial of (1.77), its analytical solution will be given by

$$\phi_m=C_1+C_2\left(\frac{a_1}{a_3}\right)^m\tag{1.79}$$

where C_1 and C_2 are constants depending on the boundary conditions. If x_m is the abcissa of the mth nodal point and $\phi(x_m)$ the value of the exact solution of problem (1.66)-(1.67) at this node, it can be readily seen that $\phi_m = \phi(x_m)$ if, and only if

$$\frac{a_1}{a_3} = e^{2\gamma} \tag{1.80}$$

Now, assuming that (1.80) holds, from (1.76) one finds that $\phi(x_{m+\frac{1}{2}}) = \phi_{m+\frac{1}{2}}$ if, and only if,

$$e^{\gamma} = \frac{4 + \gamma(2 + 4\beta) + e^{2\gamma}[4 - \gamma(2 - 4\beta)]}{8 + 8\gamma\beta} \tag{1.81}$$

Assume $\gamma \neq 0$. From (1.81)

$$\beta(\gamma) = \frac{1}{2}\left(\coth\frac{\gamma}{2} - \frac{2}{\gamma}\right) \tag{1.82}$$

and from (1.80)

$$\alpha(\gamma) = \frac{(3 + 3\gamma\beta)\tanh\gamma - (3\gamma + \gamma^2\beta)}{(2 - 3\beta\tanh\gamma)\gamma^2} \tag{1.83}$$

The expressions of α and β given by (1.83) and (1.82) are the sought upwind functions. Unfortunately, these expressions look rather more complicated than the corresponding function for linear elements (1.12).

Remarks 1.5

(1) In the first section it has been seen that the use of the SD formulation with linear elements for homogeneous equations and neglecting the contribution of $\nabla\cdot(\mathbf{k}\cdot\nabla\phi)$ in (1.39) may be interpreted simply as the introduction of numerical diffusion along the streamlines. However, this is not exactly the case for quadratic elements for two reasons: first, the mentioned term $\nabla\cdot(\mathbf{k}\cdot\nabla\phi)$ cannot be neglected, and secondly, the existence of two optimal upwind functions would imply a non-constant added diffusion.

(2) It can be easily seen that the functions α and β are skew-symmetric. Remember that they had to include the sign of the velocity. In multidimensional situations γ^e will be positive (cf. Eqn. (1.30)) and the direction of the flow will be taken into account by the perturbation (1.37). □

When linear elements are used, it has already been explained why the function $\alpha(\gamma)$ given by (1.12) is approximated by the function (1.16). For the functions $\alpha(\gamma)$ and $\beta(\gamma)$ given by (1.83) and (1.82) a straightforward computation reveals that

$$\lim_{\gamma\to\infty}\alpha(\gamma) = 1 \quad \text{and} \quad \lim_{\gamma\to\infty}\beta(\gamma) = \frac{1}{2} \tag{1.84}$$

Expanding $\alpha(\gamma)$ and $\beta(\gamma)$ in Taylor series in the neighborhood of $\gamma = 0$, the following expressions are found

$$\alpha(\gamma) = \frac{\gamma}{12} + O(\gamma^3) \quad \text{and} \quad \beta(\gamma) = \frac{\gamma}{12} + O(\gamma^3) \tag{1.85}$$

Having this limits in mind, (1.83) and (1.82) can be approximated respectively by

$$\alpha_a(\gamma) = \begin{cases} \frac{\gamma}{12} & \text{if } 0 \leq |\gamma| \leq 12 \\ \text{sgn}\gamma & \text{if } |\gamma| > 12 \end{cases} \tag{1.86}$$

$$\beta_a(\gamma) = \begin{cases} \frac{\gamma}{12} & \text{if } 0 \le |\gamma| \le 6 \\ \frac{1}{2}\text{sgn}\gamma & \text{if } |\gamma| > 6 \end{cases} \tag{1.87}$$

However, from Figure 1.3 it is seen that (1.86) and (1.87) do not give such a good approximation to (1.83) and (1.82), respectively, as (1.16) does to (1.12). In Figure 1.3, the upwind functions for linear elements are labelled '*l*'. Functions (1.16), (1.86) and (1.87) are called *asymptotic approximations.*

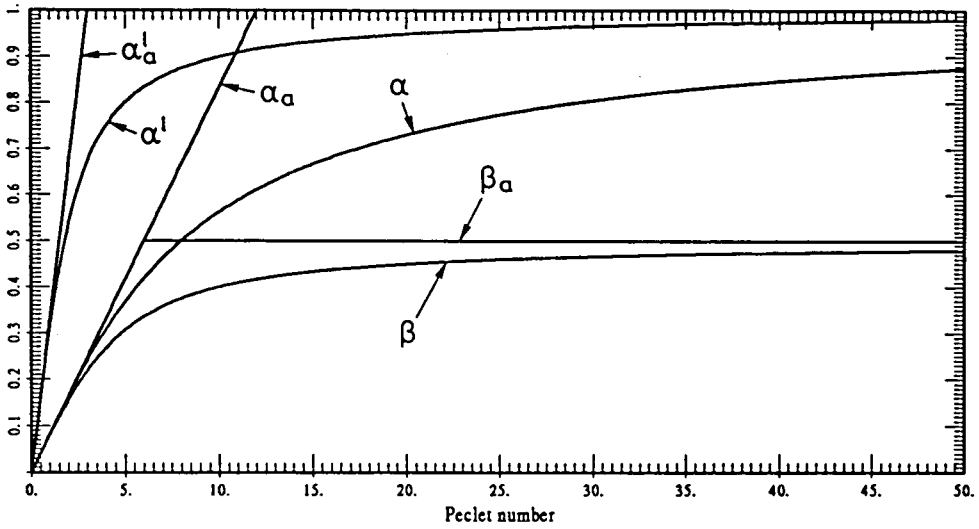

Figure 1.3 Upwind functions for linear and quadratic elements and their asymptotic approximations

It is important to remark that the functions α, β and their asymptotic approximations, as well as the upwind functions we will find for other cases below, *satisfy condition (1.65).* We will not allude to this point any more.

We have seen that nodally exact results for the solution of (1.66)–(1.67) using the SD formulation can only be obtained if the weighting functions (1.73) and (1.74) are used, with α and β given by (1.83) and (1.82). However, one could try to find a unique intrinsic time for all the nodes of the element (i.e., $\alpha = \beta$) and to relax the definition of 'optimality'. An obvious design criterion for the upwind function is that it must not be strongly dependent on the boundary conditions, in the sense that the difference between the values of this function and the functions that give nodally exact results for different boundary conditions should be bounded and as small as possible.

From the expression of the solution at the nodes (1.79) it is seen that if Eqn. (1.80) holds, the constants C_1 and C_2 that depend on the boundary conditions happen to be the same as the corresponding constants for the analytical solution of (1.66) that are determined from (1.67). Thus, although with a unique upwind function (1.80) will not be satisfied, we can try to find this function, say $\alpha^1(\gamma)$, by minimizing the difference

$$a_1 - a_3 e^{2\gamma} \tag{1.88}$$

If in expressions (1.78) we set $\alpha = \beta$ and try to satisfy (1.80), we are led to the following equation:

$$P(\alpha) := \alpha^2 + b\alpha + c = 0 \tag{1.89}$$

whith b and c defined by

$$b := \frac{1}{\gamma} - \coth\gamma \tag{1.90}$$

$$c := \frac{1}{\gamma^2} - \frac{1}{\gamma}\coth\gamma + \frac{1}{3} \tag{1.91}$$

The discriminant $\Delta := b^2 - 4c$ of equation (1.89) is plotted in Figure 1.4. Since Δ can be negative, (1.89) does not have real roots for all values of γ. However, we could try to minimize $P(\alpha)$. For a given value γ_0 of the Péclet number, the minimum of $P(\alpha)$ is attained at the point

$$\alpha_0 = -\frac{1}{2}b = \frac{1}{2}\left(\coth\gamma_0 - \frac{1}{\gamma_0}\right) \tag{1.92}$$

From Eqns. (1.90)–(1.92) it is easy to see that $b \to -1$, $c \to \frac{1}{3}$ and $\alpha_0 \to \frac{1}{2}$ as $\gamma_0 \to \infty$, and therefore we will have that

$$\lim_{\gamma\to\infty} P(\alpha_0) = \frac{1}{12}$$

So the function

$$\alpha^1(\gamma) = \frac{1}{2}\left(\coth\gamma - \frac{1}{\gamma}\right) \tag{1.93}$$

seems to be a good candidate for use as the upwind function since, although (1.89) is not fulfilled, $P(\alpha^1)$ remains small for all values of γ. In Figure 1.4 this value is represented against the Péclet number.

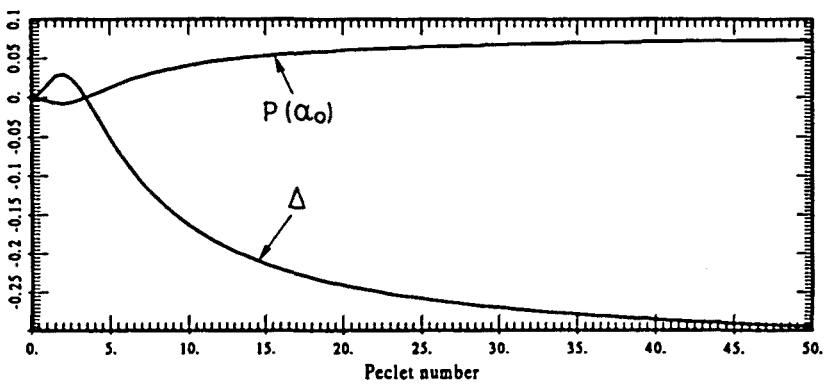

Figure 1.4 Discriminant Δ of equation (1.89) and values of $P(\alpha_0)$ for α_0 given by (1.92)

Like the function $\alpha^l(\gamma)$ given by (1.12), $\alpha^1(\gamma)$ can be approximated by

$$\alpha_a^1(\gamma) = \begin{cases} \frac{\gamma}{6} & \text{if } 0 \le |\gamma| \le 3 \\ \frac{1}{2} & \text{if } |\gamma| > 3 \end{cases} \tag{1.94}$$

The function $\alpha^1(\gamma)$ given by (1.93) is represented in Figure 1.5, together with $\alpha(\gamma)$ and $\beta(\gamma)$ of (1.83) and (1.82), for purposes of comparison. As it has already been said, this function had been proposed before by Shakib [Sh] as the result of numerical

experiments using quadratic elements. Now we have a justification of its use. Our numerical experiments also indicate that (1.93) is the best choice when a unique upwind function is to be used. The way these experiments have been performed is the following. For a test case in which the analytical solution is known (see, e.g., the first example of Section 1.5), the error of the numerical solution using the upwind function $K\alpha^1(\gamma)$ has been computed for different values of the constant K. For all the examples carried out, $K \approx 1$ happened to be the optimal choice.

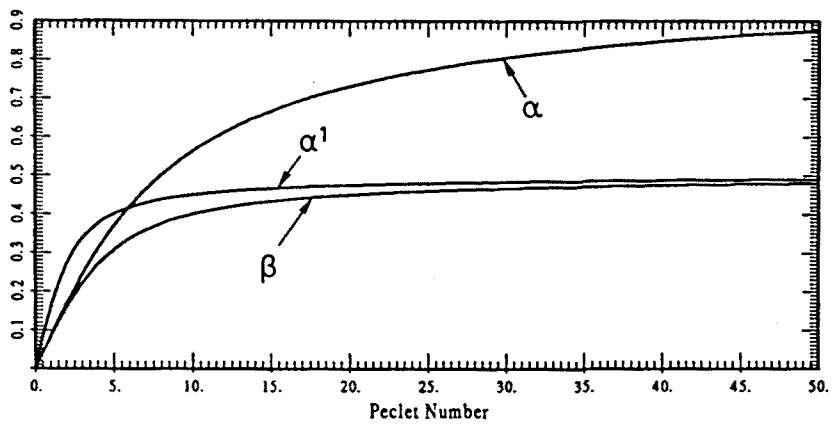

Figure 1.5 Upwind functions for quadratic elements

1.3.3 Hierarchic formulation of the SD method

The objective of this section is to investigate how sensitive the optimal upwind functions are to the interpolation used within each element. This sensitivity is a clear handicap when the previous concepts have to be applied in multidimensional situations, in which case the expressions of the shape functions take different forms depending on the direction one considers. This will be discussed in more detail in Section 1.4.

Now, let us consider the unknown function $\phi(x)$ interpolated within each element as

$$\phi(x) \approx N_1(x)\phi_1 + N_2(x)\Delta\phi_2 + N_3(x)\phi_3 \tag{1.95}$$

where N_1, N_2 and N_3 are the shape functions shown in Figure 1.6, ϕ_1, ϕ_2 and ϕ_3 the nodal values of ϕ and $\Delta\phi_2$ the difference between ϕ_2 and the linear interpolation at node 2 using the values ϕ_1 and ϕ_3.

A similar analysis to that made in Subsection 1.3.2 shows that the optimal upwind functions are now given by

$$\beta^h(\gamma) = \frac{1}{2} + \frac{1}{e^\gamma - 1} - \frac{1}{\gamma} \tag{1.96}$$

$$\alpha^h(\gamma) = \left(1 + \frac{1}{\gamma\beta}\right)\coth\gamma - \left(\frac{1}{3\beta} + \frac{1}{\gamma^2\beta} + \frac{1}{\gamma}\right) \tag{1.97}$$

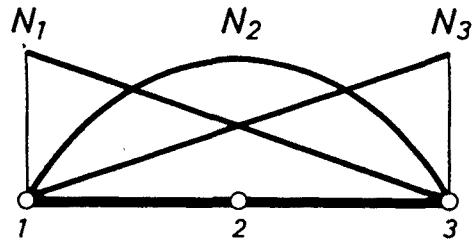

Figure 1.6 Hierarchic shape functions for three noded quadratic elements

where the label 'h' refers to the hierarchic formulation of the element. The asymptotic behavior of the upwind function α^h is completely different from the corresponding α given by (1.83), whereas the asymptotic behavior of β^h is similar to that of the function β in (1.82). In fact, now we have that

$$\lim_{\gamma\to\infty} \alpha^h(\gamma) = \frac{1}{3} \quad \text{and} \quad \lim_{\gamma\to\infty} \beta^h(\gamma) = \frac{1}{2}$$

and that

$$\alpha^h(\gamma) = \frac{\gamma}{15} + O(\gamma^2) \quad \text{and} \quad \beta^h(\gamma) = \frac{\gamma}{12} + O(\gamma^2)$$

in the neighborhood of $\gamma = 0$. Therefore, the asymptotic approximations for α^h and β^h will be

$$\alpha_a^h(\gamma) = \begin{cases} \frac{\gamma}{15} & \text{if } 0 \le |\gamma| \le 5 \\ \frac{1}{3} & \text{if } |\gamma| > 5 \end{cases} \tag{1.98}$$

$$\beta_a^h(\gamma) = \begin{cases} \frac{\gamma}{12} & \text{if } 0 \le |\gamma| \le 6 \\ \frac{1}{2} & \text{if } |\gamma| > 6 \end{cases} \tag{1.99}$$

We see that $\beta_a^h(\gamma) = \beta_a(\gamma)$ (cf. Eqn. (1.87)) but $\alpha_a^h(\gamma)$ and $\alpha_a(\gamma)$ differ totally (cf. Eqn. (1.86)). In Figure 1.7 the functions $\alpha^h(\gamma)$ of (1.97) and $\beta^h(\gamma)$ of (1.96) are represented.

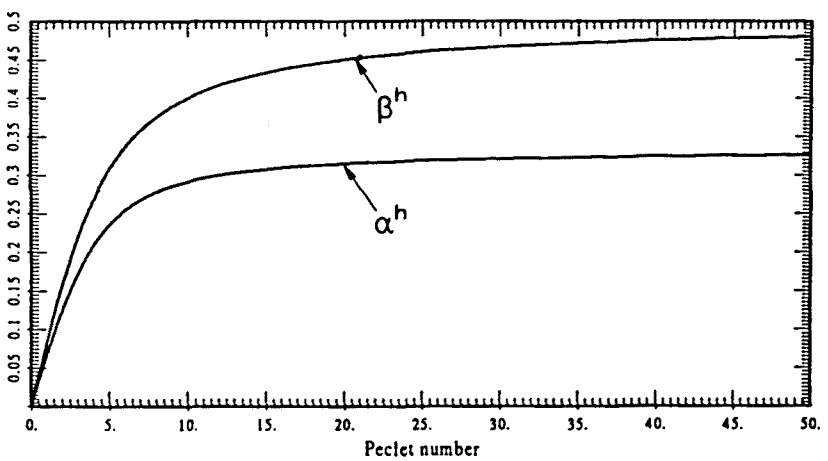

Figure 1.7 Upwind functions for hierarchic quadratic elements

An interesting point is that if a unique upwind function is sought using the same considerations as in the last subsection, this upwind function happens to be the same as for the standard formulation, i.e., the function $\alpha^1(\gamma)$ given by (1.93).

The main conclusion of this analysis is that the optimal upwind functions are very sensitive to the finite element interpolation chosen. This should be kept in mind since it has already been said that results are too diffusive if the upwind function is overestimated, whereas oscillations may occur if it is underestimated.

1.3.4 Standard formulation of the GLS method

In this subsection we again consider that the finite element interpolation is done with the shape functions depicted in Figure 1.1 but now that the GLS formulation is employed. According to expression (1.42), now the weighting functions will be:

$$\begin{aligned} W(x) &= N(x) + \frac{\alpha\, h}{2|u|}\left(u\frac{dN}{dx} - k\frac{d^2N}{dx^2}\right) \\ &= N(x) + \frac{\alpha\, h}{2}\mathrm{sgn}(u)\frac{dN}{dx} - \frac{\alpha\, h}{2|u|}k\frac{d^2N}{dx^2} \\ &= N(x) + \frac{\alpha\, h}{2}\mathrm{sgn}(u)\frac{dN}{dx} - \frac{\alpha}{\gamma}\frac{h^2}{4}\frac{d^2N}{dx^2} \end{aligned} \tag{1.100}$$

Observe that when $\gamma \to 0$ we have to have $\alpha \to 0$. If $\alpha = C\gamma$ as $\gamma \to 0$, then a term of the form $Ch^2/4$ will multiply the diffusion operator applied to the shape function. The constant C will appear naturally from what follows. Observe also that when linear elements are used, the SD and the GLS methods coincide.

As before, the sign of u will be considered to be included in γ and thus in the function α. Recall also that the perturbation of $N(x)$ that appears in (1.100) is only applied to the element interiors, in the sense explained in Section 1.1.

As for the SD method, no single upwind function will give nodally exact answers for problem (1.66)–(1.67). Using the same notation as before, the weighting functions for each element will be taken as (see Figure 1.1):

$$W_i(x) = N_i(x) + \frac{\bar{\alpha}h}{2}\frac{dN_i}{dx} - \frac{\bar{\alpha}}{\gamma}\frac{h^2}{4}\frac{d^2N_i}{dx^2} \quad \text{for } i = 1,3 \tag{1.101}$$

$$W_2(x) = N_2(x) + \frac{\bar{\beta}h}{2}\frac{dN_2}{dx} - \frac{\bar{\beta}}{\gamma}\frac{h^2}{4}\frac{d^2N_2}{dx^2} \tag{1.102}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are the upwind functions to be determined. The use of the weighting functions (1.101)–(1.102) for problem (1.66)–(1.67) with $f(x) \equiv 0$ leads to the following system of difference equations:

$$\begin{aligned} &[1 + \gamma(1+\bar{\alpha}) + 12\frac{\bar{\alpha}}{\gamma}]\phi_{m-1} - [8 - 12\bar{\alpha} + \gamma(4+8\bar{\alpha}) + 24\frac{\bar{\alpha}}{\gamma}]\phi_{m-\frac{1}{2}} \\ &+ [14 + 14\bar{\alpha}\gamma + 24\frac{\bar{\alpha}}{\gamma}]\phi_m + [-8 - 12\bar{\alpha} + \gamma(4-8\bar{\alpha}) - 24\frac{\bar{\alpha}}{\gamma}]\phi_{m+\frac{1}{2}} \\ &+ [1 + \gamma(-1+\bar{\alpha}) + 12\frac{\bar{\alpha}}{\gamma}]\phi_{m+1} = 0 \end{aligned} \tag{1.103}$$

for the extreme nodes and

$$\begin{aligned}&[-4-\gamma(2+4\bar{\beta})-6\bar{\beta}-12\frac{\bar{\beta}}{\gamma}]\phi_m+[8+8\gamma\bar{\beta}+24\frac{\bar{\beta}}{\gamma}]\phi_{m+\frac{1}{2}}\\&+[-4+\gamma(2-4\bar{\beta})+6\bar{\beta}-12\frac{\bar{\beta}}{\gamma}]\phi_{m+1}=0\end{aligned}\tag{1.104}$$

for the central nodes. Working out the explicit solution of this system of equations (1.103)-(1.104) and imposing $\phi_m = \phi(x_m)$, $\phi(x)$ being the solution of the continuous problem, the following expressions for $\bar{\alpha}$ and $\bar{\beta}$ are found:

$$\bar{\beta} = \frac{\gamma^2\left(\coth\frac{\gamma}{2}-\frac{2}{\gamma}\right)}{6-3\gamma\coth\frac{\gamma}{2}+2\gamma^2}\tag{1.105}$$

$$\bar{\alpha} = \frac{\tanh\gamma\left(3+\gamma^2+6\gamma\bar{\beta}+9\frac{\bar{\beta}}{\gamma}\right)-\left(3\gamma+9\bar{\beta}+\gamma^2\bar{\beta}\right)}{2\gamma^2-3\bar{\beta}\gamma^2\tanh\gamma}\tag{1.106}$$

As before, we might be interested in using the simpler expressions resulting from the asymptotic approximation of these two functions. Now we have that

$$\lim_{\gamma\to\infty}\bar{\alpha}(\gamma)=1 \quad \text{and} \quad \lim_{\gamma\to\infty}\bar{\beta}(\gamma)=\frac{1}{2}\tag{1.107}$$

and in the neighborhood of $\gamma = 0$

$$\bar{\alpha}(\gamma)=\frac{\gamma}{9}+O(\gamma^3) \quad \text{and} \quad \bar{\beta}(\gamma)=\frac{\gamma}{9}+O(\gamma^3)\tag{1.108}$$

So, the asymptotic aproximation of the functions $\bar{\alpha}$ and $\bar{\beta}$ will be

$$\bar{\alpha}_a(\gamma)=\begin{cases}\frac{\gamma}{9} & \text{if } 0\le|\gamma|\le 9\\ \text{sgn}\gamma & \text{if } |\gamma|>9\end{cases}\tag{1.109}$$

$$\bar{\beta}_a(\gamma)=\begin{cases}\frac{\gamma}{9} & \text{if } 0\le|\gamma|\le\frac{9}{2}\\ \frac{1}{2}\text{sgn}\gamma & \text{if } |\gamma|>\frac{9}{2}\end{cases}\tag{1.110}$$

If the method for obtaining a unique upwind function used in Subsection 1.3.2 is now applied, the optimal choice is

$$\bar{\alpha}^1(\gamma)=\left(\frac{3}{2\gamma^2}+\frac{1}{2}\right)\left(\coth\gamma-\frac{1}{\gamma}\right)-\frac{1}{2\gamma}$$

and its asymptotic approximation is

$$\bar{\alpha}^1_a(\gamma)=\begin{cases}\frac{17}{240}\gamma & \text{if } 0\le|\gamma|\le\frac{120}{17}\\ \frac{1}{2} & \text{if } |\gamma|>\frac{120}{17}\end{cases}$$

since $\bar{\alpha}^1=\frac{17}{240}\gamma+O(\gamma^3)$ in the neighborhood of $\gamma=0$ and $\bar{\alpha}^1\to\frac{1}{2}$ as $\gamma\to\infty$.

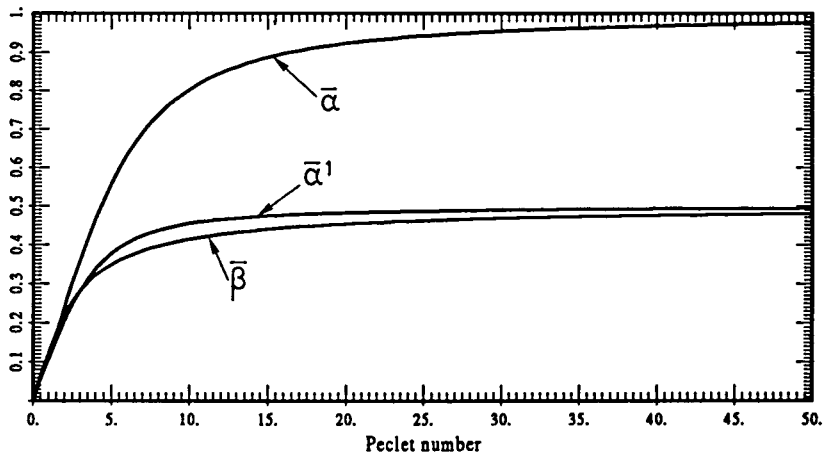

Figure 1.8 Upwind functions for quadratic elements using the GLS method

In Figure 1.8 the functions $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\alpha}^1$ have been plotted.

1.3.5 Introduction of source terms

Up to now, we have only considered the homogeneous equation (1.66), i.e., with $f(x) \equiv 0$. We have found the upwind functions that give nodally exact solutions for three different cases using quadratic elements, namely, the SD method using the canonical and the hierarchic bases and the GLS method using the canonical basis. In Section 1.1 it was also explained how nodally exact solutions could be obtained using linear elements. Now we can prove that for certain functions $f(x)$ we still do have solutions exact at the nodes.

Recall the definition of the function spaces Ψ_h and Φ_h given by (1.27) and (1.28), respectively. Let $\hat{a}$ and $\hat{l}$ be the linear forms that define the variational method employed (Galerkin, SD, GLS, etc.) and consider the following three problems:

(P.1) Case $g \neq 0$, $f \neq 0$: Find $\phi_{h,1} \in \Phi_h$ such that $\hat{a}(\phi_{h,1}, \psi_h) = \hat{l}(\psi_h) \quad \forall \psi_h \in \Psi_h$.
(P.2) Case $g = 0$, $f \neq 0$: Find $\phi_{h,2} \in \Psi_h$ such that $\hat{a}(\phi_{h,2}, \psi_h) = \hat{l}(\psi_h) \quad \forall \psi_h \in \Psi_h$.
(P.3) Case $g \neq 0$, $f = 0$: Find $\phi_{h,3} \in \Phi_h$ such that $\hat{a}(\phi_{h,3}, \psi_h) = 0 \quad \forall \psi_h \in \Psi_h$.

As before, g denotes the prescribed Dirichlet boundary condition. The following discrete space will also be needed:

$$H_h := \{\psi_h \in H^1(\Omega) \mid \psi_h|_{\Omega^e} \in P_m(\Omega^e)\} \tag{1.111}$$

The continuous problems corresponding to P.1, P.2 and P.3 are simply obtained by replacing the spaces Ψ_h and Φ_h by Ψ and Φ, respectively (cf. (1.22) and (1.23)). The solution of these problems will be denoted by dropping the subscript h in $\phi_{h,i}$, $i = 1, 2, 3$.

Now, let $\pi_h : H^1(\Omega) \longrightarrow H_h$ be the canonical projection onto the finite element space H_h, defined by $\pi_h(\psi) = \tilde{\psi}_h$, the finite element interpolant of ψ. A solution of any of the problems P.1, P.2, P.3 will be nodally exact whenever $\phi_{h,i} = \pi_h(\phi_i)$, $i = 1, 2, 3$.

So far, we only know how to get nodally exact solutions for problem P.3 when dealing with Eqn. (1.66) with $f(x) \equiv 0$. But we can prove the following:

Theorem 1.2 *Assume that for all functions g the solution of problem P.3 is nodally exact, i.e., $\phi_{h,3} = \pi_h(\phi_3)$. If the function f is such that there exists $\omega_h \in H_h$ satisfying*

$$\hat{a}(\omega_h, \psi) = \hat{l}(\psi) \quad \forall \psi \in H^1(\Omega) \tag{1.112}$$

then the solution of P.1 is also nodally exact.

Proof: We have to prove that $\phi_{h,1} = \pi_h(\phi_1)$. Observe first that

$$\phi_1 = \phi_2 + \phi_3 \quad \text{and} \quad \phi_{h,1} = \phi_{h,2} + \phi_{h,3}$$

Since $\phi_{h,3} = \pi_h(\phi_3)$ and π_h is linear, we will have that $\phi_{h,1} = \pi_h(\phi_1)$ iff $\phi_{h,2} = \pi_h(\phi_2)$. By condition (1.112) and using the fact that $\Psi_h \subset H_h$, $\phi_{h,2}$ will be the solution of the problem: Find $\phi_{h,2} \in \Psi_h$ such that

$$\hat{a}(\phi_{h,2} - \omega_h, \psi_h) = 0 \quad \forall \psi_h \in \Psi_h$$

and ϕ_2 the solution of a similar problem replacing Ψ_h by Ψ. Define now $\delta_h := \phi_{h,2} - \omega_h$, $\delta := \phi_2 - \omega_h$ and $g := -\omega_h$. The function δ_h will be the solution of a problem of type P.3: Find $\delta_h \in \Phi_h$ such that

$$\hat{a}(\delta_h, \psi_h) = 0 \quad \forall \psi_h \in \Psi_h$$

and similarly for δ. By hypothesis, $\delta_h = \pi_h(\delta)$ and this gives

$$\begin{aligned} \pi_h(\phi_2) &= \pi_h(\delta + \omega_h) && \text{(definition of } \delta\text{)} \\ &= \pi_h(\delta) + \omega_h && (\pi_h \text{ is linear and } \omega_h \in H_h) \\ &= \delta_h + \omega_h && (\delta_h \text{ is nodally exact}) \\ &= \phi_{h,2} && \text{(definition of } \delta_h\text{)} \end{aligned}$$

i.e., $\phi_{h,2}$, and hence $\phi_{h,1}$, will be exact at the nodes. □

Roughly speaking, condition (1.112) means that the equation of the continuous problem has a solution, not necessarily satisfying the boundary conditions, that belongs to the space of interpolation functions. We can now apply this general result to the problem that has been considered throughout this section:

Theorem 1.3 *Let problem (1.66)–(1.67) be solved numerically by one of the following four methods described above:*

(i) The SD method using linear elements and the upwind function (1.12)
(ii) The SD method using standard quadratic elements and the upwind functions (1.83), (1.82)
(iii) The SD method using hierarchic quadratic elements and the upwind functions (1.97), (1.96)
(iv) The GLS method using standard quadratic elements and the upwind functions (1.106), (1.105)

Then, if $f(x)$ is constant, the numerical solution is nodally exact in all the cases. If $f(x)$ is continuous and piecewise linear, the solution is nodally exact for methods (ii)–(iv).

Proof: We know from the results of this section that all the methods yield nodally exact solutions when $f(x) \equiv 0$. Suppose now that within each element f has the form $f(x) = ax + b$, with a and b constants. In this situation, for $u \neq 0$ the general solution of Eqn. (1.66) is:

$$\phi(x) = C_1 + C_2 \exp\left(\frac{u}{k}x\right) + \frac{a}{2u}x^2 + \left(\frac{b}{u} + \frac{ak}{u^2}\right)x$$

for each element, where C_1 and C_2 are constants to be determined from the boundary conditions and the continuity of ϕ. Setting $C_2 = 0$ and choosing C_1 in order to have continuity, we get a function that belongs to the interpolation space of quadratic finite elements. Thus, from Theorem 1.2 it follows that methods (ii)–(iv) will yield nodally exact solutions. When $a = 0$ we get a linear function. The solutions obtained using (i) will also be nodally exact in this case. □

1.4 Numerical implementation

In order to compute the intrinsic time given by (1.36) for each element in the multi-dimensional convection-diffusion equation (1.19), the values of h^e, k^e and $\mathbf{u}^e$ that give the Péclet number (1.30) are needed. We must also know which is the expression of the upwind function $\alpha^e = \alpha(\gamma^e)$ that corresponds to the node under consideration.

We compute the velocity $\mathbf{u}^e$ simply as the average of the nodal velocities of the element and k^e as the diffusion along the flow direction. Since we have assumed that $\mathbf{k}$ in (1.19) is a second order tensor, this diffusion will be

$$k^e = \frac{\mathbf{u}^e \cdot \mathbf{k} \cdot \mathbf{u}^e}{|\mathbf{u}^e|^2} \tag{1.113}$$

This value will be positive since $\mathbf{k}$ is positive-definite.

The computation of h^e and the choice of the upwind function will be explained in more detail.

1.4.1 The characteristic length

To simplify the notation we will consider the two-dimensional case, although what follows is completely general.

Let $\mathcal{D}$ be a convex domain in $\mathbb{R}^2$ transformed into $\mathcal{D}' \subset \mathbb{R}^2$ by an affine mapping $\mathbf{f} = (f_1, f_2)$.

Using the notation of Figure 1.9, let

$$\ell = |B - A|, \qquad \ell' = |B' - A'| \tag{1.114}$$

and $\mathbf{v}' = (\mathbf{Df})\mathbf{v}$, where $\mathbf{Df}$ is the Jacobian matrix of $\mathbf{f}$. Since

$$\begin{aligned} \mathbf{f}(B) &= \mathbf{f}(A) + \ell' \frac{\mathbf{v}'}{|\mathbf{v}'|} \\ &= \mathbf{f}(A) + (\mathbf{Df})(B - A) \end{aligned} \tag{1.115}$$

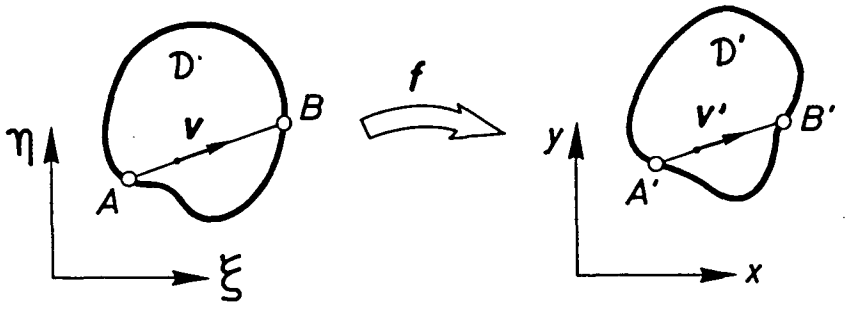

Figure 1.9 Transformation of a domain in $\mathbb{R}^2$ by an affine mapping

we have that

$$\ell' \frac{\mathbf{v}'}{|\mathbf{v}'|} = (\mathbf{Df})(B - A)$$

and multiplying this equation by $\mathbf{Df}^{-1}$ we get

$$\ell'(\mathbf{Df})^{-1}\mathbf{v}' = |\mathbf{v}'|(B - A) \tag{1.116}$$

Taking the Euclidian norm on both sides of (1.116) and considering that $\mathbf{Df}^{-1}\mathbf{v}' = \mathbf{v}$ we finally get

$$\ell' = \frac{|\mathbf{v}'|}{|\mathbf{v}|}\ell \tag{1.117}$$

Formula (1.117) allows to compute the characteristic length in the flow direction as

$$h^e = \frac{|\mathbf{u}^e|}{|\mathbf{u}_0^e|}h_0 \tag{1.118}$$

where subscript naught indicates that the value corresponds to the parent domain of the element with 'natural' coordinates (ξ, η). Equation (1.118) reduces the computation of h^e to that of h_0, which can be easily estimated since the geometry is now very simple. In our computations we have taken, for the parent domains of Figure 1.10:

$h_0 = 2$ for quadrilateral elements
$h_0 = 0.7$ for triangular elements

Remarks 1.6

(1) The length h^e defined by (1.118) depends on the point (x, y) of Ω^e. Thus, it will be numerically different at each integration point. Also, the exact value of h_0 depends on each point, although the assumption of a constant value seems reasonable.

(2) From (1.115) it can be seen that (1.118) will be exact whenever the mapping $\mathbf{f}$ can be considered affine. This will always be the case with straightsided triangles and parallelograms in two dimensions. □

1.4.2 Assignment of upwind functions

In Section 1.3, the expressions of the upwind functions α and β for quadratic elements were obtained. The weighting function of a certain node of an element will be obtained

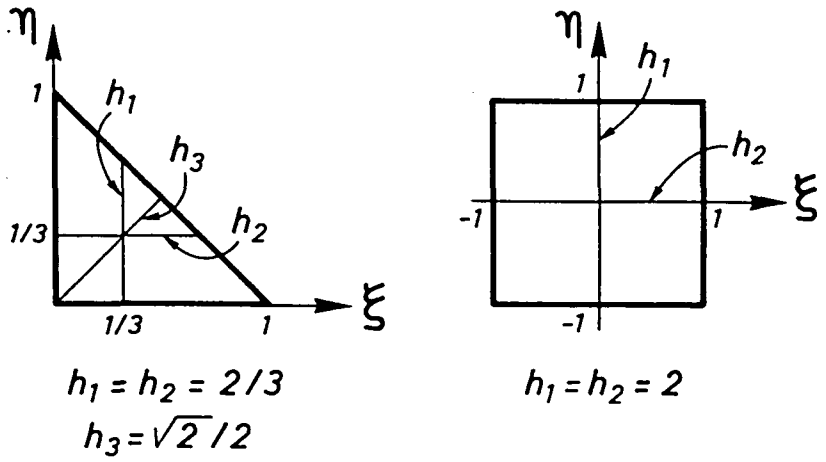

Figure 1.10 Parent domains for triangular and quadrilateral elements

using α or β depending on the position of the node. Clearly, in multidimensional situations this position is relative to the direction of the flow, which complicates the definition of a node as 'extreme' or 'central'. This, of course, is an important drawback for the use of different upwind functions.

The heuristic criterion we have followed is based on the assignment of upwind functions taking into account whether a node is extreme or central for certain directions of the flow. For 2D elements, we have taken these directions as those defined by the coordinates ξ, η (see Figure 1.11) for the nine-noded Lagrangian element and those defined by the area coordinates $1-\xi-\eta$, ξ and η for the six-noded triangle. For the corner nodes of the elements the function α has been chosen and for the interior node of the nine-noded element the function β. The problem arises when the upwind function for the midside nodes must be determined. For example, the shape function of node 5 for the nine-noded element (see Figure 1.11) along the $\eta = -1$ line corresponds to the shape functions of node 2 in Figure 1.1, i.e. a central node, whereas along the $\xi = 0$ line the corresponding shape function is that of node 1 in Figure 1.1, an extreme node. So, the upwind function of node 5, say δ_5, will be taken as a combination of functions α and β. In Figure 1.11, the nodal numbering in the parent domain and the chosen upwind functions are indicated.

The best numerical results have been obtained taking δ_i as the functions

$$\delta_i = f_i(\theta)\alpha + [1 - f_i(\theta)]\beta \tag{1.119}$$

where for the six-noded element

$$\begin{aligned} f_4(\theta) &= \sin^2\theta \\ f_5(\theta) &= f_4(\theta + \tfrac{\pi}{4}) \\ f_6(\theta) &= \cos^2\theta \end{aligned} \tag{1.120}$$

and for the nine noded element

$$\begin{aligned} f_5(\theta) &= f_7(\theta) = \sin^2\theta \\ f_6(\theta) &= f_8(\theta) = \cos^2\theta \end{aligned} \tag{1.121}$$

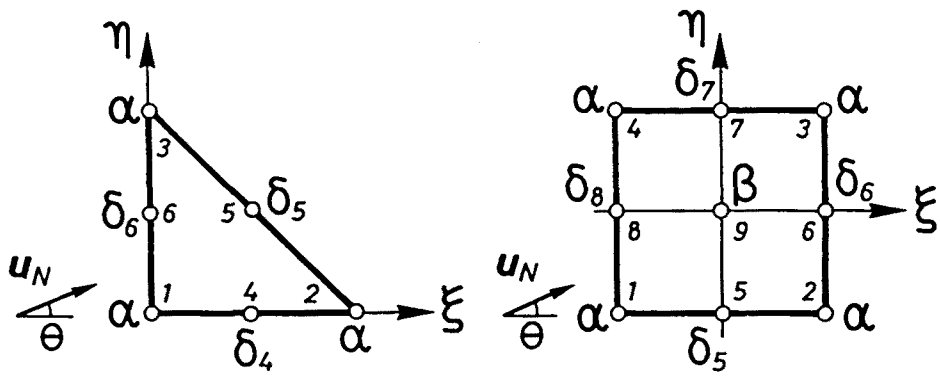

Figure 1.11 Assignment of upwind functions for the 6-noded and 9-noded elements

In (1.119)–(1.121), θ is the angle shown in Figure 1.11. Clearly, these expressions yield the expected upwind functions in the cases $\theta = 0$ and $\theta = \frac{\pi}{2}$.

1.5 Numerical examples

In this section, some very simple tests cases are presented in order to assess the performance of quadratic finite elements for stationary convection-diffusion problems when the SD method is employed. In all the cases, the standard finite element interpolation has been used. For another simple example, not presented here, the reader may consult [CO2]. Results obtained using the standard Galerkin formulation are also presented in this Reference.

Example 1.1 In this example we solve the one dimensional problem (1.66)–(1.67) with $u = 1$, $k = 0.01$, $f(x) = \sin(\pi x)$, $\ell = 1$ and $\phi_0 = \phi_\ell = 0$. The interval $[0, 1]$ is discretized using ten quadratic elements of equal length 0.1. This gives the value $\gamma = 5$ for the Péclet number. The analytical solution is

$$\phi(x) = C_1 + C_2\, e^{\frac{u}{k}x} + \frac{k}{u^2 + k^2\pi^2}\,[\sin(\pi x) - \frac{u}{k\pi}\cos(\pi x)] \tag{1.122}$$

with

$$C_2 = \frac{2u}{(u^2\pi + k^2\pi^3)(1 - e^{-\frac{u}{k}})} \tag{1.123}$$

$$C_1 = -\frac{1}{2}(1 + e^{\frac{u}{k}})C_2 \tag{1.124}$$

The hypothesis of Theorem 1.2 is not fulfilled and in fact the nodal values of the numerical solution are not exact. However, the use of the optimal upwind functions of (1.82) and (1.83) gives results (Figure 1.12.*a*) that cannot be distinguished from those of the

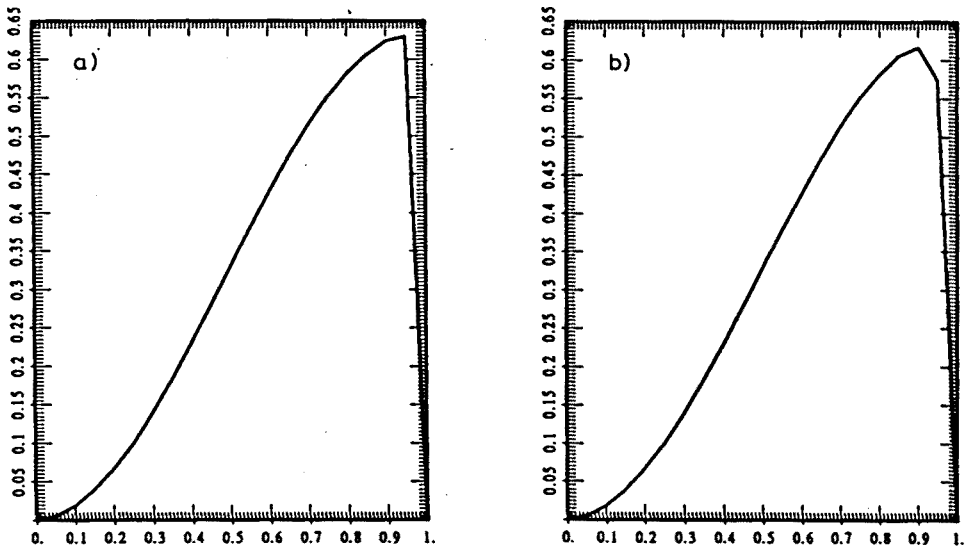

Figure 1.12 Solutions of Example 1.1. *a*) Using the upwind functions (1.82) and (1.83). *b*) Using the unique upwind function (1.93)

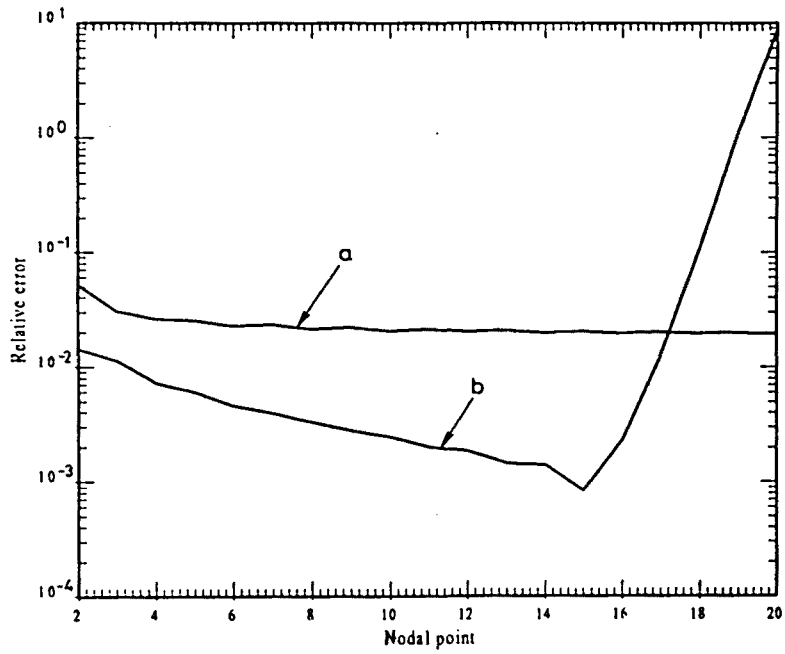

Figure 1.13 Relative errors for the solutions *a* and *b* of Figure 1.12

analytical solution (linear interpolation between nodes has been used in the plots). In Figure 1.12.*b* the solution obtained using the unique function (1.93) is plotted and Figure 1.13 shows the relative error (in percentage) obtained using the two methods.

We observe that the use of (1.93) gives a solution that smooths the right boundary layer at $x = 1$, but that the error is very small far from this coordinate.

Example 1.2 In this example, problem (1.19)-(1.21) is solved. The data are:

$$\Omega = \left]\frac{-1}{2}, \frac{1}{2}\right[\times \left]\frac{-1}{2}, \frac{1}{2}\right[$$
$$\Gamma_D = \partial\Omega, \quad \Gamma_N = \emptyset$$
$$\mathbf{u}(x,y) = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$$
$$k_{ij}(x,y) = 2 \cdot 10^{-2}\delta_{ij}$$
$$f(x,y) = 5$$
$$g(x,y) = 0$$

The domain Ω has been discretized using a uniform finite element mesh with 21×21 nodes in all the cases. The resulting Péclet number is $\gamma = 2.5$ for quadratic elements ($h \approx 0.1$) and $\gamma = 1.25$ for linear elements ($h \approx 0.05$). This example was chosen for testing the adopted expressions (1.119)-(1.121). Results obtained with quadratic and linear quadrilaterals and triangular elements and using the optimal upwind functions of (1.82) and (1.83) are shown in Figure 1.14. The results obtained for quadratic elements are almost the same as for linear elements and in all the cases very accurate.

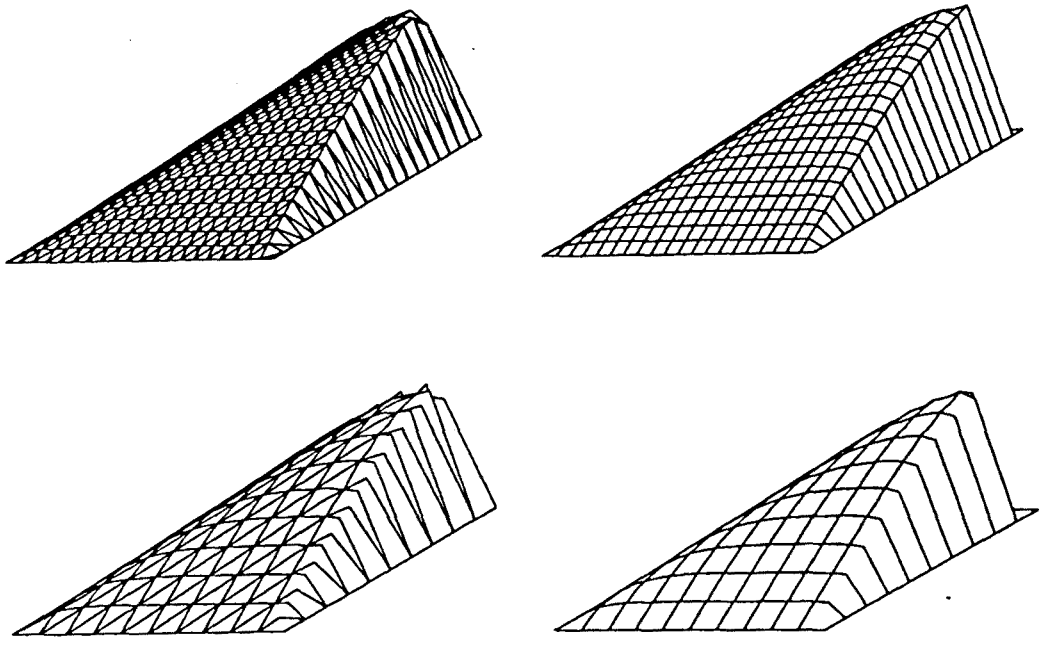

Figure 1.14 Results of Example 1.2 using triangular (3 and 6 nodes) and quadrilateral (4 and 9 nodes) elements

Example 1.3 This example and the following have been taken from reference [HMM].

Now, the data for Eqn. (1.19) are:

$$\Omega = \left]\frac{-1}{2},\frac{1}{2}\right[\times \left]\frac{-1}{2},\frac{1}{2}\right[- \{0\} \times \left]\frac{-1}{2},0\right]$$

$$\Gamma_D = \partial\Omega, \quad \Gamma_N = \emptyset$$

$$\mathbf{u}(x,y) = (-y,x)$$

$$k_{ij}(x,y) = 10^{-8}\delta_{ij}$$

$$f(x,y) = 0$$

$$g(x,y) = \begin{cases} \sin\pi(1+2y) & \text{if } x = 0 \text{ and } -\frac{1}{2} \le y \le 0 \\ 0 & \text{else} \end{cases}$$

In all the cases, 31×31 nodal points and a uniform finite element mesh have been used. For the small diffusion considered, the solution of this problem is just the advection of the sine profile. The objective of this problem was only to test the accuracy of the algorithm, since the exact solution is very smooth and the Galerkin method only produces small amplitude oscillations. Results obtained with different quadrilateral and triangular elements using the optimal upwind functions are depicted in Figure 1.15. Similar accuracy is obtained in all the cases.

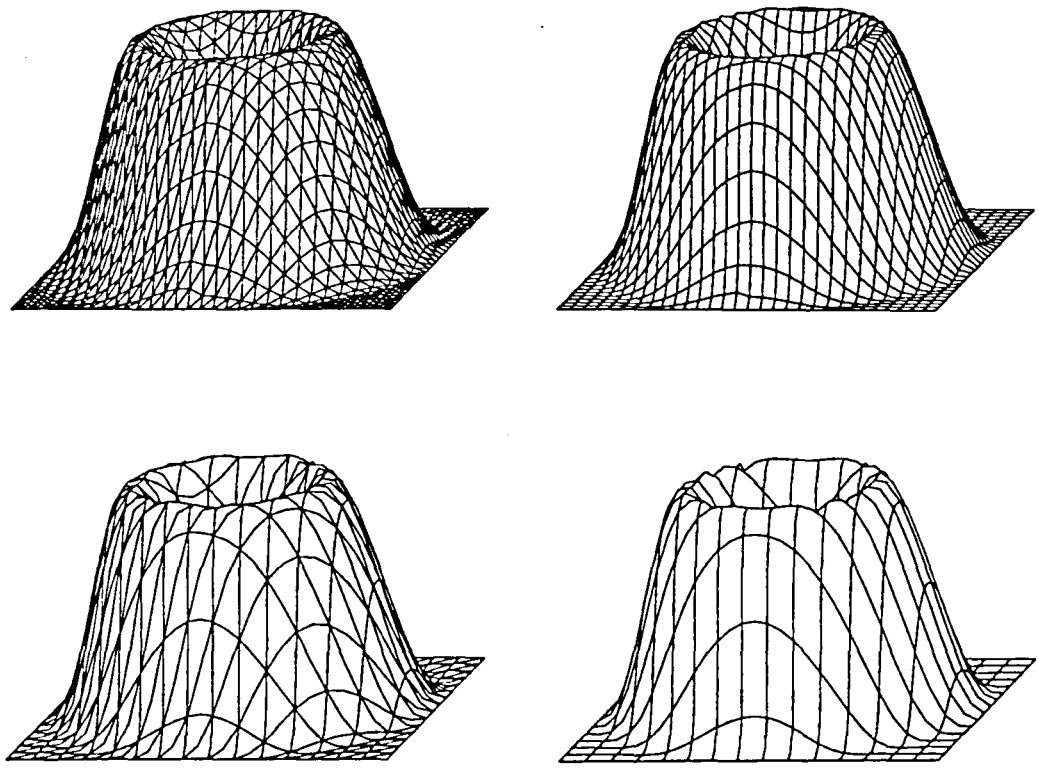

Figure 1.15 Results of Example 1.3 using triangular (3 and 6 nodes) and quadrilateral (4 and 9 nodes) elements

Example 1.4 Again, the steady-state problem (1.19)–(1.21) is solved, now with:

$$\Omega = \left]\frac{-1}{2},\frac{1}{2}\right[\times \left]\frac{-1}{2},\frac{1}{2}\right[$$
$$\Gamma_D = \partial\Omega, \quad \Gamma_N = \emptyset$$
$$\mathbf{u}(x,y) = (\cos\theta, -\sin\theta)$$
$$k_{ij}(x,y) = 10^{-6}\delta_{ij}$$
$$f(x,y) = 0$$
$$g(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \Gamma_{D1} \\ 0 & \text{if } (x,y) \in \Gamma_{D2} \end{cases}$$

with

$$\Gamma_{D1} = \left\{-\frac{1}{2}\right\} \times \left[\frac{1}{4},\frac{1}{2}\right] \cup \left]-\frac{1}{2},\frac{1}{2}\right[\times \left\{\frac{1}{2}\right\}$$
$$\Gamma_{D2} = \Gamma_D \backslash \Gamma_{D1}$$

This problem shows the inability of the SD formulation to preclude overshoots and undershoots when sharp layers are present.

We have solved this problem with the angles θ given by $\tan\theta = \frac{1}{2}$, 1 and 2. The results shown in Figures 1.16, 1.17 and 1.18 correspond to the latter case, when overshoots and undershoots are more important. However, it is seen that they are bigger using linear elements than using quadratic elements. The solution obtained using the upwind functions (1.82) and (1.83) together with (1.119)–(1.121) looks better than that obtained with the single upwind function (1.93), although the different computational effort must be also considered.

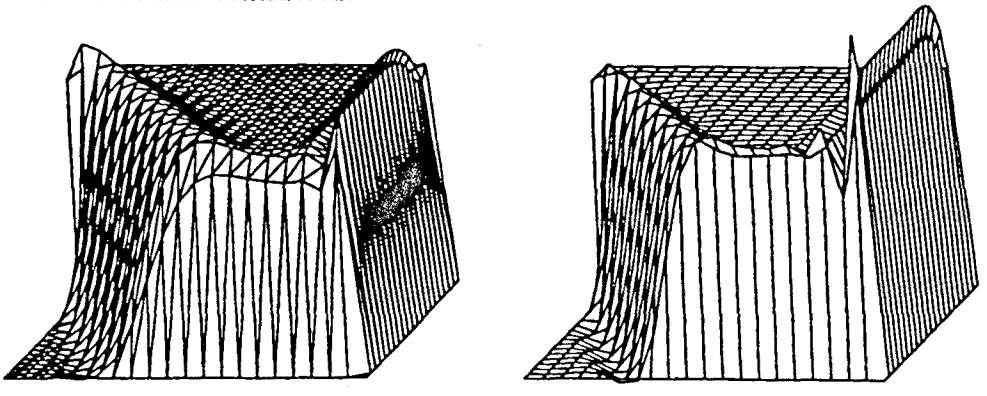

Figure 1.16 Results of Example 1.4 using 3-noded linear triangular and 4-noded bilinear quadrilateral elements

1.6 Summary and conclusions

In this chapter, a complete description of the Streamline Diffusion method for solving the stationary convection-diffusion equation has been presented. The motivation of the method has been shown for a very simple problem. Nevertheless, all the features of

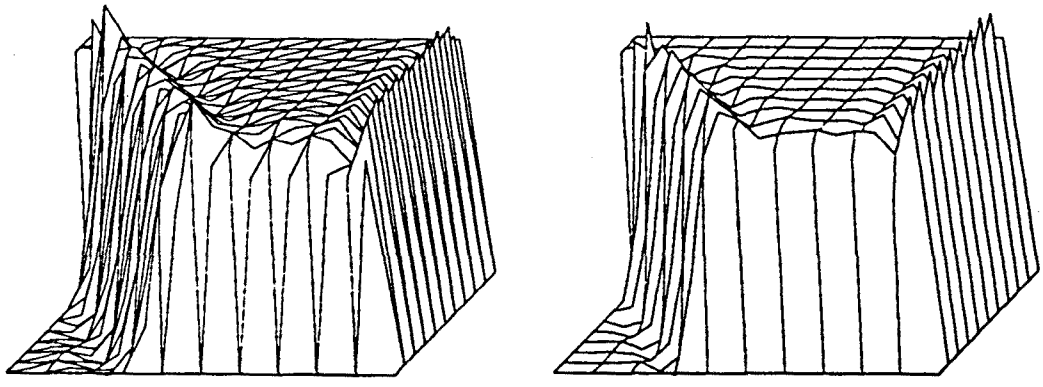

Figure 1.17 Results of Example 1.4 using quadratic triangular (6 nodes) and biquadratic quadrilateral (9 nodes) elements with the upwind functions (1.82) and (1.83)

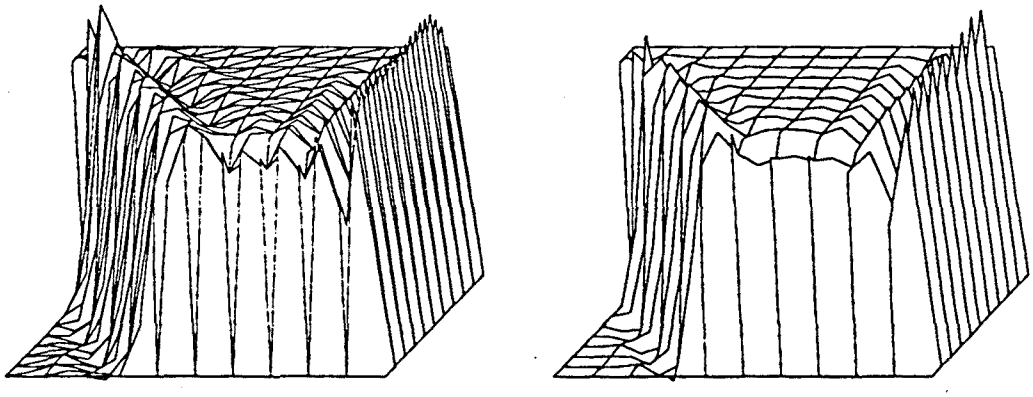

Figure 1.18 Results of Example 1.4 using quadratic triangular (6 nodes) and biquadratic quadrilateral (9 nodes) elements with the upwind function (1.93)

the misbehavior of the standard Galerkin method are present in this case. The basic literature for further discussion has also been given.

All the contributions introduced here concern the accurate calculation of the intrinsic time and, in particular, the upwind functions for quadratic elements in different cases. Numerical experiments have shown that the proposed methodology is effective. More confidence on it will also be acquired in the following chapters.

Summarizing, the specific items that have been treated are:

- *Convergence analysis.* This analysis dictated which must be the asymptotic behavior of the upwind functions. It has been seen that they have to behave as follows:

$$\alpha(\gamma) = \begin{cases} C_1\gamma & \text{as } \gamma \to 0 \\ C_2 & \text{as } \gamma \to \infty \end{cases}$$

  where γ is the Péclet number.
- *Optimal upwind functions for 1D quadratic elements.* They have been obtained for different cases. Their expressions are summarized in Box 1.1, as well as their asymptotic behavior. The functions for the extreme nodes are denoted by α in all the cases, and the functions for the central nodes by β. When a unique upwind function is to be used, it is indicated by α^1.

- *Introduction of source terms.* It has been proved for 1D problems using quadratic elements that if the source term is piecewise linear (and continuous), nodally exact results will be obtained whenever they can be found for the homogeneous equation.
- *Characteristic length.* The following expression for computing this parameter for each element has been introduced:

$$h^e = \frac{|\mathbf{u}^e|}{|\mathbf{u}_0^e|} h_0$$

  Here, $\mathbf{u}^e$ is the characteristic velocity of the element and subscript naught refers to values in the parent domain.
- *Assignment of upwind function.* A heuristic criterion has been proposed in order to compute the upwind functions taking into account the flow direction. Its performance has been checked through numerical experiments.

Box 1.1 Upwind functions for quadratic elements

Method: SD, standard basis

Expression	Limit as $\gamma \to \infty$	Behavior as $\gamma \to 0$
$\beta = \frac{1}{2}\left(\coth\frac{\gamma}{2} - \frac{2}{\gamma}\right)$	$\frac{1}{2}$	$\frac{\gamma}{12} + O(\gamma^3)$
$\alpha = \frac{(3+3\gamma\beta)\tanh\gamma - (3\gamma+\gamma^2\beta)}{(2-3\beta\tanh\gamma)\gamma^2}$	1	$\frac{\gamma}{12} + O(\gamma^3)$
$\alpha^1 = \frac{1}{2}\left(\coth\gamma - \frac{1}{\gamma}\right)$	$\frac{1}{2}$	$\frac{\gamma}{6} + O(\gamma^3)$

Method: SD, hierarchic basis

Expression	Limit as $\gamma \to \infty$	Behavior as $\gamma \to 0$
$\beta = \frac{1}{2} + \frac{1}{e^\gamma - 1} - \frac{1}{\gamma}$	$\frac{1}{2}$	$\frac{\gamma}{12} + O(\gamma^2)$
$\alpha = \left(1 + \frac{1}{\gamma\beta}\right)\coth\gamma - \left(\frac{1}{3\beta} + \frac{1}{\gamma^2\beta} + \frac{1}{\gamma}\right)$	$\frac{1}{3}$	$\frac{\gamma}{15} + O(\gamma^2)$
$\alpha^1 = \frac{1}{2}\left(\coth\gamma - \frac{1}{\gamma}\right)$	$\frac{1}{2}$	$\frac{\gamma}{6} + O(\gamma^3)$

Method: GLS, standard basis

Expression	Limit as $\gamma \to \infty$	Behavior as $\gamma \to 0$
$\beta = \frac{\gamma^2\left(\coth\frac{\gamma}{2} - \frac{2}{\gamma}\right)}{6 - 3\gamma\coth\frac{\gamma}{2} + 2\gamma^2}$	$\frac{1}{2}$	$\frac{\gamma}{9} + O(\gamma^3)$
$\alpha = \frac{\tanh\gamma\left(3 + \gamma^2 + 6\gamma\beta + 9\frac{\beta}{\gamma}\right) - (3\gamma + 9\beta + \gamma^2\beta)}{2\gamma^2 - 3\beta\gamma^2\tanh\gamma}$	1	$\frac{\gamma}{9} + O(\gamma^3)$
$\alpha^1(\gamma) = \left(\frac{3}{2\gamma^2} + \frac{1}{2}\right)\left(\coth\gamma - \frac{1}{\gamma}\right) - \frac{1}{2\gamma}$	$\frac{1}{2}$	$\frac{17\gamma}{240} + O(\gamma^3)$

References

[Ax] O. Axelson. Stability and error estimates of Galerkin finite element approximations for convection-diffusion equations. *IMA J. Numer. Anal.*, vol. 1 (1981), 329–345

[BBF] R.E. Bank, J.F. Burgler, W. Fichtner and R.K. Smith. Some upwinding tech-

niques for finite element approximations of convection-diffusion equations. *Numer. Math.*, vol. 58 (1990), 185–202

[Ba] K.E. Barret. The numerical solution of singular-perturbation boundary-value problems. *Q. Jl. Mech. appl. Math.*, vol. 27 (1974), 57–68

[BH] A.N. Brooks and T.J.R. Hughes. Streamline Upwind/Petrov-Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 32 (1982), 199–259

[CJ] G.F. Carey and B.N. Jiang. Least-squares finite elements for first-order hyperbolic systems. *Int. J. Numer. Meth. Engrg.*, vol. 26 (1988), 81–93

[CM] I. Christie and A.R. Mitchell. Upwinding of high order Galerkin methods in conduction-convection problems. *Int. J. Numer. Meth. Engrg.*, vol. 14 (1978), 1764–1771

[CGM] I. Christie, D.F. Griffiths, A.R. Mitchell and O.C. Zienkiewicz. Finite element methods for second order differential equations with significant first derivatives. *Int. J. Numer. Meth. Engrg.*, vol. 10 (1976), 1389–1396

[Ci] P.G. Ciarlet. *The finite element method for elliptic problems.* (North-Holland, 1978)

[Co] R. Codina. *Dues formulacions numèriques per al problema de flux incompressible* (in Catalan). Grade thesis, Universitat Politècnica de Catalunya (1989)

[CO1] R. Codina, E. Oñate and M. Cervera. The intrinsic time for the SUPG formulation using quadratic elements. *Comput. Meths. Appl. Mech. Engrg.*, vol. 94 (1992), 239–262

[CO2] R. Codina, E. Oñate, M. Cervera and K. Eckstein. Una formulación de Petrov-Galerkin para el análisis de problemas de convección-difusión con elementos finitos cuadráticos (in Spanish). *Proc. of the First Conference on Numerical Methods in Engineering, SEMNI*, Gran Canaria, Spain, 1990.

[Do] J. Donea. A Taylor-Galerkin method for convective transport problems. *Int. J. Numer. Meth. Engrg.*, vol. 20 (1984), 101–119

[DBS] J. Donea, T. Belytschko and P. Smolinski. A generalized Galerkin method for steady convection-diffusion problems with application to quadratic shape function elements. *Comput. Meths. Appl. Mech. Engrg.*, vol. 48 (1985), 25–43

[DR] J. Douglas and T.F. Russell. Numerical methods for convection dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. *SIAM J. Numer. An.*, vol. 19 (1982), 871–885

[FH] L. Franca and T.J.R. Hughes. Two classes of mixed finite element methods. *Comput. Meths. Appl. Mech. Engrg.*, vol. 69 (1988), 89–129

[GH] P.P.N. de Groen and P.W. Hemker. Error bounds for exponentially fitted Galerkin methods applied to stiff two-point boundary-value problems, in: *Numerical analysis of singular perturbation problems*, P.W. Hemker and J.J.H. Miller (eds.) (Academic Press, 1979)

[GP] W.G. Gray and G.F. Pinder. An analysis of the numerical solution of the transport equation. *Water Resources Research*, vol. 12 (1976), 547–555

[He] J.C. Heinrich. On quadratic elements in finite element solutions of steady-state convection-diffusion equations. *Int. J. Numer. Meth. Engrg.* vol. 15 (1980), 1041–1052

[HHZ] J.C. Heinrich, P.S. Huyakorn and O.C. Zienkiewicz. An 'upwind' finite element

scheme for two-dimensional convective transport equation. *Int. J. Numer. Meth. Engrg.*, vol. 11 (1977), 131–143

[HZ1] J.C. Heinrich and O.C. Zienkiewicz. Quadratic finite element schemes for two-dimensional convective-transport problems. *Int. J. Numer. Meth. Engrg.*, vol. 11 (1977), 1831–1844

[HZ2] J.C. Heinrich and O.C. Zienkiewicz. The finite element method and 'upwinding' techniques in the numerical solution of convection dominated flow problems, in: *Finite Element Methods for Convection Dominated Flows*, T.J.R. Hughes (ed.) ASME, New York (1979)

[Hu1] T.J.R. Hughes. A simple scheme for developing 'upwind' finite elements. *Int. J. Numer. Meth. Engrg.*, vol. 12 (1978), 1359–1365

[Hu2] T.J.R. Hughes. Recent progress in the development and understanding of SUPG methods with special reference to the compressible Euler and Navier-Stokes equations, in: *Finite Elements in Fluids*, vol. 7, R.H. Gallagher; R. Glowinski, P.M. Gresho, J.T. Oden and O.C. Zienkiewicz (eds.) (1987)

[HB1] T.J.R. Hughes and A. Brooks. A multi-dimensional upwind scheme with no crosswind diffusion, in: *FEM for convection dominated flows*, T.J.R. Hughes (ed.) ASME, New York (1979)

[HB2] T.J.R. Hughes and A.N. Brooks. A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: applications to the streamline upwind procedure, in: *Finite Elements in Fluids*, R.H. Gallagher, D.M. Norrie, J.T. Oden and O.C. Zienkiewicz (eds.), vol. IV, (Wiley, London, 1982) 46–65

[HF] T.J.R. Hughes and L.P. Franca. A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Meths. Appl. Mech. Engrg.*, vol. 65 (1987), 85–96

[HM] T.J.R. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective-diffusive systems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 58 (1986), 305–328

[HFB] T.J.R. Hughes, L.P. Franca and M. Balestra. A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuška-Brezzi condition: a stable Petrov- Galerkin formulation for the Stokes problem accommodating equal- order interpolations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 59 (1986), 85–99

[HFH] T.J.R. Hughes, L.P. Franca and G.M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 73 (1989), 173–189

[HFM] T.J.R. Hughes, L.P. Franca and M. Mallet. A new finite element formulation for computational fluid dynamics: VI. Convergence analysis of the generalized SUPG formulation for linear time-dependent multidimensional advective-diffusive systems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 63 (1987), 97–112

[HMM] T.J.R. Hughes, M. Mallet and A. Mizukami. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Comput. Meths. Appl. Mech. Engrg.*, vol. 54 (1986), 341–355

[IK] Isaacson, E. and H.E. Keller. *Analysis of numerical methods.* (John Wiley &

Sons, 1966)

[Jo1] C. Johnson. Finite element methods for convection-diffusion problems, in: *Computing methods in applied sciences and engineering*, R. Glowinski and J.L. Lions (eds.) (North-Holland, 1982)

[Jo2] C. Johnson. *Numerical solution of partial differential equations by the finite element method.* (Cambridge University Press, 1986)

[Jo3] C. Johnson. Adaptive finite element methods for diffusion and convection problems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 82 (1990), 301–322

[JNP] C. Johnson, U. Nävert and J. Pitkäranta. Finite element methods for linear hyperbolic equations *Comput. Meths. Appl. Mech. Engrg.*, vol. 45 (1984), 285–312

[KNZ] D.W. Kelly, S. Nakazawa, O.C. Zienkiewicz and J.C. Heinrich. A note on upwinding and anisotropic balancing dissipation in finite element approximations to convective diffusion problems. *Int. J. Numer. Meth. Engrg.*, vol. 15 (1980), 1705–1711

[Ki] F. Kikuchi. Discrete maximum principle and artificial viscosity in finite element approximations of convective diffusion equations. *ISAS Report No. 550* (vol. 42, No. 5), Tokyo (1977)

[LPZ] J.H.W. Lee, J. Peraire and O.C. Zienkiewicz. The characteristic-Galerkin method for advection-dominated problems- An Assessment. *Comput. Meths. Appl. Mech. Engrg.*, vol. 61 (1987), 359–369

[Na] U. Nävert. *A finite element method for convection-diffusion problems.* Thesis. Chalmers University of Technology, Göteborg, Sweden (1982)

[NRe] H. Nguyen and J. Reynen. A space-time least-square finite element scheme for advection-diffusion equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 42 (1984), 331–342

[NRi] J. von Neumann and R.D. Richtmyer. A method for the numerical calculation of hydrodynamical shocks. *J. Appl. Phys.*, vol. 21 (1950), 232

[Pi] O. Pironneau. *Finite element methods for fluid flow.* (John Wiley & Sons, 1989)

[RM] R.D. Richtmyer and K.W. Morton. *Difference methods for initial value problems.* (Interscience, New York, 1967)

[Ro] P.J. Roache. On artificial viscosity. *J. Comput. Phys.*, vol. 10 (1972), 169–184

[Sh] F. Shakib. *Finite element analysis of the compressible Euler and Navier-Stokes equations.* Ph.D. Thesis. Stanford University (1988).

[VF] E. Varoglu and W.D. Finn. A finite element method for the diffusion-convection equation with constant coefficients. *Advances in Water Resources*, vol. 1 (1978), 337–343

[Wa] L.B. Wahlbin. A dissipative Galerkin method applied to some quasilinear hyperbolic equations. *RAIRO Numer. Anal.*, vol. 8:2 (1974), 109–117

CHAPTER 2

TRANSIENT ALGORITHMS— STABILITY ANALYSIS OF AN EXPLICIT SCHEME

2.1 Introduction

In the previous chapter, we have only considered the steady-state convection diffusion equation. Now we turn to transient problems. Once again, the study of the simple convection-diffusion problem gives an indication of what might be used in more complicated situations where the analysis is more difficult and sometimes even intractable.

The transient equation is parabolic when the diffusion coefficient is strictly positive. Most finite element methods for solving this type of equations are based upon semidiscretization. The nodal unknowns are considered to be time dependent and the procedure applied for the stationary problem is repeated in this case. This leads to a system of ordinary differential equations (initial value problem) that is usually discretized in time using finite differences. Finite elements may be applied to this semidiscretized system as well, although the power of finite elements is not patent since the time domain is cylindrical in nature and finite differences are well suited for this geometry. Nevertheless, many finite difference schemes may be rederived from the finite element point of view, that is, using a finite element interpolation in time (see, e.g. [ZT]). In this approach towards the formulation of space-time finite element methods, one works directly with the differential equation, not with a weak or variational statement of the problem. Early space-time finite element formulations applied to elastodynamics can be found in References [AS], [Od]. In this case, a variational principle (in the classical sense) does exist, viz. the Hamilton principle, and the extension of finite elements to the time domain is straightforward.

Another different approach has been developed more recently. Trial solutions are allowed to be discontinuous in time, leading to the so called *discontinuous Galerkin method.* Continuity across time slabs is only enforced weakly. This method was originally designed for first-order hyperbolic equations by Lesaint & Raviart [LeR] and was used for the time discretization of the convection-diffusion equation by Johnson, Nävert & Pitkäranta [Jo1], [JNP], [Na], since their analysis [JP] showed that the convergence properties of this formulation are the same as those encountered for the Streamline Diffusion method in the steady-state case, that is, it is of order $h^{m+\frac{1}{2}}$ when a finite element partition of diameter h and a piecewise polynomial interpolation of degree m are used. In the above mentioned references, it is proved that this order of convergence

is maintained for the fully discrete transient convection-diffusion equation. Later, the method has been applied to a wide variety of transient problems, ranging from the Navier-Stokes equations to elastodynamics by Hughes *et al.* (see [HH1], [HH2], [Sh], [SHu] and references therein). The resulting scheme is unconditionally stable, in contrast to the stability requirements found for some time-continuous formulations [Ba]. The improved stability of the discontinuous Galerkin method is not its only advantadge. Unstructured meshes both in space and time can be easily accomodated, thus allowing the tracking of sharp fronts in the space-time domain through the use of mesh adaptivity techniques [Ha]. Furthermore, the fact that the method is based on an integral from of the differential equation allows for simpler convergence proofs and error estimates.

The misbehavior of centered schemes observed when the first spatial derivatives of the differential equation are important may also be present in time. A remedy may be devised for each of the three approaches described above, namely, the finite difference method, the continuous in time finite element interpolation and the discontinuous Galerkin method. If a two-level finite difference scheme is employed, this remedy is quite simple: the use of the backward Euler scheme introduces numerical dissipation in time that precludes the oscillations in a natural way. However, it must be noted that this method is only first order accurate and numerical results are somehow overdamped. Using a continuous in time finite element approach, Yu & Heinrich [YH1], [YH2] developed a Petrov-Galerkin method with the weighting functions depending also on the time discretization. Although results were very accurate, the formulation was found to be too expensive from the computational point of view. Also in the context of finite differences, the temporal discretization may be taken into account [TG]. This point will be discussed in more detail in the next section. Finally, when the discontinuous Galerkin method is used, the way to overcome oscillations is obvious: just do the same as for the stationary problem. Hughes *et al.* used the Galerkin/least-squares method for both space and time [HFH], [HH2], [Sh]. The analogue of this approach using continuous in time interpolations was introduced earlier by Nguyen & Reynen [NR].

The time discretization that will be used in this work is based on a finite difference scheme, the generalized trapezoidal rule described in the next section. In spite of the advantatges of the discontinuous Galerkin method reported earlier, there is still room for simpler schemes as the one that will be used here. First, it is easier to obtain higher accuracy in regions where the solution is smooth (in the space-time domain). Second order accuracy is obtained with the Crank-Nicolson algorithm, which is easy to program and cheaper than the linear in time interpolation that must be used for the discontinuous Galerkin method if a similar accuracy is desired. This linear interpolation doubles the number of nodal unknowns with respect to the Crank-Nicolson algorithm. On the other hand, sometimes it might be useful to use an explicit scheme, perhaps only as a way to reach the steady state.

This last point is the main concern of this chapter. In Section 2.3, the stability of the forward Euler scheme is analyzed both for linear and quadratic elements, assuming that the space discretization has been carried out using the SD method. This stability analysis is done for the one-dimensional equation using the classical von Neumann stability criterion. The extension to multidimensional situations and general meshes discussed in Subsection 2.3.5 is necessarily *ad hoc*, although the time step limitation found in the former case gives an estimate of the critical time step above which the algorithm becomes unstable. A vast amount of numerical experiments support this idea in many situations [MG]. For the present case, several numerical experiments have

also been conducted (Section 2.4) confirming the theoretical predictions obtained here and from which some practical conclusions may be drawn.

2.2 The generalized trapezoidal rule

2.2.1 The continuous problem

The notation already introduced in Chapter 1 will be kept in what follows. Let $[0,T]$ be a given time interval, with $T > 0$. If for a given $t \in [0,T]$ a function $\psi(\mathbf{x},t)$ of the space variable $\mathbf{x}$ and the time t belongs to a space H of functions defined on the domain Ω, the mapping $t \mapsto \psi(\cdot,t)$ from $[0,T]$ to H will also be denoted by $\psi(t)$. The transient convection-diffusion problem that will be considered can be written as follows: Find a function $\phi = \phi(\mathbf{x},t)$ such that

$$\frac{\partial \phi}{\partial t} + \mathbf{u}\cdot\nabla\phi - \nabla\cdot(\mathbf{k}\cdot\nabla\phi) = f, \qquad \text{in } \Omega\times(0,T) \tag{2.1}$$

$$\phi = g, \qquad \text{on } \Gamma_D\times(0,T) \tag{2.2}$$

$$\mathbf{n}\cdot\mathbf{k}\cdot\nabla\phi = r, \qquad \text{on } \Gamma_N\times(0,T) \tag{2.3}$$

$$\phi = \phi^0, \qquad \text{on } \Omega\times\{0\} \tag{2.4}$$

where now the velocity field $\mathbf{u}$, the diffusion tensor $\mathbf{k}$ and the given functions f, g and r may also be time-dependent. For simplicity, $\mathbf{u}$ and $\mathbf{k}$ will be assumed to be constant in time and $\mathbf{u}$ divergence free. The function $\phi^0 = \phi^0(\mathbf{x})$ is a given initial condition. Since we are interested in the semidiscrete formulation of problem (2.1)–(2.4), the spaces of test functions Ψ and of trial solutions Φ that will be needed are

$$\Psi := \{\psi \in H^1(\Omega) \mid \psi = 0 \text{ on } \Gamma_D\} \tag{2.5}$$

$$\Phi := \{\phi : (0,T) \to H^1(\Omega) \mid \phi = g \text{ on } \Gamma_D\times(0,T)\} \tag{2.6}$$

The weak form of problem (2.1)–(2.4) is: Find $\phi \in \Phi$ such that

$$\frac{d}{dt}(\phi(t),\psi) + a(\phi(t),\psi) = l(\psi) \qquad \forall\psi\in\Psi, \quad t\in(0,T) \tag{2.7}$$

$$(\phi(0),\psi) = (\phi^0,\psi) \qquad \forall\psi\in\Psi \tag{2.8}$$

where the bilinear form a and the linear form l are the same as for the steady-state problem, viz.

$$a(\phi,\psi) := \int_\Omega (\psi\mathbf{u}\cdot\nabla\phi + \nabla\psi\cdot\mathbf{k}\cdot\nabla\phi)\,d\Omega \tag{2.9}$$

$$l(\psi) := \int_\Omega \psi f\,d\Omega + \int_{\Gamma_N} \psi r\,d\Gamma \tag{2.10}$$

Existence and uniqueness of solutions for problem (2.7)–(2.8) can be proved under certain regularity assumptions on the data. First observe that (2.7) makes sense for $\phi \in L^2(0,T;H^1(\Omega))$, i.e., the H^1–norm of ϕ is square integrable with respect to t. If the components of $\mathbf{u}$, $\mathbf{k}$ and $\nabla\mathbf{u}$ are bounded (i.e., they belong to $L^\infty(\Omega)$), $f(\cdot,t)$ and

ϕ^0 belong to $L^2(\Omega)$ and $g \in L^2(0,T;H^{\frac{1}{2}}(\Gamma_D))$, then a unique $\phi \in \Phi \cap L^2(0,T;H^1(\Omega))$ exists satisfying (2.7)–(2.8) (see, e.g., [La]).

If we define the operator $A\phi \in \Psi'$, the dual space of Ψ, by $A\phi(\psi) = a(\phi,\psi)$, problem (2.7)–(2.8) may be written as: Find $\phi \in \Phi$ such that

$$\begin{aligned} \frac{d\phi}{dt} + A\phi &= l \qquad \text{in } \Psi', \quad t \in (0,T) \\ \phi(0) &= \phi^0 \qquad \text{in } \Psi' \end{aligned} \tag{2.11}$$

This form of writing problem (2.7)–(2.8) has a direct translation in the semidiscretized equations. The idea is to discretize the operator A using finite elements. After this is done (or before), $\frac{d}{dt}$ is discretized using finite differences.

2.2.2 Discretization in space and time

We consider first the discretization in space. This is done in a manner similar to the stationary problem, from which the notation is kept. The discrete counterparts of the spaces Ψ and Φ are

$$\Psi_h := \{\psi \in \Psi \mid \psi(\cdot)|_{\Omega^e} \in P_m(\Omega^e)\} \subset \Psi \tag{2.12}$$

$$\Phi_h := \{\phi \in \Phi \mid \phi(\cdot,t)|_{\Omega^e} \in P_m(\Omega^e), t \in (0,T)\} \subset \Phi \tag{2.13}$$

The Streamline Diffusion method (SD) will be used for the space discretization of problem (2.7)–(2.8). This leads to the following system of ordinary differential equations: Find $\phi_h \in \Phi_h$ such that

$$\frac{d}{dt}(\phi_h(t),\psi_h)_{sd} + a_{sd}(\phi_h(t),\psi_h) = l_{sd}(\psi_h) \qquad \forall \psi_h \in \Psi_h, \quad t \in (0,T) \tag{2.14}$$

$$(\phi_h(0),\psi_h) = (\phi^0,\psi_h) \qquad \forall \psi_h \in \Psi_h \tag{2.15}$$

Here, the bilinear form a_{sd} and the linear l_{sd} are those given by Eqns. (1.39) and (1.40):

$$a_{sd}(\phi_h,\psi_h) := a(\phi_h,\psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} \zeta_h \left[\mathbf{u}\cdot\nabla\phi_h - \nabla\cdot(\mathbf{k}\cdot\nabla\phi_h)\right] d\Omega \tag{2.16}$$

$$l_{sd}(\psi_h) := l(\psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} \zeta_h \, f \, d\Omega \tag{2.17}$$

and the modified inner product $(\cdot,\cdot)_{sd}$ is

$$(\phi_h,\psi_h)_{sd} := (\phi_h,\psi_h) + \sum_{e=1}^{N_{el}} \int_{\Omega^e} \zeta_h \, \phi_h \, d\Omega \tag{2.18}$$

It is clear that this method is consistent, in the sense that whenever ϕ is a solution of (2.1)–(2.4), then ϕ will also satify Eqn. (2.14) and the initial condition (2.15).

The perturbation ζ_h of the test function ψ_h will be

$$\zeta_h = \tau^e \mathbf{u}^e \cdot \nabla\psi_h \tag{2.19}$$

for each element. The meaning of the terms appearing in this expression has been already explained in Chapter 1.

The time discretization of problem (2.14)–(2.15) will be carried out using the generalized trapezoidal rule (see any book on finite difference methods, for example [MG], [MM], [RM]). In order to simplify the notation, subscript h will be omitted in what follows. Let $\theta \in [0,1]$ be given and let Δt be the time step size of a partition of the interval $[0,T]$. This time step may be either constant or variable depending on the time slab $[t^n, t^{n+1}]$, with $n = 0,1,2,...,N-1$ and $t^N = T$. When all the terms appearing in the bilinear form a_{sd} are evaluated at time $t = t^n$, we will write a^n_{sd} and similarly for the linear form l_{sd}. We will also write ϕ^n for the approximation to $\phi(t^n)$. The generalized trapezoidal rule for problem (2.14)–(2.15) reads:

For $n = 0,1,2,...,N-1$, given ϕ^n find ϕ^{n+1} such that

$$\begin{aligned}\frac{1}{\Delta t}(\phi^{n+1} - \phi^n, \psi)_{sd} + \theta\, a^{n+1}_{sd}(\phi,\psi) + (1-\theta)\, a^n_{sd}(\phi,\psi) \\ = \theta\, l^{n+1}_{sd}(\psi) + (1-\theta)\, l^n_{sd}(\psi)\end{aligned} \tag{2.20}$$

This equation (2.20) may be rewritten as

$$\begin{aligned}(\phi^{n+1},\psi)_{sd} + \Delta t\ \theta\ a^{n+1}_{sd}(\phi,\psi) = \Delta t\ \theta\ l^{n+1}_{sd}(\psi) + \Delta t\ (1-\theta)\ l^n_{sd}(\psi) \\ - \Delta t\ (1-\theta)\ a^n_{sd}(\phi,\psi) + (\phi^n,\psi)_{sd}\end{aligned} \tag{2.21}$$

For $n = 0$, Eqn. (2.15) has to be solved. Observe that if the given initial condition $\phi^0(\mathbf{x})$ belongs to the discrete finite element space, ϕ^n will simply be $\phi^0(\mathbf{x})$ for $n = 0$. For simplicity, we will assume that this is the case. In fact, the usual methodology in practice is to interpolate $\phi^0(\mathbf{x})$ in the discrete space and take this interpolation as the effective initial condition.

Once a basis in chosen for the finite element space (i.e., shape functions), system (2.14)–(2.15) can be written in the standard matrix form

$$\begin{aligned}\mathbf{M}\dot{\boldsymbol{\phi}} + \mathbf{K}\boldsymbol{\phi} = \mathbf{f} \quad t \in (0,T) \\ \boldsymbol{\phi}(0) = \boldsymbol{\phi}^0\end{aligned} \tag{2.22}$$

where the dot denotes the temporal derivative and $\boldsymbol{\phi}$ is the vector of nodal unknowns of the function ϕ. The components of the matrices $\mathbf{M}$ and $\mathbf{K}$ and the vector $\mathbf{f}$ are the linear forms $(\cdot,\cdot)_{sd}$, a_{sd} and l_{sd} applied to the basis (shape) functions, respectively. For the vector $\mathbf{f}$, the Dirichlet boundary conditions (2.2) have to be taken into account. As usual, all this arrays are constructed for each element e and then they are assembled (see, e.g. [Hu], [ZT]). System (2.22) is the discrete analogue of the abstract evolution problem (2.11). The matrix form of Eqn. (2.21) is

$$\begin{aligned}(\mathbf{M} + \Delta t\ \theta\ \mathbf{K})\boldsymbol{\phi}^{n+1} = \Delta t\ \theta\ \mathbf{f}^{n+1} + \Delta t\ (1-\theta)\ \mathbf{f}^n \\ - \Delta t\ (1-\theta)\ \mathbf{K}\boldsymbol{\phi}^n + \mathbf{M}\boldsymbol{\phi}^n\end{aligned} \tag{2.23}$$

The general stability and convergence properties of algorithm (2.20) are well known and can be found in any standard text book, both on finite difference and on finite element methods. For example, this algorithm is described in the context of finite elements and for the diffusion equation (without convection) in References [Hu], [Jo2], [RT], [SF] and for the convection-diffusion equation (and other problems) in [CO],

[CSS], [ZT]. Error estimates for the diffusion problem using the Galerkin formulation are given in [Jo2], [RT], [SF]. More information on the Galerkin method for parabolic problems can be found in [LuR] and [Th].

It is well known that algorithm (2.20) is in general first order accurate in time for any $\theta \in [0,1]$. Second order accuracy is obtained only if $\theta = \frac{1}{2}$, the Crank-Nicolson scheme. The algorithm is unconditionally stable (and hence convergent) for $\theta \geq \frac{1}{2}$. For lower values of θ, there is a time step limitation for stability. Besides $\theta = \frac{1}{2}$, the other two interesting cases are $\theta = 1$ (backward Euler) and $\theta = 0$ (forward Euler). The former is unconditionally stable and has an important numerical damping. As it was already said in the introduction to this chapter, this is an important attribute when the solution develops sharp gradients in the space-time domain, since they will be authomatically smoothed and oscillations will be precluded. The use of $\theta = 1$ is also useful for the first time steps. The reason is that it can be shown that the Fourier series expansion of the solution of the continuous problem (2.1)–(2.4) has rapidly oscillating harmonics, i.e., with high frequency, that are quickly damped when time increases. Many of these harmonics cannot be reproduced by the time discretization, no matter how small Δt is. So, the best one can do is to damp them out by using $\theta = 1$, at least for the first time steps. See Reference [Jo2] for further discussion.

The case $\theta = 0$ is of interest since it yields an explicit scheme, in the sense that if the *mass matrix* M appearing in (2.23) is approximated by a diagonal matrix, then the solution of this equation is trivial and no solver for algebraic systems is needed. The stability requirements of this scheme is the object of the next section.

The description of the fully discretized problem is now complete. It only remains to discuss the choice of the intrinsic time τ^e in Eqn. (2.19). A possible choice would be the same as for the stationary problem, i.e.

$$\tau^e = \frac{\alpha^e h^e}{2|\mathbf{u}^e|} \tag{2.24}$$

Following Tezduyar & Ganjoo [TG], suppose that instead of τ^e we take

$$\tau_t^e = C_{2\tau}\tau^e$$

where $C_{2\tau}$ depends on the *Courant number* defined by

$$c := \frac{|\mathbf{u}|\Delta t}{h} \tag{2.25}$$

This dimensionless number can also be defined for each element. For one-dimensional pure convection problems, Tezduyar & Ganjoo found that the best time accuracy was obtained selecting $C_{2\tau}$ as

$$C_{2\tau} = \frac{2}{\sqrt{15}} + \left(1 - \frac{2}{\sqrt{15}}\right) c$$

when linear elements are used. In fact, when $C_{2\tau} = \frac{2}{\sqrt{15}}$, fourth order phase accuracy is obtained for the forward Euler scheme [RG]. This choice was used for general transient problems in [BH] and [HT1]. However, we have found from several numerical experiments that when diffusion is present or quadratic elements are used, $C_{2\tau} = 1$ is optimum when errors and mesured in the discrete L^2 norm [Co1]. So, in what follows

the choice of τ^c and therefore the perturbation of the test functions (2.19) that defines the SD method will be taken the same for transient problems as for the steady-state case.

2.3 Stability analysis of the forward Euler scheme

2.3.1 General considerations

Let A be an $n \times n$ matrix and $x = x(t)$ an n–dimensional vector function of the time variable t. When the numerical solution of a system of linear ordinary differential equations $\dot{x} = Ax$ is attempted using a two-level finite difference scheme, the final algebraic system to be solved at each time step will have the form $x^{n+1} = Ex^n$, the superscript denoting the time step level and E being a certain $n \times n$ matrix. Since the error z will also satisfy the difference scheme, we will have that $z^{n+1} = Ez^n$. Stability requires that $\|z^n\|$ must be bounded for any n. Since

$$\|z^{n+1}\| = \|Ez^n\| = \|E^{n+1}z^0\| \leq \|E\|^{n+1}\|z^0\|$$

stability will hold whenever

$$\|E\| \leq 1 \tag{2.26}$$

In fact, one can show that only

$$\|E\| \leq 1 + O(\Delta t)$$

is necessary [MG], [RM], where Δt is the time step size. However, in practice one neglects the term $O(\Delta t)$ and what is really checked is condition (2.26).

The symbol $\|\cdot\|$ used above denotes *any* vector norm when it is applied to a vector and *the associated* matrix norm when it is applied to a matrix. When the differential system comes from the space discretization of a partial differential equation, matrix A and hence E will depend on the space discretization size h, as well as on the boundary conditions. Checking (2.26) for any h, Δt, space geometry and boundary conditions is in general intractable. Alternatively, the condition

$$\varrho(E) \leq 1 \tag{2.27}$$

is verified. Here, $\varrho(E)$ denotes the spectral radius of E (i.e., the maximum absolute value of the eigenvalues of E). It is known that, for any matrix norm $\|\cdot\|$,

$$\varrho(E) \leq \|E\| \tag{2.28}$$

Thus, condition (2.27) is obviously necessary, but *not sufficient* for stability. It can be proved (see, e.g. [MG]) that the equality holds when E is symmetric or similar to a symmetric matrix, that is, there exists a non-singular matrix P such that $P^{-1}EP$ is symmetric. When the differential system results from the space discretization of the convection-diffusion equation, matrix E is not symmetric (the convection operator is skew-symmetric for divergence free velocity fields). Historically, the use of (2.27) has caused confusion since it leads to misleading results. See Reference [SG] and the discussion originated in References [HGG] and [Mo].

Another way to study the stability of finite difference schemes is the von Neumann method, based on a Fourier mode analysis of the error z. It is assumed that this error can be expanded in Fourier series. This requires periodicity of the problem in the space domain. When this condition does not hold, the von Neumann criterion only gives necessary conditions for stability. Nevertheless, experience and also some heuristic considerations show that the necessary condition for stability obtained using the Fourier analysis is much more precise and useful than the one based on the spectral radius (2.27). An interesting discussion of this fact can be found in References [HGG] and [Mo]. In the former, several simple cases with different boundary conditions have been studied by checking the stability condition (2.26) and showing the effectiveness of the von Neumann criterion. It is also argued that the reason why this method works so well in situations where *a priori* it is not sufficient for stability is that instabilities are generated far from the boundary. Thus, boundary conditions do not play a decisive role on the stable or the unstable behavior of the scheme.

Having this considerations in mind, in this section we will consider the following initial-boundary-value problem: Find a scalar function $\phi = \phi(x,t)$ satisfying the differential equation

$$\frac{\partial \phi}{\partial t} + u\frac{\partial \phi}{\partial x} - k\frac{\partial^2 \phi}{\partial x^2} = 0, \quad 0 < x < \ell, \quad t > 0 \tag{2.29}$$

as well as the initial condition

$$\phi(x,0) = \phi^0(x), \quad 0 < x < \ell \tag{2.30}$$

and the periodic boundary conditions

$$\phi(0,t) = \phi(\ell,t) \quad \text{and} \quad \frac{\partial \phi}{\partial x}(0,t) = \frac{\partial \phi}{\partial x}(\ell,t) \tag{2.31}$$

for $t > 0$. The constant diffusion k is positive and, without loss of generality, we will assume the constant velocity u to be also positive. For the sake of clarity, source terms have also been omitted. The stability of the forward Euler scheme in time and the Streamline Diffusion method in space will be analyzed for both linear and quadratic finite elements using mainly the von Neumann method. This analysis is obviously restricted to this simple one-dimensional problem using a uniform finite element partition. Nevertheless, it gives a critical time step that provides an estimate of what might be used in general situations. The extension to multidimensional problems will be briefly discussed in Subsection 2.3.5.

Let $N(x)$ denote a generic shape function. It has been seen in Chapter 1 that the SD method for problem (2.29)–(2.31) consists in taking the weighting functions as $N(x)$ plus a perturbation $P(x)$ of the form

$$P(x) = \frac{\alpha h}{2}\frac{dN}{dx}$$

where α is the upwind function, depending on the Péclet number

$$\gamma := \frac{uh}{2k} \tag{2.32}$$

In order to simplify the exposition, here we will consider that the asymptotic approximation to the optimal upwind functions is used. Box 2.1 summarizes their expressions for linear and quadratic elements, both using the standard shape functions

(canonical basis) and the hierarchic approach. Recall that for quadratic elements a different upwind function is needed for the extreme nodes (α) and for the central nodes (β).

Box 2.1 Upwind functions

Element	Extreme node	Central node
Linear	$\alpha = \min(\frac{\gamma}{3}, 1)$	
Quadratic standard	$\alpha = \min(\frac{\gamma}{12}, 1)$	$\beta = \min(\frac{\gamma}{12}, \frac{1}{2})$
Quadratic hierarchic	$\alpha = \min(\frac{\gamma}{15}, \frac{1}{3})$	$\beta = \min(\frac{\gamma}{12}, \frac{1}{2})$

Let $0 = x_0 < x_1 < ... < x_{N_{el}} = \ell$ be a uniform partition of diameter h of the interval $[0, \ell]$. Element e, $e = 1, 2, ..., N_{el}$, will be defined by the nodes placed at the abcissa x_{e-1} and x_e. If quadratic elements are used, the central node will be placed at $x_{e+\frac{1}{2}} := \frac{1}{2}(x_e + x_{e+1})$. If $\boldsymbol{\phi} = \boldsymbol{\phi}(t)$ is the vector containing the nodal unknowns of $\phi(x,t)$, the application of the SD method to (2.29)–(2.31) will lead to an initial value problem of the form (2.22):

$$\begin{aligned} \mathbf{M}\dot{\boldsymbol{\phi}} + \mathbf{K}\boldsymbol{\phi} = 0 \quad t > 0 \\ \boldsymbol{\phi}(0) = \boldsymbol{\phi}^0 \end{aligned} \tag{2.33}$$

If the forward Euler scheme is now employed, once $\boldsymbol{\phi}$ is known at time level t^n, $\boldsymbol{\phi}^{n+1}$ will be found by solving

$$\mathbf{M}\boldsymbol{\phi}^{n+1} = \mathbf{M}\boldsymbol{\phi}^n - \Delta t \mathbf{K}\boldsymbol{\phi}^n \tag{2.34}$$

This scheme is only useful when the matrix $\mathbf{M}$ is approximated by a diagonal matrix $\mathbf{M}_d$. In this case, $\boldsymbol{\phi}^{n+1}$ can be obtained explicitly from $\boldsymbol{\phi}^n$:

$$\boldsymbol{\phi}^{n+1} = \boldsymbol{\phi}^n - \Delta t \mathbf{M}_d^{-1}\mathbf{K}\boldsymbol{\phi}^n \tag{2.35}$$

Observe that $E = \mathbf{I} - \Delta t \mathbf{M}_d^{-1}\mathbf{K}$, with the notation introduced earlier. Clearly, matrix $\mathbf{M}_d$ must be nonsingular and positive-definite, since otherwise the analogue of (2.33) obtained by replacing $\mathbf{M}$ by $\mathbf{M}_d$ would be unstable. Thus, the diagonal entries of $\mathbf{M}_d$ must be positive. This matrix $\mathbf{M}_d$ can be easily obtained by using the classical row-sum lumping technique or through nodal integration when the standard shape functions are used [Hu], [ZT], i.e., when the shape function $N_i(x)$ associated to a node i satisfies $N_i(x_j) = \delta_{ij}$, the Kronecker symbol, when applied to a node of abcissa x_j. In this case $\sum_{e=1}^{N_{el}} N_e(x) \equiv 1$. However, the situation is more delicate when the hierarchic approach is used. We will come back to this fact later.

The purpose of what follows is to analyse the stability and accuracy of (2.35). First, linear elements will be considered. The stability condition in this case is well

known (see, e.g. [HGG], [Mo]), but this will serve us as an introduction to the somehow more complicated situation encountered when quadratic elements are treated. Moreover, our interest here is to discuss what happens when the SD method is used.

The use of quadratic elements deserves an explanation. The best space accuracy one can hope for is an error of order $O(h^3)$ in the L^2 norm. On the other hand, the forward Euler scheme is only first order accurate in time, i.e., errors are of order $O(\Delta t)$. Thus, it is apparent that space and time errors will not be properly compensated unless Δt be very small. Nevertheless, the Euler scheme may be useful to solve a steady-state problem. One may think that the time steps are in fact iteration steps of a relaxation procedure.

2.3.2 Linear elements

When linear elements are used, matrices $\mathbf{M}$ and $\mathbf{K}$ appearing in Eqn. (2.34) will be obtained by assembling the element matrices

$$\mathbf{M}^e = \frac{h}{2}\begin{pmatrix} \frac{2}{3}-\frac{1}{2}\alpha & \frac{1}{3}-\frac{1}{2}\alpha \\ \frac{1}{3}+\frac{1}{2}\alpha & \frac{2}{3}+\frac{1}{2}\alpha \end{pmatrix}$$

$$\mathbf{K}^e = \frac{2k}{h}\begin{pmatrix} \frac{1}{2}(1+\alpha\gamma-\gamma) & -\frac{1}{2}(1+\alpha\gamma-\gamma) \\ -\frac{1}{2}(1+\alpha\gamma+\gamma) & \frac{1}{2}(1+\alpha\gamma+\gamma) \end{pmatrix}$$

Matrix $\mathbf{M}^e$ may be diagonalized by using the row-sum lumping technique (see [HT2] for different choices of $\mathbf{M}^e$ arising from numerical integration). Once this is done and the element matrices are assembled, a typical algorithmic equation for an internal node m resulting from Eqn. (2.35) is

$$\phi_m^{n+1} = \phi_m^n + \Delta t\left[\left(\frac{k}{h^2}+\frac{\alpha u}{2h}\right)\left(\phi_{m+1}^n - 2\phi_m^n + \phi_{m-1}^n\right) - \frac{u}{2h}\left(\phi_{m+1}^n - \phi_{m-1}^n\right)\right] \quad (2.36)$$

Stability and accuracy

The analytical solution of problem (2.29)–(2.31) may be expanded in Fourier series, each mode having the form

$$\hat{\phi}(x,t) = a\, e^{-(\xi+i\omega)t}\, e^{iKx} \quad (2.37)$$

where a is the amplitude of the mode, K the wave number, $\xi := kK^2$ the damping, $\omega := Ku$ the frequency and $i := \sqrt{-1}$. Let

$$\hat{\phi}_m^n = a\, e^{-(\xi^h+i\omega^h)n\Delta t}\, e^{iKmh} \quad (2.38)$$

be the harmonic corresponding to (2.37) evaluated at $(x,t) = (x_m, t^n) = (mh, n\Delta t)$ for the discrete problem. Here, ξ^h is the algorithmic damping and ω^h the algorithmic frequency. Although only discrete values of K can be reproduced by the discretization, we will consider as usual that K is any real number.

The amplification factor arising from scheme (2.36) is

$$\begin{aligned} A^h := \frac{\hat{\phi}_m^{n+1}}{\hat{\phi}_m^n} &= 1 + \left(\frac{2k\Delta t}{h^2}+\alpha\frac{u\Delta t}{h}\right)(\cos Kh - 1) - i\frac{u\Delta t}{h}\sin Kh \\ &= 1 + \left(\frac{c}{\gamma}+\alpha c\right)(\cos Kh - 1) - i\, c\, \sin Kh \end{aligned} \quad (2.39)$$

where $c := u\Delta t/h$ is the Courant number.

The von Neumann stability criterion requires that $|A^h| \leq 1$ for any K. Define $z := \cos Kh$, $z \in [-1, 1]$. The stability limit will be found by examining under which conditions the function

$$|A^h|^2(z) = \left[1 + \left(\frac{c}{\gamma} + \alpha c\right)(z-1)\right]^2 + c^2\left(1 - z^2\right)$$

is ≤ 1 for $-1 \leq z \leq 1$. Using the abbreviation

$$b := \frac{c}{\gamma} + \alpha c$$

inequality $|A^h|^2(z) \leq 1$ reduces to

$$2b + b^2(z-1) - c^2(z+1) \geq 0 \tag{2.40}$$

for $-1 \leq z \leq 1$. Condition (2.40) holds if, and only if,

$$b \leq 1 \qquad \text{and} \qquad c^2 \leq b \tag{2.41}$$

that may be equivalently written as

$$c \leq \min\left(\frac{\gamma}{1+\alpha\gamma}, \frac{1}{\gamma} + \alpha\right) \tag{2.42}$$

It is easy to see that

$$\frac{\gamma}{1+\alpha\gamma} \leq \frac{1}{\gamma} + \alpha \tag{2.43}$$

whenever α exceeds the critical value

$$\alpha_c := 1 - \frac{1}{\gamma} \tag{2.44}$$

From the expression of the upwind function for linear elements given in Box 2.1 it follows that $\alpha \geq \alpha_c$. Therefore, inequality (2.43) holds and (2.42) is simply

$$c \leq \frac{\gamma}{1+\alpha\gamma} \tag{2.45}$$

This is the sought stability condition.

Remarks 2.1

(1) Observe that if $\alpha = 0$ (Galerkin method) the algorithm becomes unconditionally unstable when $\gamma \to \infty$, since in that case condition (2.42) requires $c = 0$.

(2) Recall that the condition $\alpha \geq \alpha_c$, with α_c given by (2.44), had already been found in the last chapter as the condition under which no oscillations appear in the numerical solution of the stationary equation.

(3) From the expression of α, it follows that (2.45) reduces to

$$c \leq 1 \qquad \text{in the advective limit } (\gamma \to \infty) \tag{2.46}$$

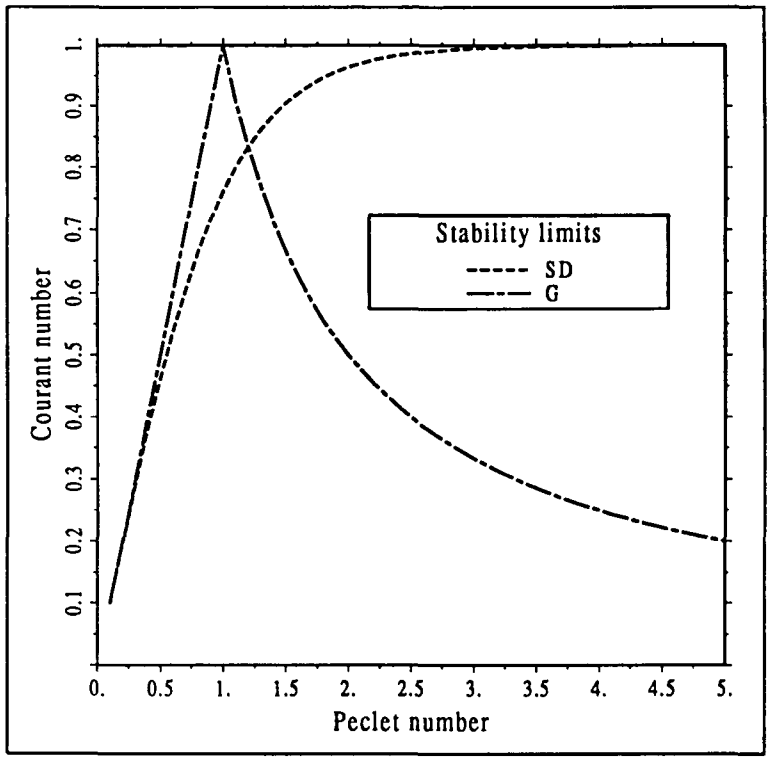

Figure 2.1 Stability limits for the convection-diffusion equation using linear elements. G: Galerkin method, SD: Streamline Diffusion method.

and

$$\Delta t \leq \frac{h^2}{2k} \qquad \text{in the diffusive limit } (u \to 0) \tag{2.47}$$

Inequality (2.46) is the CFL condition. It is well known that it is the best one can hope for. □

The stability region dictated by (2.42) when $\alpha = 0$ (Galerkin mehod) and when $\alpha = \min(\frac{\gamma}{3}, 1)$ (SD method) has been plotted in Figure 2.1

In order to determine the formal accuracy of the algorithm, both the exact and the numerical amplification factors, A and A^h, will be expanded in powers of Δt and h. Let $\hat{\phi}^n(x) = \hat{\phi}(x, t^n)$, with $\hat{\phi}$ given by (2.37). The analytical amplification factor is

$$\begin{aligned} A :&= \frac{\hat{\phi}^{n+1}(x_m)}{\hat{\phi}^n(x_m)} = e^{-(kK^2+iKu)\Delta t} \\ &= 1 - (kK^2 + iKu)\Delta t + \frac{1}{2}(k^2K^4 - K^2u^2 + 2ikK^3u)\Delta t^2 + O(\Delta t^3) \end{aligned} \tag{2.48}$$

whereas the algorithmic amplification factor given by (2.39) satisfies

$$A^h = 1 - (kK^2 + iKu)\Delta t - \frac{1}{2}\alpha u K^2 h \Delta t + O(h^2 \Delta t) \tag{2.49}$$

Since $A = e^{-(\xi+i\omega)\Delta t}$ and $A^h = e^{-(\xi^h+i\omega^h)\Delta t}$, the damping error and the frequency error will be one order of Δt less than the amplification factor error. Comparing expressions (2.48) and (2.49) we see that

- $A - A^h = O(\alpha h \Delta t)$ and hence $\xi - \xi^h = O(\alpha h)$ and $\omega - \omega^h = O(\alpha h)$. This formal estimate is clearly pessimistic, since for a properly chosen upwind function α we know that the accuracy is much higher than with $\alpha = 0$. This fact has also been observed in [SHu], where predictor-corrector algorithms for the Galerkin/least-squares method are studied.
- If $\alpha = 0$ and $h^2 = C\Delta t$, C being a constant, we have that $A - A^h = O(\Delta t^2)$ and thus $\xi - \xi^h = O(\Delta t)$ and $\omega - \omega^h = O(\Delta t)$. The algorithm is formally first order accurate in time, although it suffers from very important spatial oscillations when the Péclet number γ is high.

Remarks on the Taylor-Galerkin method

The Taylor-Galerkin method introduced by Donea [Do1], [Do2], is nothing but the finite element counterpart of the Lax-Wendroff method in the finite difference context. Although its basic motivation is quite different, the final formulation is very similar to the SD method for the problem considered here and using the forward Euler scheme in time. Basically, only the definition of the upwind function α differs. Whereas in the SD method it is designed considering the space accuracy (the error analysis dictates its asymptotical behavior) the time discretization is the starting point of the Taylor-Galerkin approach. The final algorithm results in the choice $\alpha = c$, the Courant number, for the upwind function.

Here, our aim is to point out that the previous discussion is also valid for the Taylor-Galerkin method if α is set equal to c. In particular, the results obtained in References [Do1], [Do2] and [Pe] may be easily recovered.

Recall that, for stability, conditions $b \leq 1$ and $c^2 \leq b$ are needed (cf. Eqn. (2.41)) and we have seen that if $\alpha \geq \alpha_c$, with α_c given by (2.44), the former is more restrictive than the latter. However, if

$$\alpha \geq c - \frac{1}{\gamma}$$

we have that

$$b = \frac{c}{\gamma} + \alpha c \geq \frac{c}{\gamma} + \left(c - \frac{1}{\gamma}\right) c = c^2$$

and hence $c^2 \leq b$ *always*, not only when $b \leq 1$.

The stability limit of the Taylor-Galerkin method is also (2.45). Setting $\alpha = c$ yields

$$c \leq \sqrt{\frac{1}{4\gamma^2} + 1} - \frac{1}{2\gamma} \tag{2.50}$$

It can be easily seen that this limit is slightly less restrictive than (2.45) with $\alpha = \min(\frac{\gamma}{3}, 1)$.

Concerning the accuracy, we will always have that $\xi - \xi^h = O(\Delta t)$ and $\omega - \omega^h = O(\Delta t)$. Moreover, for the particular case $k = 0$ (pure convection) and for $h^2 = C\Delta t$, with C a constant, second order accuracy is obtained, i.e., $\xi - \xi^h = O(\Delta t^2)$ and $\omega - \omega^h = O(\Delta t^2)$. This follows from the comparison of (2.48) and (2.49) with $\alpha = c$. The reader is referred to the above quoted references for further discussion.

Algorithmic damping ratio (ADR) and frequency ratio (AFR)

From the practical point of view, it is important to have a global feeling of how the algorithm behaves for the whole range of space sizes h. The algorithmic damping

ratio and the algorithmic frequency ratio are dimensionless numbers defined by

$$ADR := \frac{\xi^h}{\xi} \qquad \text{and} \qquad AFR := \frac{\omega^h}{\omega} \tag{2.51}$$

respectively. The ADR gives a mesure of the dampig error and the AFR of the phase error. These quantities are functions of the dimensionless wave number $\bar{K} := Kh$. The numerical method can only reproduce values $0 \leq \bar{K} \leq \pi$, the upper bound corresponding to two elements per wave length. For accuracy, it is commonly argued that at least ten elements per wave length are needed [BH], [SHu], corresponding to $\bar{K} \approx 0.6$.

In Reference [TG], the ADR and the AFR are plotted only for the pure convection problem, whereas the convection-diffusion case is considered in [SHu], although not for the forward Euler scheme. We will do this here, considering also the relative importance of convection in the problem.

Given a diffusion k and a velocity u, let us write the Péclet number γ as

$$\gamma = \frac{u}{2kK} Kh =: \gamma_0 \bar{K} \tag{2.52}$$

The coefficient γ_0 is proportional to the global Péclet number. We will call it *convection factor*. Low values of γ_0 will indicate that diffusion dominates, whereas convection will be dominant for high values of γ_0.

On the other hand, the Courant number will be taken as

$$c = c_0 \frac{\gamma}{1 + \alpha\gamma} \tag{2.53}$$

For stability, $c_0 \leq 1$. This value c_0 will be called *security factor*.

Having introduced γ_0 and c_0, the analytical amplification factor and the algorithmic amplification factor will be

$$A = \exp\left(-\frac{c}{2}\frac{\bar{K}}{\gamma_0} - ic\bar{K}\right)$$
$$A^h = 1 + c_0(\cos\bar{K} - 1) - ic\sin\bar{K}$$

with c given by (2.53), $\gamma = \gamma_0\bar{K}$ and $\alpha = \min(\frac{\gamma}{3}, 1)$. The factors A and A^h will be a function of the convection factor γ_0 and the security factor c_0. Hence

$$ADR = \left(\log|A^h|\right)\left(\log|A|\right)^{-1} = ADR(c_0, \gamma_0, \bar{K})$$
$$AFR = \left(\arg A^h\right)\left(\arg A\right)^{-1} = AFR(c_0, \gamma_0, \bar{K})$$

We have considered the cases $\gamma_0 = 0.1$, 1 and 10 as representative of problems with different importance of convection. For each case, the ADR and the AFR have been plotted for $c_0 = 0.25$, 0.5, 0.75 and 0.95. Results are shown in Figure 2.2. Since the sign of ω^h only affects the imaginary part of A^h, the absolute value of AFR has been plotted.

The conclusions that may be drawn from these plots can be predicted considering the mode $\bar{K} = \pi$. In this case

$$|A| = \exp\left(-\frac{\pi}{2\gamma_0}\frac{c_0\gamma_0\pi}{1 + \gamma_0\pi}\right) \qquad (\text{for } \alpha = 1)$$
$$|A^h| = |1 - 2c_0|$$

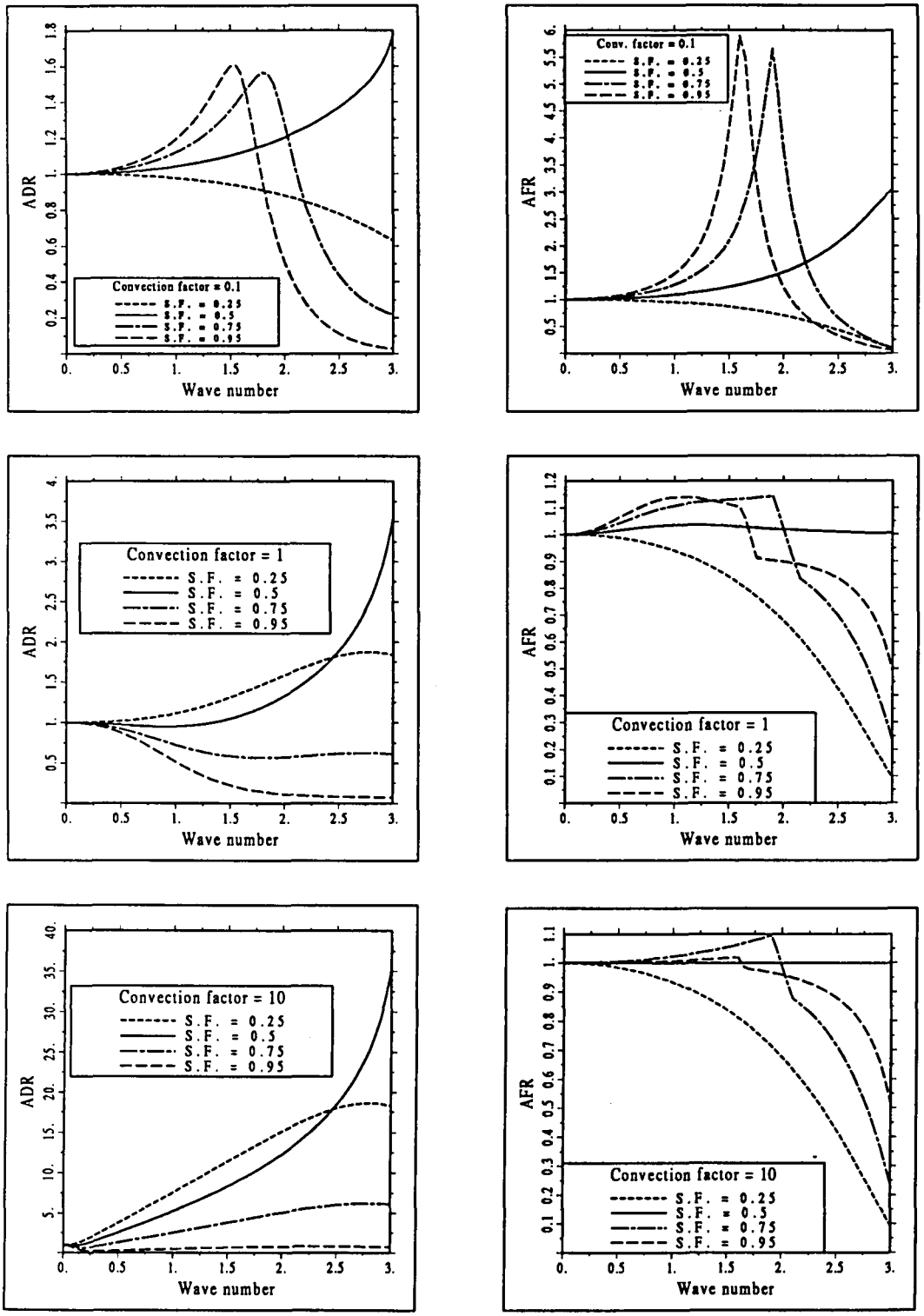

Figure 2.2 Algorithmic damping ratio (ADR) and algorithmic frequency ratio (AFR) for the SD method using linear elements and for different values of the security factor (S.F.) c_0 and the convection factor γ_0.

We see that when $\gamma_0 \to \infty$, then $|A| \to 1$ and hence $\xi \to 0$. In order to obtain values of $ADR \geq 1$ (for precluding oscillations), security factors close to 1 (in which case also ξ^h is small) may be used. But if γ_0 is fixed ($< \infty$) and $c_0 = 1$, the mode $\bar{K} = \pi$ is not damped in the numerical solution and $ADR = 0$. Oscillations or unphysical behavior may be expected if the analytical solution exhibits this mode.

Since for $c_0 = 0.5$ it is $|A^h| = 0$, for $\gamma_0 < \infty$ we will have that $ADR \to \infty$ as $\bar{K} \to \pi$. From Figure 2.2 it is seen that $c_0 = 0.5$ gives an important damping for the whole range of $\bar{K}$ and the different values of γ_0. We conclude that this choice of c_0 is 'safe' and especially useful if only the steady-state solution is of interest. Moreover, the AFR is close to 1, showing that the phase (or dispersion) errors will be small. This fact will be confirmed by the numerical experiments presented later.

2.3.3 Quadratic elements I : canonical basis

Suppose now that the spatial discretization of the problem is performed using quadratic elements. Here we will consider that the standard shape functions are used (see Figure 1.1). The element matrices $\mathbf{M}^e$ and $\mathbf{K}^e$ are

$$\mathbf{M}^e = \frac{h}{2}\begin{pmatrix} \frac{4}{15}-\frac{\alpha}{2} & \frac{2}{15}-\frac{2\alpha}{3} & -\frac{1}{15}+\frac{\alpha}{6} \\ \frac{2}{15}+\frac{2\beta}{3} & \frac{16}{15} & \frac{2}{15}-\frac{2\beta}{3} \\ -\frac{1}{15}-\frac{\alpha}{6} & \frac{2}{15}+\frac{2\alpha}{3} & \frac{4}{15}+\frac{\alpha}{2} \end{pmatrix}$$

$$\mathbf{K}^e = \frac{2k}{h}\begin{pmatrix} \frac{7}{6}+\alpha+\gamma\left(-\frac{1}{2}+\frac{7\alpha}{6}\right) & -\frac{4}{3}-2\alpha+\gamma\left(\frac{2}{3}-\frac{4\alpha}{3}\right) & \frac{1}{6}+\alpha+\gamma\left(-\frac{1}{6}+\frac{\alpha}{6}\right) \\ -\frac{4}{3}+\gamma\left(-\frac{2}{3}-\frac{4\beta}{3}\right) & \frac{8}{3}+\gamma\frac{8\beta}{3} & -\frac{4}{3}+\gamma\left(\frac{2}{3}-\frac{4\beta}{3}\right) \\ \frac{1}{6}-\alpha+\gamma\left(\frac{1}{6}+\frac{\alpha}{6}\right) & -\frac{4}{3}+2\alpha+\gamma\left(-\frac{2}{3}-\frac{4\alpha}{3}\right) & \frac{7}{6}-\alpha+\gamma\left(\frac{1}{2}+\frac{7\alpha}{6}\right) \end{pmatrix}$$

Once again, $\mathbf{M}^e$ may be diagonalized by using the row-sum lumping technique. Observe that when this is done the effect of the SD weighting disappears in the assembled matrix $\mathbf{M}$.

The situation is now more involved than for linear elements. Two different typical algorithmic equations will be found for the internal nodes of the finite element partition, one for the extreme nodes and another one for the central nodes. The stability of each set of equations has to be studied separately, as well as its accuracy. These equations are

- Central nodes:

$$\phi^{n+1}_{m+\frac{1}{2}} = \left[c(1+2\beta)+2\frac{c}{\gamma}\right]\phi^n_m + \left[1-4\left(\frac{c}{\gamma}+\beta c\right)\right]\phi^n_{m+\frac{1}{2}} + \left[-c(1-2\beta)+2\frac{c}{\gamma}\right]\phi^n_{m+1} \tag{2.54}$$

- Extreme nodes:

$$\phi^{n+1}_m = \left[-3\frac{c}{\gamma}\left(\frac{1}{6}-\alpha\right)-3c\left(\frac{1}{6}+\frac{1}{6}\alpha\right)\right]\phi^n_{m-1} + \left[3\frac{c}{\gamma}\left(\frac{4}{3}-2\alpha\right)+3c\left(\frac{2}{3}+\frac{4}{3}\alpha\right)\right]\phi^n_{m-\frac{1}{2}}$$

$$+\left[1-7\left(\frac{c}{\gamma}+\alpha c\right)\right]\phi_m^n$$
$$+\left[3\frac{c}{\gamma}\left(\frac{4}{3}+2\alpha\right)-3c\left(\frac{2}{3}-\frac{4}{3}\alpha\right)\right]\phi_{m+\frac{1}{2}}^n$$
$$+\left[-3\frac{c}{\gamma}\left(\frac{1}{6}+\alpha\right)+3c\left(\frac{1}{6}-\frac{1}{6}\alpha\right)\right]\phi_{m+1}^n \tag{2.55}$$

The Courant number used in these expressions is also $u\Delta t/h$, where h is the total length of the elements.

Stability and accuracy

Define $\hat{\phi}(x,t)$ and $\hat{\phi}_{m+q}^n$ as before, with $q=0$ or $q=\frac{1}{2}$. We first consider the stability and accuracy for Eqn. (2.54) (central nodes), i.e., $q=\frac{1}{2}$. The amplification factor in this case will be denoted by A_c^h. It is given by

$$A_c^h := \frac{\hat{\phi}_{m+\frac{1}{2}}^{n+1}}{\hat{\phi}_{m+\frac{1}{2}}^n} = 1+4\left(\frac{c}{\gamma}+\beta c\right)\left(\cos K\frac{h}{2}-1\right)-i\,2c\,\sin K\frac{h}{2} \tag{2.56}$$

To obtain the von Neumann stability condition in this case is an easy task. It can be done exactly as for linear elements. Omitting the details, the stability limit is found to be

$$c \le \min\left(\frac{\gamma}{4(1+\beta\gamma)}, \frac{1}{\gamma}+\beta\right) \tag{2.57}$$

It can be easily verified that

$$\frac{\gamma}{4(1+\beta\gamma)} \le \frac{1}{\gamma}+\beta \tag{2.58}$$

whenever $\beta \ge \beta_c$, with

$$\beta_c := \frac{1}{2}-\frac{1}{\gamma} \tag{2.59}$$

From the expression of the function β given in Box 2.1, it can be shown that $\beta \ge \beta_c$. Since (2.58) holds, (2.57) reduces to

$$c \le \frac{\gamma}{4(1+\beta\gamma)} \tag{2.60}$$

This is the stability requirement for the central nodes.

The expansion of A_c^h in powers of Δt and h yields

$$A_c^h = 1-(kK^2+iKu)\Delta t + O(\beta h\Delta t)+O(h^2\Delta t) \tag{2.61}$$

The analytical amplification factor is given again by (2.48). Exactly the same comments concerning the accuracy of linear elements may be done for this case. As it could be expected, the accuracy of the algorithm is driven by the time discretization. No improvement is obtained because of the use of quadratic elements.

Consider now the algorithmic equation for the extreme nodes (2.55). The corresponding amplification factor will be denoted by A_e^h. Its expression is

$$A_e^h := \frac{\hat{\phi}_m^{n+1}}{\hat{\phi}_m^n} = 1-7\left(\frac{c}{\gamma}+\alpha c\right)+8\left(\frac{c}{\gamma}+\alpha c\right)\cos K\frac{h}{2}-\left(\frac{c}{\gamma}+\alpha c\right)\cos Kh$$
$$+i\left[-\left(6\frac{c}{\gamma}\alpha-c\right)\sin Kh+4\left(3\frac{c}{\gamma}\alpha-c\right)\sin K\frac{h}{2}\right] \tag{2.62}$$

Let us introduce the abbreviations

$$\begin{aligned} z &:= \cos K\frac{h}{2} \\ b &:= \frac{c}{\gamma} + \alpha c \\ d &:= 6\frac{c}{\gamma}\alpha - c \end{aligned} \tag{2.63}$$

which after using some elementary trigonometric relations allow to write $|A_e^h|^2$ as

$$\begin{aligned} |A_e^h|^2(z) = 1 &- 4b(z-1)(z-3) + 4b^2(z-1)^2(z-3)^2 \\ &- (z-1)(z+1)[2(d-c) - 2dz]^2 \end{aligned} \tag{2.64}$$

The von Neumann criterion $|A_e^h|^2 \leq 1$ for any $z \in [-1,1]$ leads to the inequality $p(z) \geq 0$ in the interval $[-1,1]$, where $p(z)$ is the third degree polynomial

$$p(z) := -b(z-3) + b^2(z-1)(z-3)^2 - (z+1)(d-c-dz)^2 \tag{2.65}$$

There are two obvious necessary conditions for having $p(z) \geq 0$ in $[-1,1]$, viz.

$$p(1) \geq 0 \iff c^2 \leq b \tag{2.66}$$

$$p(-1) \geq 0 \iff b \leq \frac{1}{8} \tag{2.67}$$

We now prove that (2.66) and (2.67) are also sufficient. Let us write $p(z)$ as $p(z) = a_0z^3 + a_1z^2 + a_2z + a_3$. Suppose that $a_0 \neq 0$. This polynomial may have two local extrema, located at the abcissa z_1 and z_2 given by

$$z_1 = \zeta + \sigma \qquad \text{and} \qquad z_2 = \zeta - \sigma \tag{2.68}$$

where the notation

$$\zeta := -\frac{a_1}{3a_0}; \qquad \sigma := \frac{1}{3a_0}\left(a_1^2 - 3a_0a_2\right)^{\frac{1}{2}}$$

has been introduced. Assume now that the following two conditions hold: *(i)* $a_0 > 0$, *(ii)* $-a_1 \geq 3a_0$.

If $a_0 > 0$, then $p(z) \to +\infty$ as $z \to +\infty$ and $p(z) \to -\infty$ as $z \to -\infty$. Hence, if $p(z)$ has a local minimum located at z_m and a local maximum located at z_M then it must be $z_m \geq z_M$. From (2.68) it follows that $z_1 = z_m$ and $z_2 = z_M$. If condition *(ii)* holds, then $\zeta \geq 1$. Since $z_1 \geq \zeta$, we have that if $p(z)$ has a local minimum it is located out of the interval $[-1,1]$ and conditions (2.66) and (2.67) will suffice for stability.

It only remains to check inequalities *(i)* and *(ii)*. Expanding the polynomial $p(z)$ given by (2.65) it is found that

$$\begin{aligned} a_0 &= b^2 + d^2 > 0 \\ -a_1 - 3a_0 &= \frac{4c^2}{\gamma^2}[(1+\alpha\gamma)^2 + 6\alpha(\gamma - 6\alpha)] \end{aligned}$$

Since the upwind function α is $\leq \gamma/6$ we have that $-a_1 - 3a_0 \geq 0$, i.e., condition *(ii)* holds.

Inequalities (2.66) and (2.67) can be written as

$$c \leq \min\left(\frac{\gamma}{8(1+\alpha\gamma)}, \frac{1}{\gamma} + \alpha\right) \tag{2.69}$$

As for the cases considered before, we have that

$$\frac{\gamma}{8(1+\alpha\gamma)} \leq \frac{1}{\gamma} + \alpha \tag{2.70}$$

for $\alpha \geq \alpha_c$, where now the 'critical' value α_c is

$$\alpha_c := \frac{\sqrt{2}}{4} - \frac{1}{\gamma} \tag{2.71}$$

The upwind function α given in Box 2.1 verifies $\alpha \geq \alpha_c$. Inequality (2.70) holds and hence (2.69) may be simplified to

$$c \leq \frac{\gamma}{8(1+\alpha\gamma)} \tag{2.72}$$

This is the sought stability limit for the extreme nodes.

The algorithm will be stable only if both (2.60) and (2.72) hold. Since $\alpha \geq \beta$, we have that (2.72) is more restrictive than (2.60). Therefore, we finally obtain that (2.72) is the necessary and sufficient condition needed to satisfy the von Neumann stability criterion.

Remarks 2.2

(1) The key steps in which the behavior of the upwind functions α and β has been needed in the above development are $\beta \geq \beta_c$, $\alpha \geq \alpha_c$, $\alpha \leq \frac{1}{6}\gamma$ and $\alpha \geq \beta$, with α_c and β_c given by (2.71) and (2.59), respectively. In the previous chapter we have shown that the choice $\alpha = \beta = \min(\frac{\gamma}{6}, \frac{1}{2})$ may also be used. Clearly, for this unique upwind function also (2.72) is the stability condition.

(2) From (2.57) and (2.69) it follows that the Galerkin method ($\alpha = \beta = 0$) becomes unconditionally unstable when $\gamma \to \infty$.

(3) From the expression given for α, condition (2.72) reduces to

$$c \leq \frac{1}{8} \qquad \text{in the advective limit } (\gamma \to \infty) \tag{2.73}$$

and

$$\Delta t \leq \frac{h^2}{16k} \qquad \text{in the diffusive limit } (u \to 0) \tag{2.74}$$

These conditions are clearly far from being optimal. Instead of (2.73) one would hope $c \leq 1/2$ for the definition of the Courant number we have used. This limit depends on the upwind function α and it could be thought that this lack of 'optimality' is due to the choice of this function. However, (2.74) is independent of α and is also suboptimal if the result obtained for linear elements is taken as a reference. In particular, for a given set of nodes, the critical time step for stability will be higher using linear elements than quadratic elements (twice or four times, according to (2.74) or (2.73)). Nevertheless, the numerical results presented later

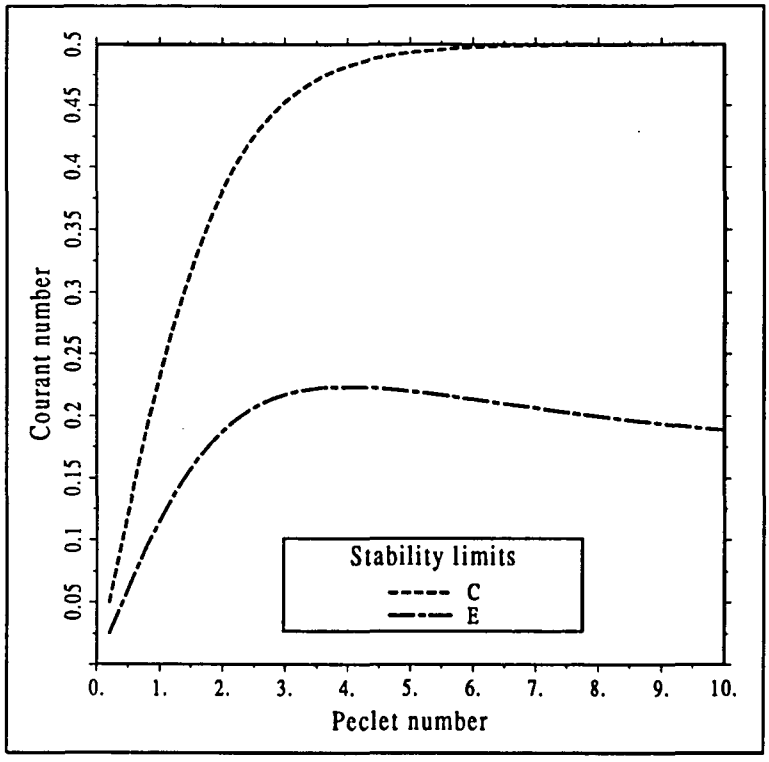

Figure 2.3 Stability limits for the convection-diffusion equation using quadratic elements. C: Central nodes, E: extreme nodes.

show that the steady-state is reached in a similar number of time steps using linear and quadratic elements. □

Figure 2.3 shows the stability limits dictated by (2.60) and (2.72). It is observed that the stability restriction imposed by the extreme nodes is much more severe than the one imposed by the central nodes.

So far, we have only considered *necessary* conditions for stability. However, we can prove that the diffusive limit (2.74) is also sufficient. To see this, we apply now the matrix method. Since for $\gamma = 0$ (and hence $\alpha = \beta = 0$) matrices $\mathbf{M}_d$ and $\mathbf{K}$ are symmetric, the spectral radius of $\mathbf{I} - \Delta t \mathbf{M}_d^{-1}\mathbf{K}$ is equal to its L^2–matrix norm,

$$\varrho_0 := \varrho(\mathbf{I} - \Delta t \mathbf{M}_d^{-1}\mathbf{K}) = \|\mathbf{I} - \Delta t \mathbf{M}_d^{-1}\mathbf{K}\|_2$$

The necessary and sufficient condition for stability will be $\varrho_0 \leq 1$. Since $\varrho_0 = |1 - \Delta t \varrho(\mathbf{M}_d^{-1}\mathbf{K})|$ this leads to

$$\Delta t \leq \frac{2}{\varrho(\mathbf{M}_d^{-1}\mathbf{K})} \tag{2.75}$$

Applying Irons' Theorem and solving an elementary eigenvalue problem we obtain

$$\begin{aligned}\varrho(\mathbf{M}_d^{-1}\mathbf{K}) &= \max_{\lambda}\{\lambda \mid \det(\mathbf{K}-\lambda\mathbf{M}_d)=0\}\\ &\le \max_{e}\ \max_{\lambda^e}\{\lambda^e \mid \det(\mathbf{K}^e-\lambda^e\mathbf{M}_d^e)=0\}\\ &= \max\left\{0,\frac{12k}{h^2},\frac{32k}{h^2}\right\}\\ &= \frac{32k}{h^2}\end{aligned}$$

where $\mathbf{M}_d^e = \frac{h}{2}\mathrm{diag}(\frac{1}{3},\frac{4}{3},\frac{1}{3})$ comes from the 'lumping' of $\mathbf{M}^e$. Since

$$\frac{2}{\varrho(\mathbf{M}_d^{-1}\mathbf{K})} \ge \frac{h^2}{16k}$$

it follows that (2.74) implies (2.75). Stability is ensured.

The results obtained up to now are summarized next.

Proposition 2.1 *Consider the forward Euler scheme defined by the algorithmic equations (2.54) and (2.55). Let the upwind functions be* $\alpha = \min(\frac{\gamma}{12},1)$ *and* $\beta = \min(\frac{\gamma}{12},\frac{1}{2})$. *Then, the algorithm satisfies the von Neumann stability condition if, and only if,*

$$c \le \frac{\gamma}{8(1+\alpha\gamma)}$$

Moreover, for $u = 0$ *the condition*

$$\Delta t \le \frac{h^2}{16k}$$

is both necessary and sufficient for stability. □

Not much new can be said about the accuracy when the extreme nodes are considered. The expansion of the amplification factor A_e^h in powers of Δt and h is

$$A_e^h = 1-(kK^2+iKu)\Delta t + O(\alpha h\Delta t) + O(h^2\Delta t)$$

that has the same form as Eqn. (2.61). What was said for the central nodes also applies here.

Algorithmic damping ratio (ADR) and frequency ratio (AFR)

As for linear elements, the *ADR* and the *AFR* have been plotted for values $c_0 =$ 0.25, 0.5, 0.75 and 0.95 of the security factor and $\gamma_0 = 0.1$, 1 and 10 for the convection factor. Here, the *ADR* and the *AFR* are referred to the extreme nodes. Now the numerical method can reproduce values of $\bar{K}$ in the whole interval $[0,2\pi]$, the upper bound corresponding to a single element per wavelength (three nodes).

From the plots shown in Figure 2.4, it is seen that the choice $c_0 = 0.5$ is also 'safe', as it happened to be for linear elements. The *ADR* is always ≥ 1. However, now the phase errors for this value of the security factor are higher than for linear elements. The salient point of these results is that $c_0 = 0.95$ gives values of the *ADR* higher than 1 in a range much wider than for linear elements, especially for $\gamma_0 = 10$. This indicates

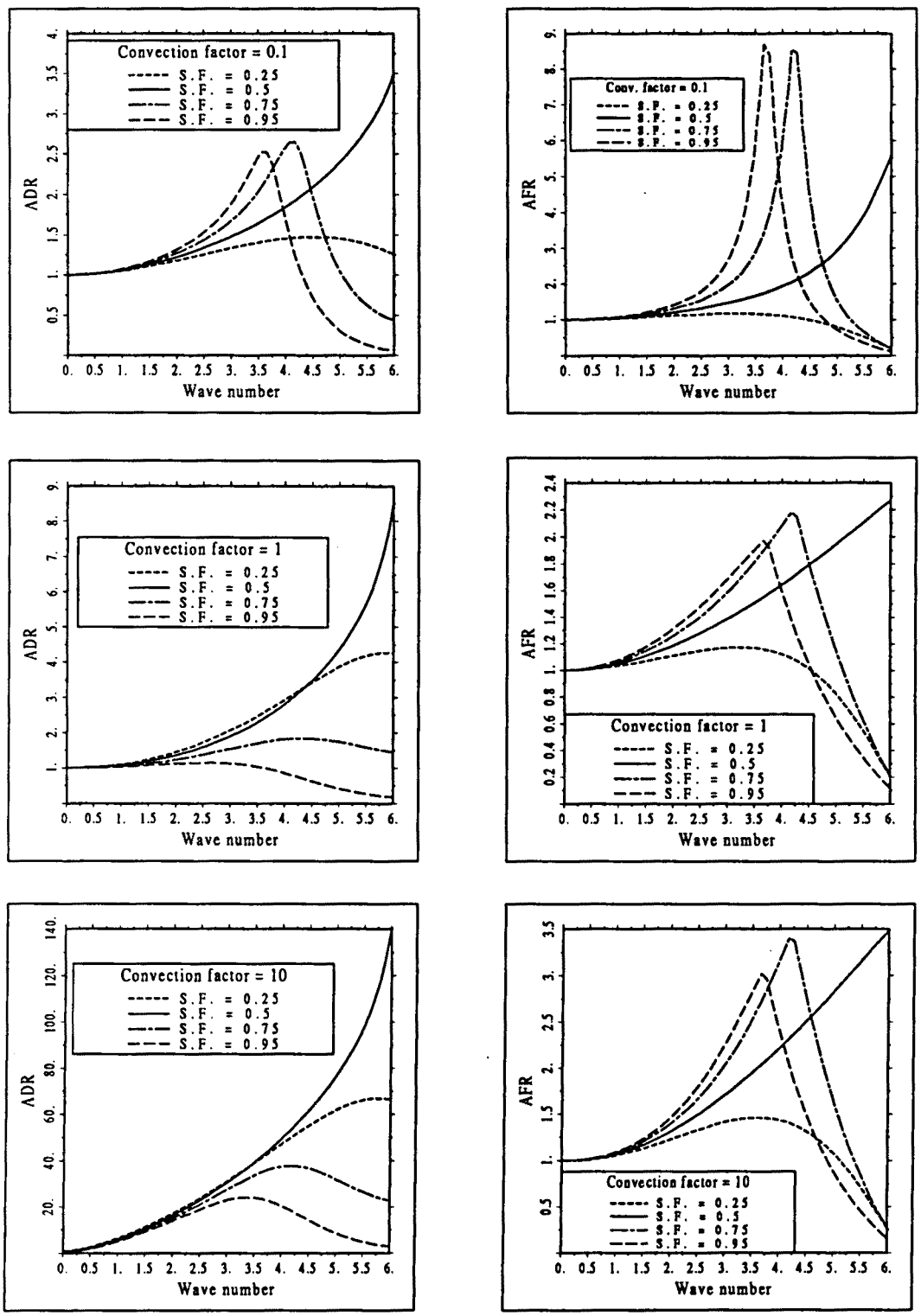

Figure 2.4 Algorithmic damping ratio (ADR) and algorithmic frequency ratio (AFR) for the SD method using quadratic elements and for different values of the security factor (S.F.) c_0 and the convection factor γ_0.

that the algorithm will have an important amount of numerical damping. Although this fact has a negative connotation when the time accuracy is crucial, it is beneficial if only the steady-state is sought. This explains why the number of time steps needed to reach the stationary solution are similar for linear and quadratic elements, even though the critical time step for stability is smaller using quadratic than linear interpolations.

2.3.4 Quadratic elements II : hierarchic approach

Now we consider that the quadratic finite element interpolation is done using the hierarchic basis. The shape functions for each element are shown in Figure 1.6 (Chapter 1).

The matrices $\mathbf{M}^e$ and $\mathbf{K}^e$ obtained using the SD method are

$$\mathbf{M}^e = \frac{h}{2}\begin{pmatrix} \frac{2}{3}-\frac{\alpha}{2} & \frac{2}{3}-\frac{2\alpha}{3} & \frac{1}{3}+\frac{\alpha}{6} \\ \frac{2}{3}+\frac{2\beta}{3} & \frac{16}{15} & \frac{2}{3}-\frac{2\beta}{3} \\ \frac{1}{3}+\frac{\alpha}{2} & \frac{2}{3}+\frac{2\alpha}{3} & \frac{2}{3}+\frac{\alpha}{2} \end{pmatrix}$$

$$\mathbf{K}^e = \frac{2k}{h}\begin{pmatrix} \frac{1}{2}+\gamma\left(-\frac{1}{2}+\frac{\alpha}{2}\right) & -2\alpha+\gamma\frac{2}{3} & -\frac{1}{2}+\gamma\left(\frac{1}{2}-\frac{\alpha}{2}\right) \\ -\gamma\frac{2}{3} & \frac{8}{3}+\gamma\frac{8\beta}{3} & \gamma\frac{2}{3} \\ -\frac{1}{2}+\gamma\left(-\frac{1}{2}-\frac{\alpha}{2}\right) & 2\alpha-\gamma\frac{2}{3} & \frac{1}{2}+\gamma\left(\frac{1}{2}+\frac{\alpha}{2}\right) \end{pmatrix}$$

where the upwind functions α and β are given in Box 2.1.

Diagonalization of $\mathbf{M}^e$

For simplicity, we will consider the matrix $\mathbf{M}^e$ with $\alpha = \beta = 0$. It is not clear how to obtain a diagonal matrix $\mathbf{M}_d^e$ that approximates $\mathbf{M}^e$. Now the property $\sum_{e=1}^{N_{el}} N_e(x) \equiv 1$ does *not* hold and the row-sum lumping technique does not make sense. If a nodal quadrature rule is used, the resulting matrix will not be diagonal, since now $N_i(x_j) \neq \delta_{ij}$, with the notation used earlier.

Let $\mathbf{N}^S$ be the vector whose components are the standard shape functions of a quadratic element and $\mathbf{N}^H$ the vector containing the hierarchic shape functions. Since $\mathbf{N}^S = \mathbf{T}\mathbf{N}^H$, with

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & -\frac{1}{2} \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}$$

the matrix $\tilde{\mathbf{M}}^e$ obtained with the standard formulation will be related to $\mathbf{M}^e$ by

$$\tilde{\mathbf{M}}^e = \mathbf{T}\mathbf{M}^e\mathbf{T}^T \tag{2.76}$$

Suppose that $\mathbf{M}^e$ is approximated by a diagonal matrix $\mathbf{M}_d^e$ of the form

$$\mathbf{M}_d^e = \frac{h}{3}\mathrm{diag}(\mu, \mu', \mu) \tag{2.77}$$

Formula (2.76) with $\mathbf{M}^e$ replaced by $\mathbf{M}_d^e$ given by (2.77) yields

$$\tilde{\mathbf{M}}^e = \frac{h}{3}\begin{pmatrix} \mu+\frac{1}{4}\mu' & -\frac{1}{2}\mu' & \frac{1}{4}\mu' \\ -\frac{1}{2}\mu' & \mu' & -\frac{1}{2}\mu' \\ \frac{1}{4}\mu' & -\frac{1}{2}\mu' & \mu+\frac{1}{4}\mu' \end{pmatrix} \tag{2.78}$$

This expression gives an idea of how the 'masses' of the extreme nodes μ and of the central node μ' have to be splitted in order to obtain a diagonal hierarchic matrix $\mathbf{M}_d^e$.

Now, let us try to obtain (2.78) by using a certain three point quadrature rule. Let $-\eta$, 0, η be the position of the integration points in the reference interval $[-1,1]$ and w, w', w the corresponding weights. Since the quadrature rule has to integrate exactly polynomials of degree at least one, w and w' must verify $2w+w'=2$. Moreover, if a matrix $\tilde{\mathbf{M}}_d^e$ of the form (2.78) is to be found, the following relations are needed

$$\begin{aligned} \frac{3}{4}w\eta^2(1+\eta^2) &= \mu + \frac{1}{4}\mu' \\ 3w\eta^2(1-\eta^2) &= -\mu' \\ 3 + 3w\eta^2(\eta^2-2) &= \mu' \end{aligned}$$

from where it follows that

$$\mu = \frac{3}{2} \qquad \text{and} \qquad \mu' = -3(1-\eta^2)$$

Since $0 \leq \eta \leq 1$, then $\mu' \leq 0$. Hence, matrix $\mathbf{M}_d^e$ and the assembled matrix $\mathbf{M}_d$ will be *not definite* and cannot be used to solve a transient problem $\mathbf{M}_d\dot{\boldsymbol{\phi}} + \mathbf{K}\boldsymbol{\phi} = \mathbf{0}$. However, the proposed itegration rule might be useful for other purposes such as the least-squares smoothing of a discontinuous function (see, e.g. [Hu], [ZT]).

Stability

Having in mind the previous considerations, a matrix $\mathbf{M}_d^e$ of the form (2.77) will be taken, with $\mu > 0$ and $\mu' > 0$. The transient evolution of the system $\mathbf{M}_d\dot{\boldsymbol{\phi}} + \mathbf{K}\boldsymbol{\phi} = \mathbf{0}$ will not be an approximation to the real problem (2.33). Only the steady-state of both problems will (hopefully) coincide. Scheme (2.35), with $\mathbf{M}$ replaced by $\mathbf{M}_d$ can be thought of as a Jacobi-type iterative method to reach this stationary solution.

As before, two different sets of algorithmic equations will be found for the internal nodes. If we define the scaled Courant numbers

$$\bar{c}' := \frac{c}{\mu'} \qquad \text{and} \qquad \bar{c} := \frac{c}{2\mu} \tag{2.79}$$

these equations are

- Central nodes:

$$\phi_{m+\frac{1}{2}}^{n+1} = 2\bar{c}'\phi_m^n + \left[1 - 8\left(\frac{\bar{c}'}{\gamma} + \beta\bar{c}'\right)\right]\phi_{m+\frac{1}{2}}^n - 2\bar{c}'\phi_{m+1}^n \tag{2.80}$$

- Extreme nodes:

$$\begin{aligned} \phi_m^{n+1} = &\left[\frac{3}{2}\frac{\bar{c}}{\gamma} + \frac{3}{2}\bar{c}(\alpha+1)\right]\phi_{m-1}^n + \left[-6\alpha\frac{\bar{c}}{\gamma} + 2\bar{c}\right]\phi_{m-\frac{1}{2}}^n \\ &+ \left[1 - 3\frac{\bar{c}}{\gamma} - 3\alpha\bar{c}\right]\phi_m^n + \left[6\alpha\frac{\bar{c}}{\gamma} - 2\bar{c}\right]\phi_{m+\frac{1}{2}}^n \\ &+ \left[\frac{3}{2}\frac{\bar{c}}{\gamma} + \frac{3}{2}\bar{c}(\alpha-1)\right]\phi_{m+1}^n \end{aligned} \tag{2.81}$$

The amplification factor for Eqn. (2.80) is

$$A_c^h := \frac{\hat{\phi}^{n+1}_{m+\frac{1}{2}}}{\hat{\phi}^n_{m+\frac{1}{2}}} = 1 - 8\left(\frac{\bar{c}'}{\gamma} + \beta\bar{c}'\right) - i\,4\bar{c}'\sin K\frac{h}{2}$$

which verifies $|A_c^h| \leq 1$ whenever

$$\bar{c}' \leq \gamma\frac{1+\beta\gamma}{\gamma^2 + 4(1+\beta\gamma)^2} \tag{2.82}$$

Remarks 2.3

(1) We see again that when $\beta = 0$ (Galerkin method) the scheme is unconditionally unstable for $\gamma \to \infty$

(2) If the upwind function β given in Box 2.1 for hierarchic elements is used, the advective and diffusive limits have the stability conditions

$$c \leq \frac{\mu'}{4} \qquad (\text{for } \gamma \to \infty) \tag{2.83}$$

and

$$\Delta t \leq \frac{\mu' h^2}{8k} \qquad (\text{for } u \to 0) \tag{2.84}$$

(3) Clearly, the magnitude of μ' affects proportionally the time step size. What will be important is the ratio μ'/μ, and not μ or μ' themselves.

(4) The discrete modes $\hat{\phi}^n_m$ given by (2.38) have to be considered from the series expansion *of the error*, since now $\phi^n_{m+1/2}$ are not the nodal values of the unknown function. □

Let us consider now the stability for the extreme nodes. The amplification factor associated to Eqn. (2.81) is

$$A_e^h := \frac{\hat{\phi}^{n+1}_m}{\hat{\phi}^n_m} = 1 - 3\frac{\bar{c}}{\gamma} - 3\alpha\bar{c} + 3\left(\frac{\bar{c}}{\gamma} + \alpha\bar{c}\right)\cos K\frac{h}{2} + i\left[-2\bar{c}\sin Kh + \left(12\alpha\frac{\bar{c}}{\gamma} - 4\bar{c}\right)\sin K\frac{h}{2}\right] \tag{2.85}$$

Defining

$$\begin{aligned} z &:= \cos K\frac{h}{2} \\ b &:= \frac{\bar{c}}{\gamma} + \alpha\bar{c} \\ d &:= 3\frac{\bar{c}}{\gamma}\alpha - \bar{c} \end{aligned} \tag{2.86}$$

the square of the modulus of A_e^h can be written as

$$|A_e^h|^2(z) = \left[1 + 6(z^2-1)b\right]^2 + 16(1-z^2)(d - \bar{c}z)^2$$

Requiring $|A_e^h|^2 \leq 1$ for $z \in [-1,1]$ leads to $p(z) \geq 0$ for z in this interval, where $p(z)$ is the polynomial

$$p(z) = 3b + 9b^2(z^2-1) - 4(d - \bar{c}z)^2 \tag{2.87}$$

Since the upwind function satisfies $\alpha < \frac{\gamma}{3}$, we have that $d < 0$. Rewrite $p(z)$ as $p(z) = a_0 z^2 + a_1 z + a_2$, with

$$a_0 := 9b^2 - 4\bar{c}^2$$
$$a_1 := 8d\bar{c} \qquad (< 0)$$
$$a_2 := 3b - 9b^2 - 4d^2$$

Two cases will be distinguished:

- $a_0 \leq 0$. In this case, $p(z) \geq 0$ in $[-1,1]$ if $p(1) = a_0 + a_1 + a_2 \geq 0$ and $p(-1) = a_0 - a_1 + a_2 \geq 0$. Since $a_1 < 0$, the former condition is more restrictive. It leads to

$$\bar{c} \leq \frac{3\gamma(1+\alpha\gamma)}{4(3\alpha - 2\gamma)^2} \tag{2.88}$$

- $a_0 > 0$. The polynomial $p(z)$ has a local minimum located at $z_1 = -a_1/2a_0 > 0$. The value of this minimum is

$$p(z_1) = -\frac{1}{4}\frac{a_1^2}{a_0} + a_2$$

Two subcases have to be considered:

- If $z_1 \geq 1$, i.e., $-a_1 \geq 2a_0$, condition (2.88) is enough for stability.
- If $z_1 < 1$, then $p(z_1) \geq 0$ is required. This leads to

$$\bar{c} \leq \gamma \frac{9(1+\alpha\gamma)^2 - 4\gamma^2}{27(1+\alpha\gamma)^3 + 12(1+\alpha\gamma)[(3\alpha-\gamma)^2 - \gamma^2]} \tag{2.89}$$

The final stability condition is rather cumbersome to define: if $a_0 \leq 0$ or $a_0 > 0$ and $z_1 \geq 1$ then (2.88) is needed; else, (2.89) is necessary. The value of γ that determines if one or another condition has to hold will be denoted by γ_c. For the upwind function $\alpha = \min(\frac{\gamma}{15}, \frac{1}{3})$, it is found that $\gamma_c \approx 1.23$. Below this value, (2.89) is the stability condition, whereas (2.88) has to be verified for higher values of γ.

Figure 2.5 shows the stability limits for the central nodes (condition (2.82)) and for the extreme nodes of hierarchic elements, always in terms of the scaled Courant numbers (c/μ' and $c/2\mu$).

The advective stability limit for extreme nodes is found by taking $\gamma \to \infty$ in (2.88) and the diffusive limit by taking $u \to 0$ in (2.89). The results are

$$c \leq \frac{\mu}{8} \qquad \text{for } \gamma \to \infty \tag{2.90}$$

$$\Delta t \leq \frac{\mu h^2}{3k} \qquad \text{for } u \to 0 \tag{2.91}$$

Since the values of μ and μ' have been considered independent, both the two conditions (2.82) and (2.88) or (2.89) must hold in order to have a stable scheme. Concerning the diffusive limit, inequalities (2.84) and (2.91) may be written together as

$$\Delta t \leq \min\left(\frac{\mu' h^2}{8k}, \frac{\mu h^2}{3k}\right) \tag{2.92}$$

Exactly as for the case of standard shape functions, condition (2.92) is not only necessary but also sufficient for stability, since when $u = 0$ it is found that

$$\text{Spec}(\mathbf{M}_d^e)^{-1}\mathbf{K}^e = \left\{0, \frac{16k}{\mu' h^2}, \frac{6k}{\mu h^2}\right\}$$

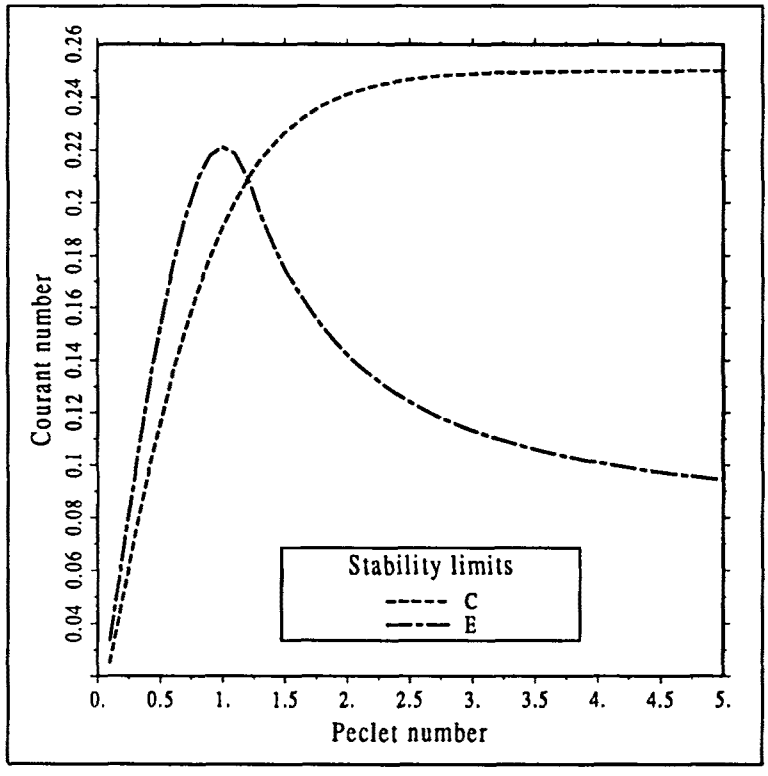

Figure 2.5 Stability limits for the convection-diffusion equation using quadratic hierarchic elements. C: Central nodes, E: extreme nodes.

The following proposition summarizes the results that have been obtained.

Proposition 2.2 *Consider the forward Euler scheme defined by the algorithmic equations (2.80) and (2.81). Let the upwind functions be* $\alpha = \min(\frac{\gamma}{15}, \frac{1}{3})$ *and* $\beta = \min(\frac{\gamma}{12}, \frac{1}{2})$, *and* γ_c *the value of the Péclet number defined above* (≈ 1.23). *Then, the algorithm satisfies the von Neumann stability condition if, and only if,*

$$c \leq \begin{cases} \min(\mu' c_1, 2\mu c_2) & \text{if } \gamma \leq \gamma_c \\ \min(\mu' c_1, 2\mu c_3) & \text{if } \gamma > \gamma_c \end{cases}$$

where c_1, c_2 *and* c_3 *are given by*

$$c_1 := \gamma \frac{1+\beta\gamma}{\gamma^2 + 4(1+\beta\gamma)^2}$$

$$c_2 := \frac{3\gamma(1+\alpha\gamma)}{4(3\alpha - 2\gamma)^2}$$

$$c_3 := \gamma \frac{9(1+\alpha\gamma)^2 - 4\gamma^2}{27(1+\alpha\gamma)^3 + 12(1+\alpha\gamma)[(3\alpha-\gamma)^2 - \gamma^2]}$$

Moreover, for $u = 0$ *the condition*

$$\Delta t \leq \min\left(\frac{\mu' h^2}{8k}, \frac{\mu h^2}{3k}\right)$$

is both necessary and sufficient for stability. □

Remark 2.4

The values of μ and μ' have been considered independent but there is a simple criterion to relate them in order to speed up the convergence of the algorithm to the steady-state. With the notation used in Proposition 2.2, we could take μ and μ' such that

$$\mu' c_1 = \begin{cases} 2\mu c_2 & \text{if } \gamma \leq \gamma_c \\ 2\mu c_3 & \text{if } \gamma > \gamma_c \end{cases} \tag{2.93}$$

This choice will not increase the critical time step, but will ensure that both the equations for the central nodes and for the extreme nodes advance in 'time' as fast as possible. In particular, for the advective limit it is easy to see that (2.93) yields $\mu = 2\mu'$, and for the diffusive limit $\mu = \frac{3}{8}\mu'$. From numerical experiments we have found that this reduces the number of required time steps about a 10% for convection dominated problems. □

2.3.5 Extension to multidimensional problems

The Fourier analysis of a difference scheme in a general multidimensional mesh is a difficult, if not impossible, goal. The expression of the critical time step in these cases is necessarily based on heuristic criteria. In Reference [HGG] the following partial result was proved for the finite difference method. Assume that the domain is two-dimensional (for simplicity), discretized using a uniform grid and that the centered five-point stencil is used to approximate both the first and the second spatial derivatives. Define

$$\gamma_x := \frac{u_x h_x}{2k_x}, \qquad \gamma_y := \frac{u_y h_y}{2k_y},$$
$$c_x := \frac{u_x \Delta t}{h_x}, \qquad c_y := \frac{u_y \Delta t}{h_y},$$

where subscripts x and y refer to the Cartesian directions. Under these conditions, the forward Euler scheme satisfies the von Neumann stability condition if, and only if

$$c_x\gamma_x + c_y\gamma_y \leq 1 \qquad \text{and} \qquad \frac{c_x}{\gamma_x} + \frac{c_y}{\gamma_y} \leq 1 \tag{2.94}$$

Let Δt_x and Δt_y be the critical time steps that would be found if the problem were one-dimensional along the x and y directions, respectively. Observe that (2.94) can be written as

$$\Delta t \Delta t_x^{-1} + \Delta t \Delta t_y^{-1} \leq 1 \tag{2.95}$$

Now suppose that a local system of coordinates σ and ν is taken on each point, σ following the streamline and ν normal to it. If we make the assumption that (2.95) still holds true in this new coordinate system, Δt must satisfy

$$\Delta t \leq \frac{\Delta t_\sigma \Delta t_\nu}{\Delta t_\sigma + \Delta t_\nu} \tag{2.96}$$

The problem is now reduced to compute Δt_σ and Δt_ν. Since the velocity follows the σ–direction, Δt_ν is calculated using the diffusive stability limits and Δt_σ using the general expressions obtained for one-dimensional convection-diffusion.

In general situations what we do is the following. Let Δt^e be the critical time step computed using (2.96) for element e, i.e., using its characteristic diffusion and element length and the Euclidian norm of the characteristic velocity within this element. The global time step is then taken as

$$\Delta t = f_t \min_e (\Delta t^e) \tag{2.97}$$

where f_t acts as a safety factor. In the numerical results presented thereafter (examples 2.4 and 2.5) we have found $f_t = 1$ effective in all the cases except for the six-noded triangular element, where $f_t < 1$ has been needed. This method has also been successfully applied to a quite different problem in Reference [Co2], where the equations arising from elliptic mesh generation are solved via a fictious transient.

Another question that arises using finite elements is the way a diagonal approximation to the mass matrix is obtained. We have used standard nodal quadrature rules for the next examples. Since the weights of the classical second order rule for the quadratic triangle are zero for the corner nodes, this element has been splitted into four linear triangles and the weights for these subelements have been utilized (see Chapter 4).

2.4 Numerical examples

Example 2.1 In this example, the transient problem (2.29)–(2.30) has been solved, not with the periodic boundary conditions (2.31) but with $\phi(0,t) = 0$ and $\phi(\ell,t) = 1$. The data of the problem are $\ell = 1$, $u = 1$, $k = 0.01$ and $\phi^0(x) = x$. The analytical solution for this problem can be expressed as [Co1]

$$\phi(x,t) = x + \sum_{n=1}^{\infty} \frac{B_n}{A_n} \left(1 - e^{-A_n t}\right) e^{\frac{u}{2k}x} \sin(n\pi x)$$

with

$$A_n = \frac{u^2}{2k} + k(n\pi)^2$$

$$B_n = \frac{2nu\pi}{\frac{u^2}{4k^2} + n^2\pi^2} \left[(-1)^n e^{-\frac{u}{2k}} - 1\right]$$

The discretization of the interval $[0,1]$ consists of ten *quadratic* elements of equal length 0.1, yielding a Péclet number $\gamma = 5$. Results are shown in Figure 2.6.

The solution obtained using the Galerkin method in space and the Crank-Nicolson scheme in time is depicted in Figure 2.6.(a) (the backward Euler method has been used for the first time step). Observe that for this rather small Péclet number, oscillations occur even at an early stage. The different curves correspond to the times $t = 0$, 0.25, 0.5, 1 and 2. Figure 2.6.(b) shows the numerical solution obtained with the SD method. This result is indistinguishable from the analytical solution given above. Both using the Galerkin and the SD method, the time step has been taken as $\Delta t = 0.05$.

The solution obtained using the forward Euler scheme in time is shown in Figure 2.6.(c). The time step that has been used is the maximum value allowed by formula (2.72). The plots correspond to $t = 0.242$, 0.506, 0.991 and 2.003. It is observed that the agreement with the previous results is excellent.

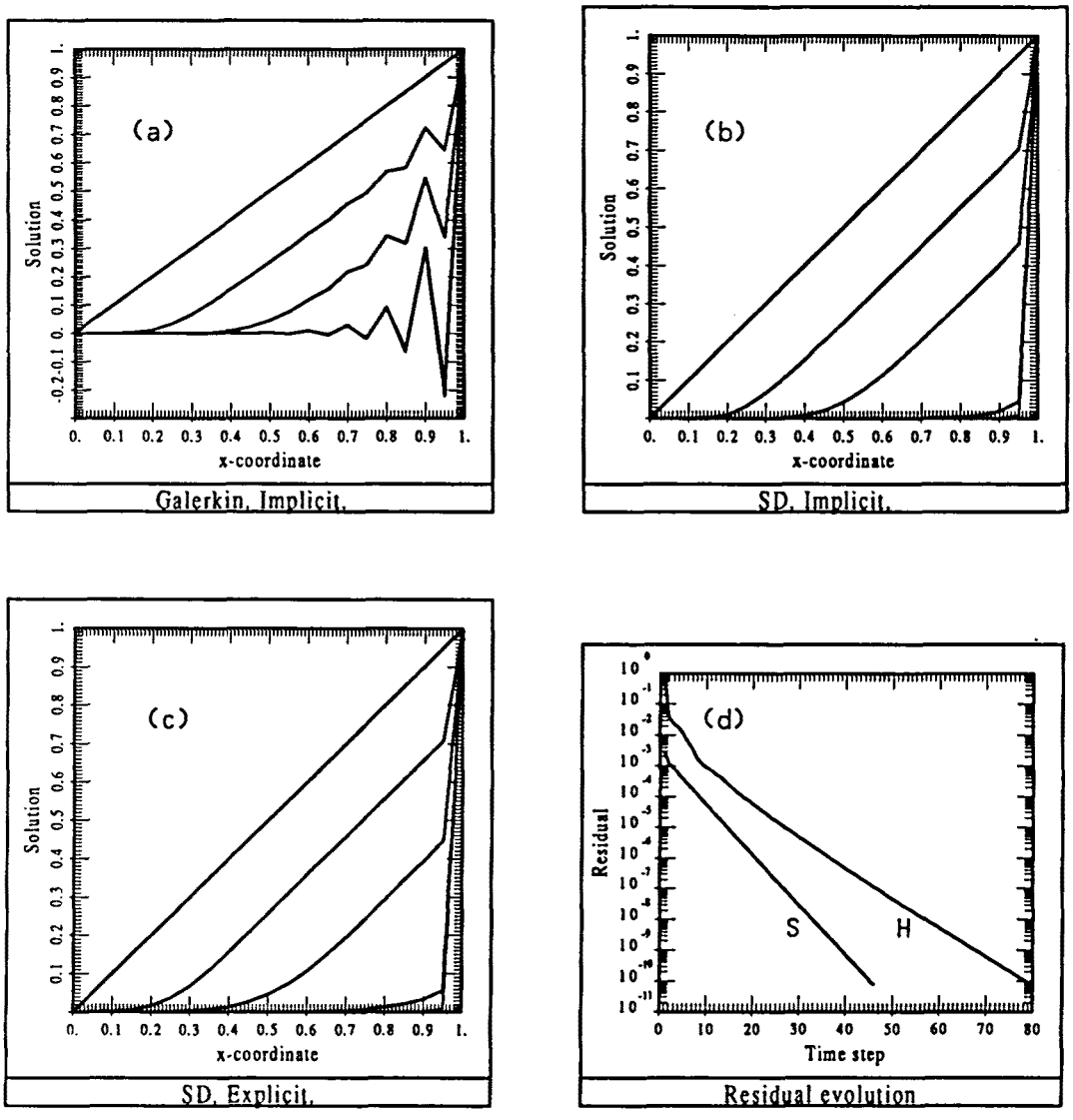

Figure 2.6 Results for example 2.1. (a): Galerkin solution using the Crank-Nicolson scheme. (b): SD solution using the Crank-Nicolson scheme. (c): SD solution using the forward Euler scheme. (d): Evolution to the steady-state using the standard (S) and the hierarchic (H) shape functions.

Finally, Figure 2.6.(d) shows the evolution of the residual $\max_m |\phi_m^{n+1} - \phi_m^n|$ as time goes on towards the steady-state using both the standard and the hierarchic shape functions. For the latter case, μ has been set equal to 1 and μ' has been calculated using (2.93). If μ' is also set to 1, the number of time steps required to reach the residual 10^{-10} is 85 instead of 79. It is seen that the standard approach reaches faster the stationary solution.

Example 2.2 This example has been taken from [YH1] and is useful to check the accuracy of the temporal discretization. Problem (2.29)–(2.30) has been solved with boundary conditions $\phi(0,t) = 0$ and $\frac{\partial \phi}{\partial x}(\ell,t) = 0$. The initial condition is

$$\phi^0(x) = e^{-(x-u)^2/4k}$$

This problem has an analytical solution given by

$$\phi(x,t) = \frac{1}{\sqrt{1+t}}\, e^{-[x-u(t+1)]^2/4k(t+1)}$$

The data are now $\ell = 2$, $u = 0.25$ and $k = 0.00125$. The space discretization has been done using 81 equal spaced nodal points. Both linear and quadratic elements have been considered (80 and 40, respectively). The resulting Péclet number is 2.5 for linear elements and 5 for quadratic elements. When the Crank-Nicolson scheme has been used, the time step has been taken as $\Delta t = 0.1$, yielding a Courant number $c = 0.5$ for quadratic elements and $c = 1$ for linear elements.

Figure 2.7.(a) shows the solution obtained using linear elements and the forward Euler scheme in time with a security factor $c_0 = 1$, as well as the analytical solution for $t = 0.197$ and 3.946. As it was already explained, some modes of the Fourier series expansion of the analytical solution are not properly damped by the numerical algorithm for this choice of c_0. The underdiffusive behavior of the numerical solution is evident, showing that the method is potentially oscillatory in more complicated situations. The results obtained under the same conditions but with $c_0 = 0.5$ are depicted in Figure 2.7.(b). As expected, the numerical answers show a much higher dissipation. In fact, they are a little overdiffusive. It is important to point out that the phase accuracy is very good, as it was predicted from the behavior of the algorithmic frequency ratio. If quadratic elements are used, this phase accuracy is not so high, as it can be observed from Figure 2.7.(c), where the solution for t =1.761 and 4.006 has been represented. Nevertheless, now a security factor $c_0 = 1$ can be used without suffering from underdiffusive behavior. All these observations confirm the theoretical predictions that had been obtained.

The results obtained using the Crank-Nicolson scheme with linear elements are plotted in Figure 2.7.(d) for $t = 2$ and 4, showing a much higher accuracy than the forward Euler method. If quadratic elements are used in this case, the numerical solution is almost the same. Also, if the Galerkin formulation is employed, only small amplitude oscillations are present (not shown), since the exact solution is now very smooth.

Example 2.3 We consider in this example the Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0, \qquad 0 < x < 1, \quad t > 0$$

with boundary and initial conditions

$$\begin{aligned} u(0,t) = u(1,t) &= 0, \qquad t > 0 \\ u(x,0) &= \sin \pi x, \qquad 0 < x < 1 \end{aligned}$$

This problem has been taken from [BDH], where different solutions obtained by different investigators using spectral methods are reported. It is considered as a test for this

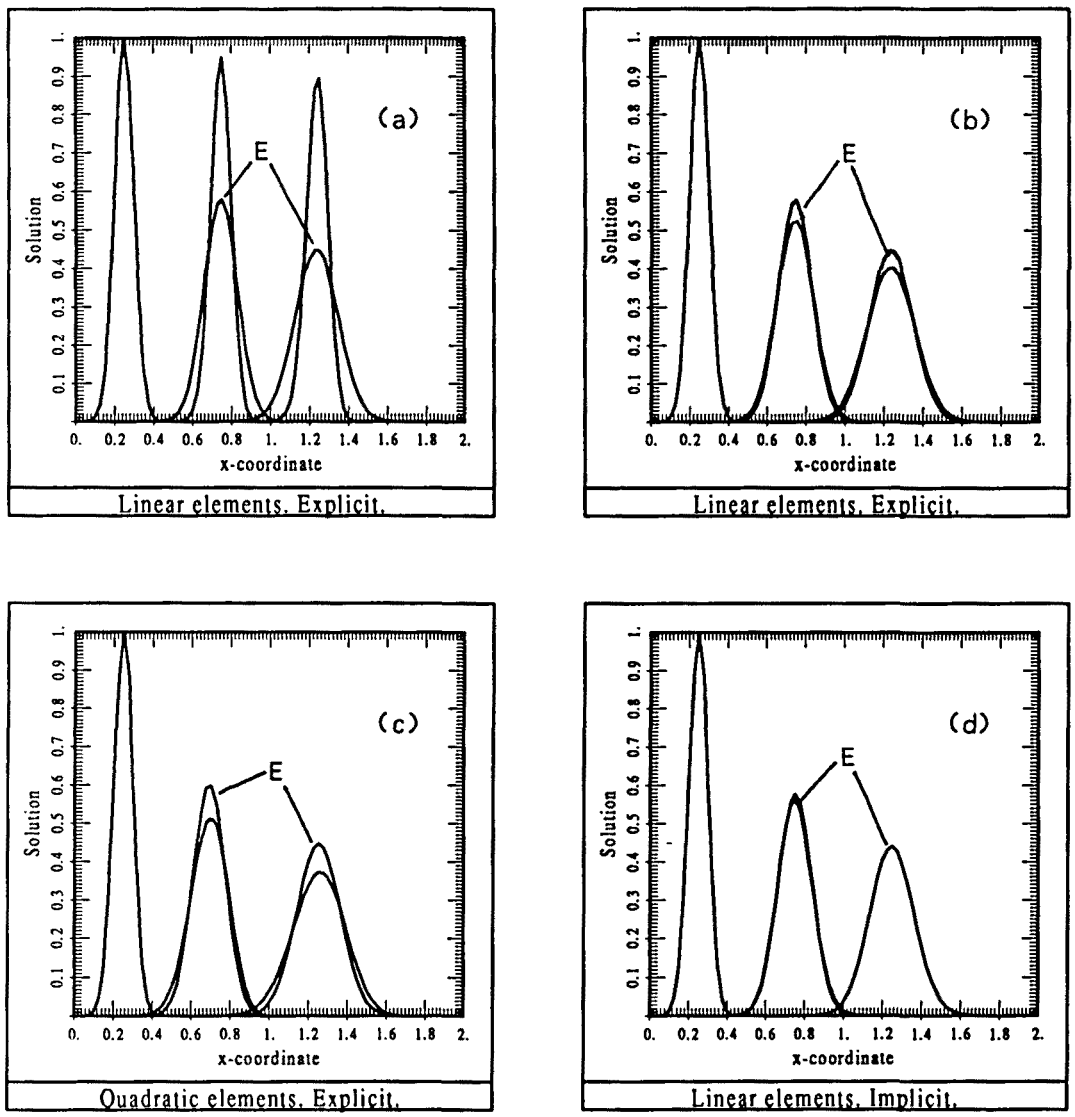

Figure 2.7 Results for example 2.2. Marker 'E' stands for exact solution. (a): Linear elements and forward Euler with $c_0 = 1$. (b): Linear elements and forward Euler with $c_0 = 0.5$. (c): Quadratic elements and forward Euler with $c_0 = 1$. (d): Linear elements and Crank-Nicolson.

type of numerical methods for problems in which the solution develops rapid variations, but not discontinuities. This happens when the viscosity ν is small. In this example, it will be taken as $\nu = 1/100\pi$.

Although in this chapter only the convection-diffusion equation has been considered, this problems fits naturally in this context once a linearization procedure for the nonlinear term has been chosen. Here, the simplest Picard method has been employed.

The discretization of the domain has been carried out using 40 quadratic elements whose lengths decrease exponentially from $x = 0$ to $x = 1$. The minimum element length is $h_1 = 0.01$, approximately. This concentration of elements at $x = 1$ is needed if the sharp profile that the solution develops there is to be reproduced accurately.

The tolerance of the iterative scheme has been taken as 10^{-8}, checking convergence in the discrete L^∞ norm. The maximum number of iterations required to reach this tolerance has been six. The time discretization uses the Crank-Nicolson method, with a time step $\Delta t = 4.67 \times 10^{-3}$, which corresponds to 150 time steps in 0.7 time units.

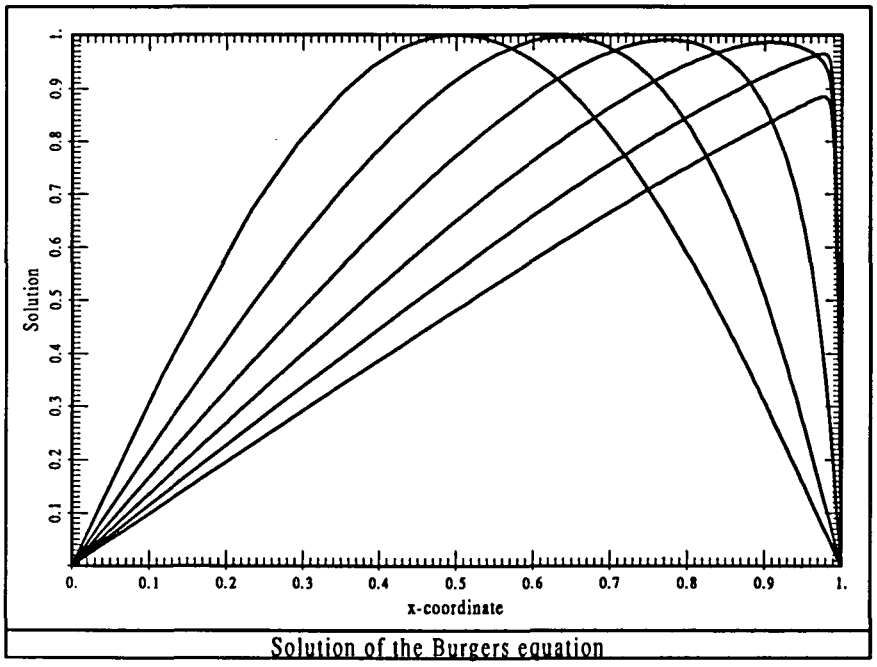

Figure 2.8 Solution of example 2.3

The solution is shown in Figure 2.8 for t =0, 0.14, 0.28, 0.42, 0.56 and 0.7. Figure 2.9 shows the time evolution of the slope of $u(x,t)$ at $x = 1$ and Figure 2.10 the time evolution of $\max_x u(x,t)$. These functions are used in [BDH] to compare the performance of different spectral methods.

The results presented here show an accuracy similar to the best result described in [BDH], except for the maximum absolute value of $\frac{\partial u}{\partial x}(1,t)$. The spectral method considered consists in a collocation procedure using Chebyshev polynomials and the ABCN scheme in time (Adams-Brashforth for the convective part, Crank-Nicolson for the viscous term). The discretization is done using 64 modes and a time step $\Delta t = 1.06 \times 10^{-3}$. Thus, the computational effort using this approach is much higher than using the finite element formulation described here (CPU times are not given in [BDH]). The only remarkable difference in the results is the maximum slope at $x = 1$. The analytical value is -152.005, obtained for $t = 0.5105$. The slope of the spectral method solution is -152.05, encountered for $t = 0.509$. The maximum slope for the finite element method is found to be -163.73 for $t = 0.513$. This inaccuracy could be expected, since the

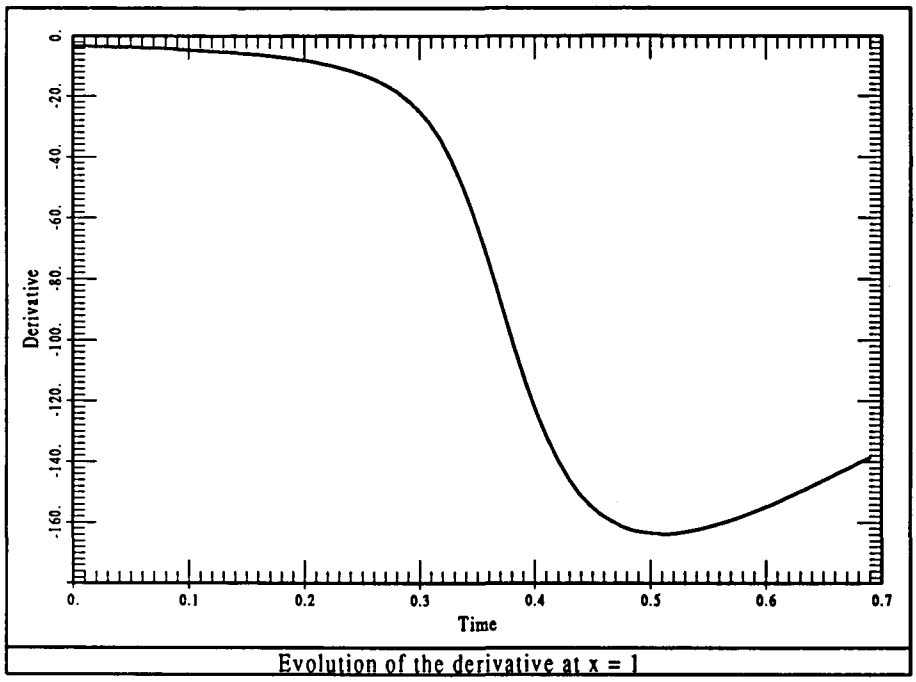

Figure 2.9 Evolution of the derivative at $x = 1$ for example 2.3

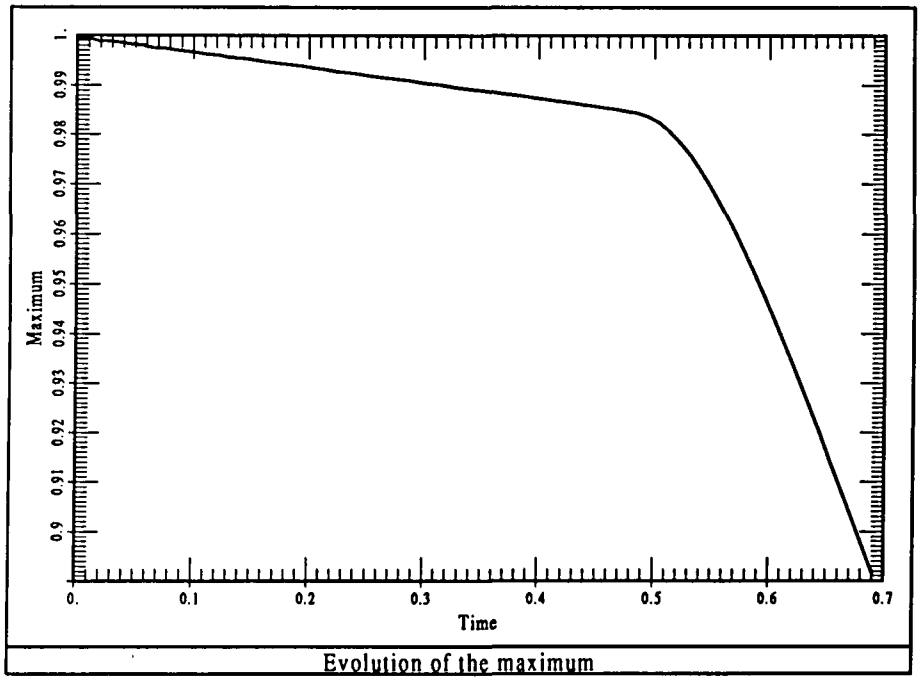

Figure 2.10 Evolution of the maximum value of the solution for example 2.3

derivatives at the nodes are not well calculated using finite elements (in fact, they are discontinuous for internal nodes). The expression we have used to compute $\frac{\partial u}{\partial x}(1,t)$ is

$$\frac{\partial u}{\partial x}(1,t) \approx \frac{2}{h_1}\left(\frac{1}{2}u_{79} - 2u_{80} + \frac{3}{2}u_{81}\right)$$

which comes from the differentiation of the finite element interpolation within the last element and the evaluation at $x = 1$.

The conclusion of this example is that the SD formulation presented here using quadratic elements and the Crank-Nicolson scheme in time is very accurate. Even the calculation of the nodal derivatives, known to be very inexact, gives reasonable results.

The reader may consult [COC] for another example in which the Burgers equation is solved and [HSG] for a Petrov-Galerkin method especially designed for this equation.

Example 2.4 The numerical test presented in example 1.2 of Chapter 1 is now solved using the forward Euler scheme for the transient equation as a way to reach the stationary solution. We are now interested in the evolution of the residual $\max_m |\phi_m^{n+1} - \phi_m^n|$ as time advances. The subscript refers to a nodal point. The results obtained for elements of 3, 4, 6 and 9 nodes are shown in Figure 2.11. Formula (2.97) has been used to compute the critical time step. In all the cases except for the six-noded triangular element, $f_t = 1$ has been used. For the quadratic triangle, $f_t = 0.9$ has been needed, since the time marching scheme has been found to be unstable for $f_t = 1$.

It is seen that the steady-state is reached faster for quadrilateral elements than for triangles. The bilinear element 'converges' slightly faster than the biquadratic one. The difference is more pronounced for the 3 and 6–noded elements.

At this point, it is interesting to make the following observation. Quadratic elements are often blamed to be more expensive than linear elements. The main argument is that the bandwidth of the final 'stiffness' matrix is larger. On the other hand, the total number of numerical quadrature points for a mesh with a given number of nodes is smaller. For example, if a 2×2 Gauss-Legendre quadrature rule is used for the bilinear element and a 3×3 rule for the biquadratic one, the ratio of total quadrature points of the former and the latter is 16/9. The important fact is that if an iterative method for solving the algebraic system of equations is used, the problem of the large bandwidth disappears and quadratic elements could be cheaper. In order to verify this hypothesis in this particular problem, the CPU time required on a CONVEX–C1 computer has been calculated. The results are the following:

No. of nodes	No. of integration points	CPU (seconds)
3	3	2.01
4	4 (2×2)	0.77
6	4	1.33
9	9 (3×3)	0.66

It is observed that the CPU time is smaller for quadratic elements than for linear elements, even though more time steps have to be performed to reach the steady-state.

Example 2.5 Here, the same test as in example 2.4 has been carried out, now for the problem presented in example 1.4 of Chapter 1. As before, the factor f_t of formula

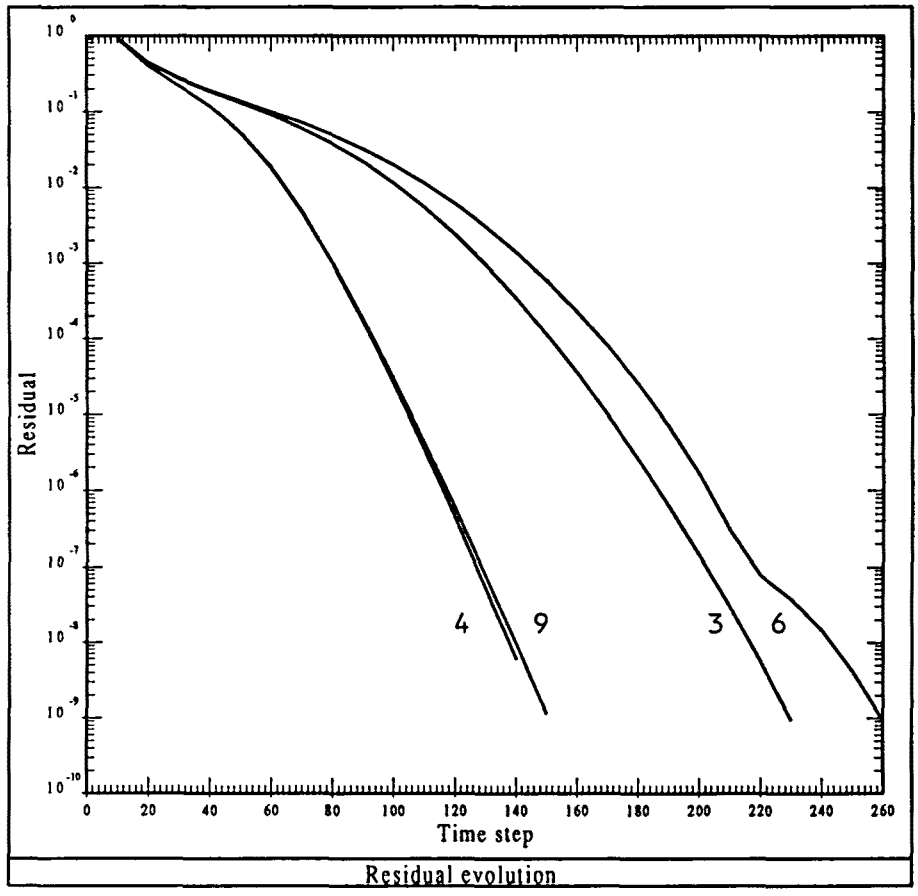

Figure 2.11 Evolution towards the steady-state in example 1.2. The number of nodes of the elements is indicated on each curve.

(2.97) has been set equal to 1 for all the elements except for the 6-noded triangle. In this case, $f_t = 0.5$ has been taken, since for higher values the forward Euler scheme happens to be unstable. From the results shown in Figure 2.12 it is observed that now the element that gives the best performance is the biquadratic one. Concerning the CPU times, the conclusions are the same as for example 2.4.

2.5 Summary and conclusions

Two main issues have been adressed in this chapter. The first is a fairly comprehensive description of the generalized trapezoidal rule applied to the convection-diffusion equation and combined with the SD method for the space discretization. This is the method that will be used in the rest of this work, although the possibility of using the discontinuous Galerkin method has been left open. The need for using the backward

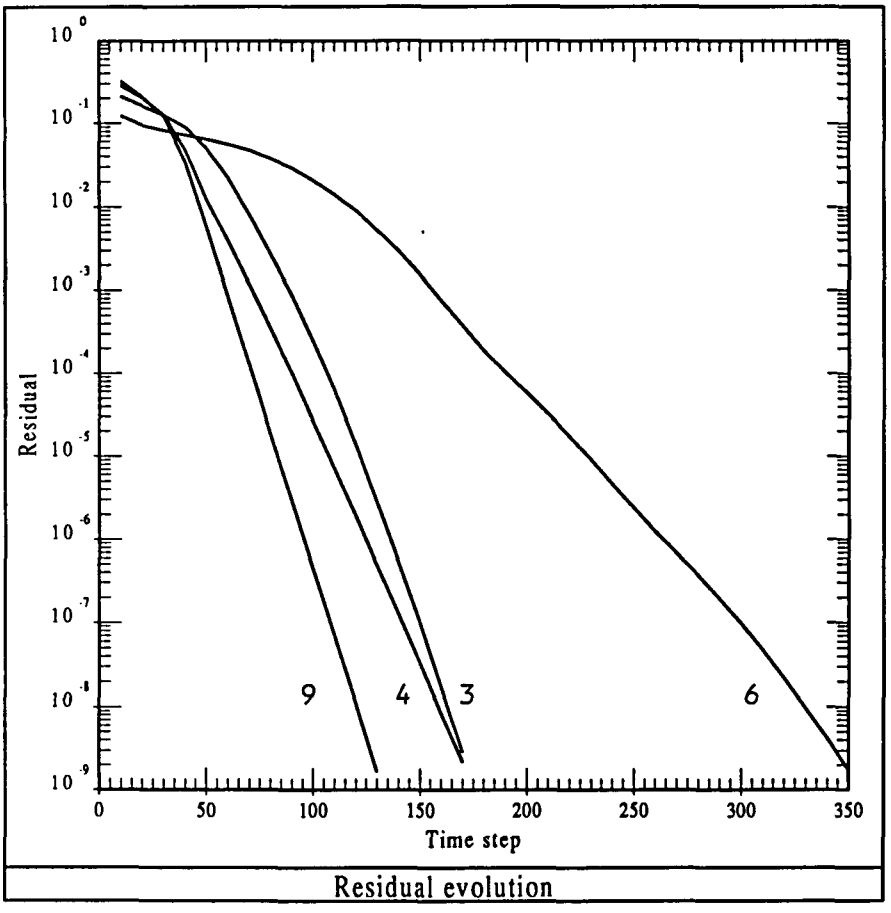

Figure 2.12 Evolution towards the steady-state in example 1.4. The number of nodes of the elements is indicated on each curve.

Euler method in some situations has been discussed, as well as the choice of the intrinsic time of the SD formulation for transient problems.

However, the main part of this chapter has been devoted to the stability and accuracy analysis of the forward Euler scheme, using both linear and quadratic finite elements. The interest of this method relies basically on the fact that it allows to obtain stationary solutions via an iterative procedure. This technique is ubiquitous in computational fluid dynamics, especially when the numerical simulation of compressible flow problems is attempted. Thus, its analysis has an inherent interest, regardless of the fact that we will not use it for incompressible flow because of the reasons that will be explained in Chapter 4.

The new results that have been obtained are now summarized:

- *Linear elements.* It has been shown that the upwind function yields optimal stability limits, in the sense that both the advective and diffusive limit cases reduce to conditions known to be optimal.
- *Accuracy.* A new methodology has been proposed in order to determine the effect

of convection and the time step size on the accuracy of the algorithm. The method is based on the representation of the ADR and the AFR for different values of the security factor c_0 and the convection factor γ_0 introduced here.

- *Time step for linear elements.* Using the technique just mentioned, it has been shown that if linear elements are used the choice $c_0 = 1$ may lead to unphysical results for the transient evolution of convection-diffusion problems. This drawback is circumvented if $c_0 = 0.5$ is selected. Moreover, an excellent phase accuracy is obtained for this value of c_0. These facts have been corroborated through numerical experiments.
- *Stability limits for quadratic elements.* They have been derived using the standard and the hierarchic shape functions. Their expressions are collected in Box 2.2.
- *Time step for quadratic elements.* If standard quadratic elements are employed, the value of the security factor $c_0 = 1$ yields dissipative results for a wide range of Fourier modes. The problem encountered with linear elements does not appear here. This compensates the fact that the critical time step is smaller for quadratic elements.
- *Diagonalization of the hierarchic mass matrix.* The possibility of approximating this matrix by a diagonal one has been discussed. A new numerical integration rule has been proposed in order to achieve this diagonal structure. Although the method happens to be useless for time-marching schemes, it can be applied to other situations in which this diagonalization is of interest. When a diagonal mass matrix is chosen, regardless of how does it approximate the exact one, a method for choosing its (positive) diagonal entries has been described.
- *Extension to multidimensional problems.* A criterion to compute the critical time step has been proposed based on heuristic grounds and a previous partial result. This method has proved to work well in practice.

Box 2.2 Stability limits for the forward Euler scheme

Linear elements

General expression	Advective limit	Diffusive limit
$c \le \dfrac{\gamma}{1+\alpha\gamma}$	$c \le 1$	$\Delta t \le \dfrac{h^2}{2k}$

Quadratic elements I : standard basis

General expression	Advective limit	Diffusive limit
$c \le \dfrac{\gamma}{8(1+\alpha\gamma)}$	$c \le \dfrac{1}{8}$	$\Delta t \le \dfrac{h^2}{16k}$

Quadratic elements II : hierarchic approach

General expression	Advective limit	Diffusive limit
$c \le \begin{cases} \min(\mu' c_1, 2\mu c_2) & \text{if } \gamma \le \gamma_c \\ \min(\mu' c_1, 2\mu c_3) & \text{if } \gamma > \gamma_c \end{cases}$	$c \le \min\left(\dfrac{\mu'}{4}, \dfrac{\mu}{8}\right)$	$\Delta t \le \min\left(\dfrac{\mu' h^2}{8k}, \dfrac{\mu h^2}{3k}\right)$

where $\gamma_c \approx 1.23$, μ and μ' are positive constants and

$$c_1 := \gamma \frac{1+\beta\gamma}{\gamma^2 + 4(1+\beta\gamma)^2}$$

$$c_2 := \frac{3\gamma(1+\alpha\gamma)}{4(3\alpha-2\gamma)^2}$$

$$c_3 := \gamma \frac{9(1+\alpha\gamma)^2 - 4\gamma^2}{27(1+\alpha\gamma)^3 + 12(1+\alpha\gamma)[(3\alpha-\gamma)^2 - \gamma^2]}$$

References

[AS] J.H. Argyris and D.W. Scharpf. Finite elements in space and time. *Nucl. Engrg. Des.*, vol. 10 (1969), 456–464

[Ba] C.I. Bajer. Notes on the stability of non-rectangular space-time finite elements. *Int. J. Numer. Meth. Engrg.*, vol. 24 (1987), 1721–1739

[BDH] C. Basdevant, M. Devile, P. Haldenwang, J.M. Lacroix, J. Ouazzani, R. Peyret, P. Orlandi and A.T. Patera. Spectral and finite difference solutions of the Burgers equation. *Comput. & Fluids*, vol. 14 (1986), 23–41

[BH] A.N. Brooks and T.J.R. Hughes. Streamline Upwind/Petrov-Galerkin formulations for convective dominated flows with particular emphasis on the incom-

pressible Navier-Stokes equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 32 (1982), 199–259

[CO] G.F. Carey and J.T. Oden. *Finite Elements: Fluid Mechanics.* The Texas Finite Element Series, vol. VI (Prentice Hall, 1986)

[Co1] R. Codina. *Dues formulacions numèriques per al problema de flux incompressible* (in Catalan). Grade thesis, Universitat Politècnica de Catalunya (1989)

[Co2] R. Codina. Generació de malles estructurades a partir d'equacions el.líptiques. (in Catalan) *CIMNE Report Num. 7* (1990)

[COC] R. Codina, E. Oñate and M. Cervera. The intrinsic time for the SUPG formulation using quadratic elements. *Comput. Meths. Appl. Mech. Engrg.*, vol. 94 (1992), 239–262

[CSS] C. Cuvelier, A. Segal and A. van Steenhoven. *Finite element methods and Navier-Stokes equations.* (Reidel, 1986)

[Do1] J. Donea. Recent advances in computational methods for steady and transient transport problems. SMIRT-7 Conference (1983).

[Do2] J. Donea. A Taylor-Galerkin method for convective transport problems. *Int. J. Numer. Meth. Engrg.*, vol. 20 (1984), 101–119

[Ha] P. Hansbo. *Adaptivity and Streamline Diffusion procedures in the finite element method.* Ph.D. Thesis. Chalmers University of Technology, Göteborg, Sweden (1989)

[HSG] B.M. Herbst, S.W. Schoombie, D.F. Griffthis and A.R. Mitchel. Generalized Petrov-Galerkin methods for the numerical solution of Burgers' equation. *Int. J. Numer. Meth. Engrg.*, vol. 20 (1984), 1273–1289

[HGG] A.C. Hindmarsh, P.M. Gresho and D.F. Griffiths. The stability of explicit Euler time-integration for certain finite-difference approximations of the multidimensional advection-diffusion equations. *Int. J. Numer. Meth. in Fluids*, vol. 4 (1984), 853–897

[Hu] T.J.R. Hughes. *The finite element method. Linear static and dynamic analysis.* (Prentice-Hall, 1987)

[HFH] T.J.R. Hughes, L.P. Franca and G.M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 73, (1989), 173–189

[HH1] T.J.R. Hughes and G.M. Hulbert. Space-time finite element methods for elastodynamics: formulation and error estimates. *Comput. Meths. Appl. Mech. Engrg.*, vol. 66 (1988), 339–363

[HT1] T.J.R. Hughes and T.E. Tezduyar. Finite element methods for first-order hyperbolic systems with particular emphasis on the compressible Euler equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 45 (1984), 217–284

[HT2] T.J.R. Hughes and T.E. Tezduyar. Analysis of some fully discrete algorithms for the one-dimensional heat equation. *Int. J. Numer. Meth. Engrg.*, vol. 21 (1985), 163–168

[HH2] G.M. Hulbert and T.J.R. Hughes. Space-time finite element methods for second order hyperbolic equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 84 (1990), 327–348

[Jo1] C. Johnson. Finite element methods for convection-diffusion problems, in: *Computing methods in applied sciences and engineering*, R. Glowinski and J.L. Lions (eds.) (North-Holland, 1982)

[Jo2] C. Johnson. *Numerical solution of partial differential equations by the finite element method.* (Cambridge University Press, 1986)

[JP] C. Johnson and J. Pitkäranta. An analysis of the discontinous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, vol. 46 (1986), 1–26

[JNP] C. Johnson, U. Nävert and J. Pitkäranta. Finite element methods for linear hyperbolic equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 45 (1984), 285–312

[La] O. Ladyzhenskaya. *The mathematical theory of viscous incompressible flow.* (Gordon-Breach, 1963)

[LeR] P. Lesaint and P.A. Raviart. On a finite element method for solving the neutron transport equation, in: C. de Boor (ed.), *Mathematical aspects of the finite element method.* (Academic Press, 1974)

[LuR] M. Luskin and R. Rannacher. On the smoothing property of the Galerkin method for parabolic equations. *SIAM J. Numer. Anal.*, vol. 19 (1981), 93–113

[MM] T. Meis and U. Marcowitz. *Numerical Solution of Partial Differential Equations.* (Springer-Verlag, 1981).

[MG] A.R. Mitchell and D.F. Griffiths. *The finite difference method in partial differential equations.* (John Wiley, 1980)

[Mo] K.W. Morton. Stability of finite difference approximations to a diffusion- convection equation. *Int. J. Numer. Meth. Engrg.*, vol. 15 (1980), 677–683

[Na] U. Nävert. *A finite element method for convection-diffusion problems.* Ph.D. Thesis. Chalmers University of Technology, Göteborg, Sweden (1982)

[NR] H. Nguyen and J. Reynen. A space-time least-square finite element scheme for advection-diffusion equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 42 (1984), 331–342

[Od] J.T. Oden. A generalized theory of finite elements II. Applications. *Int. J. Numer. Meth. Engrg.*, vol. 1 (1969), 247–259

[Pe] J. Peraire. *A finite element method for convection-dominated flows.* Ph.D. Thesis. University College of Swansea (1986).

[RT] P.A. Raviart and J.M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* (Masson, 1983)

[RG] W.H. Raymond and A. Garder. Selective damping in a Galerkin method for solving wave problems with variable grids. *Monthly Weather Review*, vol. 104 (1976), 1583–1591

[RM] R.D. Richtmyer and K.W. Morton. *Difference methods for initial value problems.* (Interscience, New York, 1967)

[Sh] F. Shakib. *Finite element analysis of the compressible Euler and Navier-Stokes equations.* Ph.D. Thesis. Stanford University (1988).

[SHu] F. Shakib and T.J.R. Hughes. A new finite element formulation for computational fluid dynamics: IX. Fourier Analysis of space-time Galerkin/least-squares algorithms. *Comput. Meths. Appl. Mech. Engrg.*, vol. 87 (1991), 35–58

[SG] J. Siemieniuch and I. Gladwell. Analysis of explicit difference methods for a diffusion-convection equation. *Int. J. Numer. Meth. Engrg.*, vol. 12 (1978), 899–916

[SF] G. Strang and G. Fix. *An analysis of the finite element method.* (Prentice-Hall, 1973)

[TG] T.E. Tezduyar and D.K. Ganjoo. Petrov-Galerkin formulations with weighting

functions dependent upon spatial and temporal dsicretization: application to transient convection-diffusion problems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 59 (1986), 49–71

[Th] V. Thomée. *Galerkin finite element methods for parabolic problems*, Lecture Notes in Mathematics 1054. (Springer, 1984)

[YH1] C.C. Yu and J.C. Heinrich. Petrov-Galerkin methods for the time-dependent convective transport equation. *Int. J. Numer. Meth. Engrg.*, vol. 23 (1986), 883–901

[YH2] C.C. Yu and J.C. Heinrich. Petrov-Galerkin methods for multidimensional time-dependent convective transport equation. *Int. J. Numer. Meth. Engrg.*, vol. 24 (1987), 2201–2215

[ZT] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method*, Fourth Edition, vols. 1 and 2 (McGraw-Hill, 1989)

PART II:

THE INCOMPRESSIBLE NAVIER–STOKES EQUATIONS

CHAPTER 3

PENALTY FINITE ELEMENT METHODS FOR THE STATIONARY NAVIER–STOKES EQUATIONS

3.1 Introduction

The numerical solution of the incompressible Navier-Stokes equations for many practical engineering applications is still far from being a reality. Assuming that the mathematical model represented by these equations is correct, several problems have to be faced, each one represented by different terms of the full system of equations (conservation of momentum and mass balance, i.e., incompressibility condition). Regarding the temporal discretization, the most common way to do it is by far the use of finite difference schemes, with the inconveniences of stability and/or accuracy discussed in the previous chapter. There is no doubt that the most difficult problem arises because of the nonlinear convective term. Loss of unicity of solution, hydrodynamical instabilities and turbulence are caused by this apparently innocent term. These physical phenomena are obviously reflected in the numerical algorithms and in the mathematical analysis of the problem. High Reynolds number flows are a tough problem from the physical, the mathematical and the numerical standpoints.

Another problem to be considered is the incompressibility constraint, closely related to the presence of the pressure forces. The way to overcome this problem will be the subject of this chapter. The mixed velocity-pressure finite element solution of the incompressible Navier-Stokes equations has several inconveniences due to the zero divergence condition for the velocity field. If the standard Galerkin formulation is used, the first problem to be faced is the use of compatible spaces for the velocity and the pressure, in the sense that they have to satisfy the *inf-sup* or Babuška-Brezzi (BB) stability condition [Ba1], [Ba2], [Br] (see also [BB] for a simple derivation of this condition for the discrete problem). A remedy that is gaining popularity is the use of the Galerkin Least Squares approach introduced by Hughes & Franca [HF], [HFB], [FH], [FHL] (see also [BD]) and already described in Chapter 1, especially effective when a continuous interpolation for the pressure is used (otherwise, high order velocity interpolation or non-standard assembly algorithms have to be employed [FS]). Recently, quasi-optimal convergence of this method has been proved by Hansbo & Szepessy [HS] for the time dependent Navier-Stokes equations using space-time linear elements. In any case, the formulation depends on an algorithmic parameter whose physical meaning and optimal values are not known yet. This, and the fact that pressures appear as nodal variables

(see below) may decide the user in favor of the Galerkin formulation, perhaps with an upwind technique for high Reynolds number flows.

If the Galerkin formulation is used, the matrix of the discrete algebraic system resulting from the finite element discretization has zero diagonal terms. The use of iterative solvers or inefficient renumbering algorithms seems to be the only available remedy for solving this system of equations. However, the penalty approach circumvents this problem and has other interesting features. If the pressure interpolation is discontinuous, one can eliminate the element unknowns of this field in terms of the velocity nodal unknowns. Substitution of the obtained expression in the momentum equation leads to a system whose only degrees of freedom are velocities. The reduction of the number of nodal unknowns and the fact that the method is known to work well, have made the penalty method very popular, especially in the engineering literature (see, e.g., References [CK1], [CK2], [ESG], [HLB], [Od], [OKS]). Perhaps the only drawback of this approach is the ill-conditioning of the stiffness matrix when the penalty parameter is very small. A lower bound is determined basically by the computer and the arithmetic precision used in the calculation.

There are basicaly two ways to penalize the incompressibility constraint: to start with the penalized differential equation or to wait until the weak form has been established. Both approaches do coincide for the continuous problem. However, when the finite element discretization is carried out the former automatically yields the pressure space as a consequence of the choice of the velocity interpolation. The resulting velocity-pressure pair will be in general unstable, in the sense that the BB condition will be violated. Hereafter, we will refer to this approach as *strong penalization*, whereas the use of the penalty method for the weak incompressibility equation (continuous or discrete) will be referred to as *weak penalization*. A discussion of the way a particular finite element can be implemented using strong penalization is the subject of Section 3.3.

The objective of Section 3.4 is to present and analyse an iterative weak penalty finite element method for the stationary Stokes and Navier-Stokes equations. The goal is the convergence of the iterates to the true incompressible solution. The main advantage of this approach is that larger penalty parameters can be used, thus alleviating the ill-conditioning mentioned above. The basic idea is solving the penalized equations in each iteration but adding a right-hand-side term that is basically the residual of the incompressibility equation of the previous iterate. For the Stokes equations, this approach is the only reason for an iterative scheme to be used and the conditions under which convergence is achieved are only determined by the iterative penalization. However, the Navier-Stokes equations must be solved iteratively. The question that naturally arises is whether the iterative scheme employed can be coupled with the iterative penalization or not. We prove that, under not very restrictive conditions, the answer is yes. The exposition of this section is organized as follows. The Stokes problem is considered in Section 3.4.1, where the idea of the iterative penalization is described in detail. Section 3.4.2 deals with the Navier-Stokes equations when the Picard (or successive substitution) algorithm is used for the nonlinear term and Section 3.4.3 when the Newton-Raphson scheme is employed. The uncoupling of the nonlinear and penalization iterative loops is then studied.

In this chapter the choice of the finite element spaces will not be the main interest and it is postponed until Chapter 4. Only when necessary we will refer to the stability of a certain finite element under consideration. As usual, whenever a certain element satifies the BB stability condition for the restriction arising form the zero divergence

constraint, we will call it *div-stable* [BN1]. The possibility of by-passing this requirement by using an exactly divergence free finite element basis for the velocity field [GP], [RT] will not be considered in this work. The main reason is that although this formulation seems to be feasible for two-dimensional flows, the construction of velocity bases in three-dimensional problems happens to be complicated [Ne1], [Ne2]. The use of formulations other than the conforming velocity-pressure will not be treated either. Background for these methods and for the techniques to be used in this chapter can be found in References [BFo], [CO1], [CO2], [CSS], [GR], [Gu], [OC], [Pi], among other standard text books.

3.2 Statement of the problem and penalty methods

3.2.1 The continuous problem

Let Ω be an open bounded domain of $\mathbb{R}^{N_{sd}}$ ($N_{sd} = 2$ or 3) and $\Gamma = \partial\Omega$ its boundary, assumed to be locally Lipschitz. The Navier-Stokes problem for an incompressible fluid moving in Ω with, for simplicity, homogeneous boundary conditions, consists in finding a velocity field $\mathbf{u}$ and a pressure p such that

$$\begin{aligned} \rho(\mathbf{u}\cdot\nabla)\mathbf{u} - \mu\Delta\mathbf{u} + \nabla p &= \rho\mathbf{f} \quad && \text{in } \Omega \\ \nabla\cdot\mathbf{u} &= 0 && \text{in } \Omega \\ \mathbf{u} &= \mathbf{0} && \text{on } \Gamma \end{aligned} \tag{3.1}$$

where $\mathbf{f}$ is a given body force, ρ is the density of the fluid and μ is its dynamical viscosity. In order to write the weak form of problem (3.1) we introduce the spaces

$$V = H_0^1(\Omega)^{N_{sd}}, \qquad Q = L^2(\Omega) \tag{3.2}$$

and the multilinear forms

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &= \mu\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega, \\ b(q,\mathbf{v}) &= \int_\Omega q\nabla\cdot\mathbf{v}\, d\Omega, \\ c(\mathbf{u},\mathbf{v},\mathbf{w}) &= \rho\int_\Omega [(\mathbf{u}\cdot\nabla)\mathbf{v}]\cdot\mathbf{w}\, d\Omega, \\ l(\mathbf{v}) &= \rho < \mathbf{f},\mathbf{v} > \end{aligned} \tag{3.3}$$

defined on $V\times V, Q\times V, V\times V\times V$ and V respectively. The symbol $<\cdot,\cdot>$ denotes the duality paring between V and its topological dual V' ($= H^{-1}(\Omega)^{N_{sd}}$). If the viscous term in (3.1) is written as $-\nabla\cdot(2\mu\varepsilon(\mathbf{u}))$, where $\varepsilon(\mathbf{u})$ is the symmetric part of $\nabla\mathbf{u}$, the bilinear form a to be considered is

$$a(\mathbf{u},\mathbf{v}) = 2\mu\int_\Omega \varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v})\, d\Omega \tag{3.4}$$

instead of that appearing in (3.3). Continuity of a, b and l is obvious. Continuity of c follows from Sobolev's imbedding Theorem (if $\mathbf{u}$ and $\mathbf{v} \in V$ then $\mathbf{u}$ and $\mathbf{v} \in L^4(\Omega)^{N_{sd}}$

for $N_{sd} = 2, 3$) and from Hölder's inequality (if $u_j, v_k \in L^4(\Omega)$, then $u_j v_k \in L^2(\Omega)$, u_j and v_k being the components of $\mathbf{u}$ and $\mathbf{v}$). See, e.g. References [GR], [Te] for details. Since a, c and l are continuous, we can define their 'norms' by

$$\begin{aligned} N_a &= \sup \frac{a(\mathbf{v}_1, \mathbf{v}_2)}{\|\mathbf{v}_1\|_V \|\mathbf{v}_2\|_V} \\ N_c &= \sup \frac{c(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)}{\|\mathbf{v}_1\|_V \|\mathbf{v}_2\|_V \|\mathbf{v}_3\|_V} \\ N_l &= \sup \frac{l(\mathbf{v}_1)}{\|\mathbf{v}_1\|_V} \end{aligned} \tag{3.5}$$

where the supremum is taken over all the $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V - \{\mathbf{0}\}$ and $\|\cdot\|_V$ denotes the usual norm in V. We will use the symbol $\|\cdot\|_Q$ for the norm in Q and $(\cdot, \cdot)$ for the inner product in the space Q. When no ambiguity be possible, we shall omit the subscripts in the norms.

Define the space

$$Z = \{q \in Q \mid b(q, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in V\} \tag{3.6}$$

For b given by (3.3), $Z = \mathbb{R}$. In the quocient space Q/Z the following norm is defined

$$\|q\|_{Q/Z} = \inf_{z \in Z} \|q + z\|_Q \tag{3.7}$$

Having introduced all this notation, the weak form of problem (3.1) can be written as follows: Find $\mathbf{u} \in V$ and $p \in Q/Z$ such that

$$\begin{aligned} c(\mathbf{u}, \mathbf{u}, \mathbf{v}) + a(\mathbf{u}, \mathbf{v}) - b(p, \mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ b(q, \mathbf{u}) &= 0 \qquad \forall q \in Q \end{aligned} \tag{3.8}$$

Besides the continuity of all the forms involved in (3.8), we will assume that the bilinear form a is coercive and that b satisfies the BB condition, i.e., there exist positive constants K_a and K_b such that

$$a(\mathbf{v}, \mathbf{v}) \geq K_a \|\mathbf{v}\|_V^2 \qquad \forall \mathbf{v} \in V \tag{3.9}$$

$$\sup_{\mathbf{v}} \frac{b(q, \mathbf{v})}{\|\mathbf{v}\|_V} \geq K_b \|q\|_{Q/Z} \qquad \mathbf{v} \in V - \{\mathbf{0}\},\ \forall q \in Q \tag{3.10}$$

Condition (3.9) follows from Poincaré-Friedrics inequality if a is given by (3.3) and from Korn's inequality if it is given by (3.4). Condition (3.10) holds for V and Q given by (3.2) [La].

For the trilinear form c it will be assumed that

$$c(\mathbf{u}, \mathbf{v}, \mathbf{v}) = 0 \qquad \forall \mathbf{v} \in V \tag{3.11}$$

If $\mathbf{u}$ is the solution of problem (3.8), it is easy to see that condition (3.11) is satisfied. However, we will be interested in velocity fields that do not exactly satisfy the incompressibility condition. In this case, instead of the form c given in (3.3), its skew-symmetrized form will be used:

$$c_\sigma(\mathbf{u}, \mathbf{v}, \mathbf{w}) = c(\mathbf{u}, \mathbf{v}, \mathbf{w}) + \frac{1}{2}\rho \int_\Omega (\nabla \cdot \mathbf{u})\mathbf{v} \cdot \mathbf{w} d\Omega$$

It can be easily checked that $c_\sigma(\mathbf{u},\mathbf{v},\mathbf{v}) = 0$ and that $c_\sigma(\mathbf{u},\mathbf{v},\mathbf{w}) = c(\mathbf{u},\mathbf{v},\mathbf{w})$ if $\mathbf{u}$ is the solution of problem (3.8). Continuity of c_σ can be proved as for c. Thus, c_σ can be used instead of c in (3.8) and condition (3.11) will hold. In any case subscript σ will be omitted.

Finally, we will assume that

$$\chi := \frac{N_c N_l}{K_a^2} < 1 \tag{3.12}$$

Under all these conditions, existence and uniqueness of solution of (3.8) can be proved [GR].

In what follows, the spaces V and Q will be those given by (3.2) or finite dimensional subspaces V_h and Q_h arising from the finite element discretization of Ω (internal approximation). Conditions (3.9), (3.11) and (3.12) will be automatically satisfied if V and Q are replaced by V_h and Q_h. However, (3.10) has to be explicitly required for each pair of finite element spaces V_h, Q_h.

The condition $\chi < 1$ is certainly restrictive. It ensures uniqueness of weak solutions. However, such unicity is not likely to hold for high Reynolds numbers and in fact examples are known for which it is not true [Te]. In these cases, a more careful analysis has to be done. Fortunately, solutions of the Navier-Stokes happen to be isolated in most of the cases. For a detailed analysis of the approximation of this type of problems, the reader is referred to [BR1-3]. See also [GR].

Concerning the no-slip boundary condition in problem (3.1), the extension to the non-homogeneous condition $\mathbf{u} = \mathbf{g}$ on Γ is straightforward and only requires some technicalities [GR]. However, the situation is somehow more involved when the traction is prescribed on a part of the boundary. An analysis of the Galerkin finite element approximation in this case can be found in [Ve2].

3.2.2 Penalty methods for the Stokes problem

In this chapter, we will deal with a class of penalty methods that will be briefly described here. In order to introduce the problem, it is enough to consider the Stokes equations, i.e., Eqns. (3.1) without the convective term $\rho(\mathbf{u}\cdot\nabla)\mathbf{u}$ or its weak form (3.8) with $c = 0$. In this case, Eqns. (3.1) are the optimality conditions for the minimization problem of finding $\mathbf{u} \in V$ such that $\nabla \cdot \mathbf{u} = 0$ and

$$\mathcal{L}_0(\mathbf{u}) = \inf_{\mathbf{v}} \mathcal{L}_0(\mathbf{v}) \qquad \text{over} \quad \{\mathbf{v} \in V \mid \nabla \cdot \mathbf{v} = 0\} \tag{3.13}$$

$$\mathcal{L}_0(\mathbf{v}) := \frac{1}{2}a(\mathbf{v},\mathbf{v}) - l(\mathbf{v}) \tag{3.14}$$

Introducing the Lagrangian multiplier p to account for the constraint $\nabla \cdot \mathbf{u} = 0$ we are led to the following saddle point problem: Find $\mathbf{u} \in V$ and $p \in Q/Z$ such that

$$\mathcal{L}_p(\mathbf{u},p) = \inf_{\mathbf{v}\in V} \sup_{q\in Q} \mathcal{L}_p(\mathbf{v},q) \tag{3.15}$$

$$\mathcal{L}_p(\mathbf{v},q) := \frac{1}{2}a(\mathbf{v},\mathbf{v}) - l(\mathbf{v}) - b(q,\mathbf{v}) \tag{3.16}$$

Now we can consider standard techniques from optimization theory. The first is the penalty method, consisting in adding to the functional $\mathcal{L}_0$ defined by (3.14) a

positive definite bilinear form evaluated on the constraint and multiplied by a large number. If, for example, we consider the L^2 inner product as the bilinear form and pick a small number $\epsilon > 0$, we are led to the penalized problem: Find $\mathbf{u}^\epsilon \in V$ such that

$$\mathcal{L}_{0,\epsilon}(\mathbf{u}^\epsilon) = \inf_{\mathbf{v}\in V} \mathcal{L}_{0,\epsilon}(\mathbf{v}) \tag{3.17}$$

$$\mathcal{L}_{0,\epsilon} := \frac{1}{2}a(\mathbf{v},\mathbf{v}) - l(\mathbf{v}) + \frac{1}{2\epsilon}(\nabla\cdot\mathbf{v}, \nabla\cdot\mathbf{v}) \tag{3.18}$$

Finally, we could also perturb the functional $\mathcal{L}_p$ given by (3.16) in order to obtain the regularized problem: Find $\mathbf{u}^\epsilon \in V$ and $p^\epsilon \in Q_0$ such that

$$\mathcal{L}_{p,\epsilon}(\mathbf{u}^\epsilon, p^\epsilon) = \inf_{\mathbf{v}\in V}\sup_{q\in Q} \mathcal{L}_{p,\epsilon}(\mathbf{v}, q) \tag{3.19}$$

$$\mathcal{L}_{p,\epsilon}(\mathbf{v}, q) := \frac{1}{2}a(\mathbf{v},\mathbf{v}) - l(\mathbf{v}) - b(q,\mathbf{v}) - \frac{1}{2}\epsilon(q,q) \tag{3.20}$$

where we have introduced the space

$$Q_0 := \{q \in Q \mid \int_\Omega q \, d\Omega = 0\} \tag{3.21}$$

that is isomorphic to Q/Z for b given in (3.3). The reason for this choice of the pressure space will be clear immediately.

The Euler-Lagrange equations for the variational problem (3.17) are

$$a(\mathbf{u}^\epsilon,\mathbf{v}) + \frac{1}{\epsilon}(\nabla\cdot\mathbf{u}^\epsilon, \nabla\cdot\mathbf{v}) = l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \tag{3.22}$$

whereas for problem (3.19) the optimality conditions are

$$a(\mathbf{u}^\epsilon,\mathbf{v}) - b(p^\epsilon,\mathbf{v}) = l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \tag{3.23}$$

$$\epsilon(p^\epsilon, q) + b(q,\mathbf{u}^\epsilon) = 0 \qquad \forall q \in Q \tag{3.24}$$

If we take $q \equiv const.$ in (3.24) we see that

$$\epsilon\int_\Omega p^\epsilon \, d\Omega + \int_\Gamma \mathbf{n}\cdot\mathbf{u}^\epsilon \, d\Gamma = 0,$$

where $\mathbf{n}$ is the unit outward normal to Γ. Since $\mathbf{u}^\epsilon = 0$ on Γ, it follows that $\int_\Omega p^\epsilon \, d\Omega = 0$, i.e., Q_0 is the right space where p^ϵ is to be sought. For non-homogeneous boundary conditions $\mathbf{u} = \mathbf{g}$ on Γ, the given function $\mathbf{g}$ must satisfy the compatibility condition $\int_\Gamma \mathbf{n}\cdot\mathbf{g} \, d\Gamma = 0$ and hence $\int_\Omega p^\epsilon \, d\Omega = 0$ still holds true. Pressures that are solution of (3.1) or (3.8) are determined up to an additive constant that can be fixed seeking p either in Q_0 or in Q/Z. From now onwards, the former choice will be employed since penalized solutions automatically belong to Q_0.

It is clear that Eqn. (3.24) implies

$$\epsilon p^\epsilon + \nabla\cdot\mathbf{u}^\epsilon = 0 \qquad \text{in the space } Q \tag{3.25}$$

Since for $\mathbf{u}^\epsilon \in V = H_0^1(\Omega)^{N_{sd}}$ it is $\nabla\cdot\mathbf{u}^\epsilon \in Q = L^2(\Omega)$, Eqn. (3.25) can be understood in the classical sense. Inserting the pressure p^ϵ obtained from (3.25) in terms of $\mathbf{u}^\epsilon$ into Eqn. (3.23) we recover problem (3.22). Therefore, we can state the following important

fact: *for the continuous case, the penalized and the perturbed variational problems are equivalent.* The crucial point is to observe that the divergence of the velocity field belongs to the pressure space. This *will not* be true for the discrete problem and the equivalence just mentioned will cease to be valid.

The reason why we have introduced first the Stokes problem is to highlight the connexion between classical optimization techniques and the penalty methods we will consider. The Navier-Stokes equations *are not* the Euler-Lagrange equations of a functional to be minimized but nevertheless we can still consider the analogues of (3.22) and (3.23)–(3.24) with $c \neq 0$. The former problem will be the weak form of the partial differential equation

$$\rho(\mathbf{u}^\epsilon \cdot \nabla)\mathbf{u}^\epsilon - \mu\Delta\mathbf{u}^\epsilon - \frac{1}{\epsilon}\nabla(\nabla \cdot \mathbf{u}^\epsilon) = \rho\mathbf{f} \tag{3.26}$$

that is obtained by replacing the pressure p in (3.1) by the expression found from the pseudo-constitutive relation

$$p^\epsilon = -\frac{1}{\epsilon}\nabla \cdot \mathbf{u}^\epsilon. \tag{3.27}$$

If the linear Elasticity problem is considered, $1/\epsilon$ has the physical meaning of being the Lamé parameter of a slightly compressible material. For a discussion about the physical meaning of several penalty methods, see Reference [HD].

Since in (3.26) the *differential equation* has been penalized, we will call this approach *strong penalization.* On the other hand, it is observed from (3.24) that the *weak form* of the incompressibility constraint has been replaced by a penalized equation. This method will be referred to as *weak penalization.* It must be stressed that both approaches are equivalent for the continuous case and that for the Stokes problem the strong penalization corresponds to the minimization of the classical penalized functional (3.18) and the weak penalization comes from seeking the saddle point of the perturbed Lagrangian (3.20).

Whichever of the two approaches just described is used, the key for proving the convergence of $\mathbf{u}^\epsilon$ and of p^ϵ to the solution $\mathbf{u}$ and p of problem (3.8) as $\epsilon \to 0$ is the BB condition (3.10) (see, e.g. [Be], [BFo], [CSS], [GR], [OKS]), that in turn happens to be the key condition, together with the coercivity of a (condition (3.9)), for proving existence and uniqueness of solution for this problem (3.8).

3.2.3 Finite element discretization

The discrete problem

Let $\{\Omega^e\}$ be a regular finite element partition of the domain Ω, with index e ranging from 1 to the number of elements N_{el}. For simplicity, we shall assume that Ω is a polyhedral domain. The diameter of $\{\Omega^e\}$ will be denoted by h, as usual.

Let now $V_h \subset V$ and $Q_h \subset Q$ be conforming finite element spaces associated to the partition $\{\Omega^e\}$. Define also

$$Z_h := \{q_h \in Q_h \mid b(q_h, \mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in V_h\} \tag{3.28}$$

$$Q_{0,h} := \{q_h \in Q_h \mid \int_\Omega q_h \, d\Omega = 0\} \tag{3.29}$$

Let us consider first the Stokes problem. The discrete version of (3.8) with $c = 0$ is: Find $\mathbf{u}_h \in V_h$ and $p_h \in Q_h/Z_h$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) - b(p_h, \mathbf{v}_h) &= l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \\ b(q_h, \mathbf{u}_h) &= 0 \qquad \forall q_h \in Q_h \end{aligned} \tag{3.30}$$

The weak penalty method applied to (3.30) will read: Find $\mathbf{u}_h^\epsilon \in V_h$ and $p_h^\epsilon \in Q_{0,h}$ such that

$$\begin{aligned} a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) - b(p_h^\epsilon, \mathbf{v}_h) &= l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \\ \epsilon(p_h^\epsilon, q_h) + b(q_h, \mathbf{u}_h^\epsilon) &= 0 \qquad \forall q_h \in Q_h \end{aligned} \tag{3.31}$$

Convergence in norm of $\mathbf{u}_h^\epsilon$ to $\mathbf{u}_h$ and of p_h^ϵ to p_h as $\epsilon \to 0$ is a well known result [CSS], [GR], [OKS].

The first question to be answered is whether the space $Q_{0,h}$ is isomorphic to Q_h/Z_h or not. Clearly, this requires that $\dim Z_h = 1$, which is not always the case and depends on the choice of the spaces V_h and Q_h. In Reference [JP] it is proved that this condition is also sufficient to assert that $Q_{0,h} \cong Q_h/Z_h$.

Some finite element spaces

There are several popular elements for which it is known that $\dim Z_h > 1$. Loosely speaking, this means that there are pressures whose discrete gradient is zero, appart from the constants. The most well known element that exhibits this pathology is the Q_1/P_0 pair, constructed using a bilinear continuous velocity interpolation (in 2D) and piecewise constant pressures. For this element it is known that when Ω is a square (in 2D) discretized with an even number of uniform elements along each direction, $\dim Z_h = 2$, the space Z_h consisting of constants and the so called 'checkboard mode'. There is a vast literature on this controversial element, apparently first used by Hughes & Allik [HA] (cf. [BFo]). See, for example, References [BN2], [CK1], [JP], [OKS], [SG1–2], among many others. Another problem encountered when this element is used is the satisfaction of the BB condition. It is known that the discrete analogue of (3.10), namely,

$$\sup_{\mathbf{v}_h} \frac{b(q_h, \mathbf{v}_h)}{\|\mathbf{v}_h\|_V} \geq K_b \|q_h\|_{Q_h/Z_h} \qquad \mathbf{v}_h \in V_h - \{\mathbf{0}\},\ \forall q_h \in Q_h \tag{3.32}$$

is satisfied but with $K_b = O(h)$, h being the diameter of the uniform mesh [CK1], [OJ1], [JP]. It is believed that this element yields stable velocity approximations on general distorted meshes, but this fact still resists analysis. Moreover, this element can be stabilized either by using macroelements composed of this element [TR], by redefining the pressure space [KOS], [OKS] or by using iterative stabilization techniques [FB]. The reader is referred to the book of Brezzi & Fortin [BFo] (Section VI.5.4) for a clarifying discussion.

Another popular element is the Q_2/Q_1 pair (continuous biquadratic velocities, discontinuous piecewise bilinear pressures, in 2D). Again, it is found that $\dim Z_h = 2$ and that the constant K_b in (3.32) is proportional to h [CK1], [OJ1].

Perhaps the quadrilateral two-dimensional element that enjoys most popularity at the present time is the Q_2/P_1 pair (continuous biquadratic velocities, discontinuous piecewise linear pressures), first proposed in Reference [NPR] and known to yield very good results for incompressible flow problems [Fo2], [FF]. This element satisfies the BB condition with K_b independent of h and $\dim Z_h = 1$, showing that no spurious pressure modes are possible. See, e.g. [GR] for a rigorous analysis of its stability. A related

element is the Q_2^-/P_1, where now a serendipid interpolation for the velocity is used. Unfortunately, this element happens to be unstable, with $\dim Z_h = 3$ and $K_b = O(h)$ [OJ1], [OJ2]. A div-stable element is obtained if the pressure is taken as piecewise constant (Q_2^-/P_0 element) [GR]. Box 3.1 summarizes the properties of the elements discussed so far (recall that $\dim Z_h$ refers to the case of a square discretized using a uniform mesh). See Chapter 4 for a schematic of the elements.

Box 3.1 Stability of some quadrilateral elements

Element	$\dim Z_h$	K_b
Q_1/P_0	2	$O(h)$
Q_2^-/P_0	1	$O(1)$
Q_2^-/P_1	3	$O(h)$
Q_2/P_0	1	$O(1)$
Q_2/P_1	1	$O(1)$
Q_2/Q_1	2	$O(h)$

Concerning the rate of convergence when the BB condition is satisfied, it can be expressed in the norm of $V \times Q$, i.e.,

$$\|(\mathbf{u} - \mathbf{u}_h, p - p_h)\|_{V\times Q} := \|\mathbf{u} - \mathbf{u}_h\|_V + \|p - p_h\|_Q \tag{3.33}$$

where $(\mathbf{u}, p)$ is the solution of the continuous problem and $(\mathbf{u}_h, p_h)$ the solution of (3.30). From Brezzi's result [Br], an estimate for (3.33) reduces to an estimate for the interpolation error, since

$$\|(\mathbf{u} - \mathbf{u}_h, p - p_h)\|_{V\times Q} \leq C \left(\inf_{\mathbf{v}_h \in V_h} \|\mathbf{u} - \mathbf{v}_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right) \tag{3.34}$$

Once this is found, an estimate for the velocity in the L^2 norm is easily obtained through the classical Aubin-Nitsche duality argument [Br], [BFo], [Ci], [OC]. The elements Q_2^-/P_0 and Q_2/P_0 yield a suboptimal rate of convergence for the velocity in the L^2 norm, that is $O(h^2)$. This is due to the poor pressure approximation, which controls the error in the $V \times Q$ norm. On the other hand, the Q_2/P_1 yields optimal rates of convergence, namely, $\|\mathbf{u} - \mathbf{u}_h\|_0 = O(h^3)$ and $\|p - p_h\|_0 = O(h^2)$.

We postpone until Chapter 4 a more detailed discussion on available finite element interpolations and convergence properties.

Strong penalization and RIP methods

When dealing with the weakly penalized problem (3.31) it is clear that the velocity and pressure spaces have to be chosen independently and they have to satisfy the BB condition. Let us consider now the discrete version of problem (3.22): Find $\mathbf{u}_h^\epsilon \in V_h$ such that

$$a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) + \frac{1}{\epsilon}(\nabla \cdot \mathbf{u}_h^\epsilon, \nabla \cdot \mathbf{v}_h) = l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \tag{3.35}$$

Implicitly, relation (3.27) has been assumed. Since only velocities appear in (3.35), the pressure space is not explicitly defined in this approach. In fact, it is determined by (3.27). Symbolically, we can write $Q_h = \mathrm{div} V_h$, meaning that pressures are the divergence of vector fields belonging to the velocity space V_h. The resulting $V_h - Q_h$ pair will be in general unstable.

The way to (partially) overcome this problem was first devised in the pioneering works of Zienkiewicz *et al.* [ZTT] in the context of plate and shell bending theory for the Reissner-Mindlin formulation and Fried [Fr] in incompressible Elasticity. In its actual form, it consists in integrating the volumetric term $(\nabla \cdot \mathbf{u}_h^\epsilon, \nabla \cdot \mathbf{v}_h)$ using a low order quadrature rule. The method was first called *selective underintegration* (or reduced underintegration, if this rule is used to compute all the terms) and was lately popularized by Oden [Od] under the acronym *RIP* (Reduced Integration Penalty) method.

To see the connexion between (3.35) and (3.31), let ξ_j^e, $j = 1, ..., N_{qp}$ be the quadrature points placed within element e, $e = 1, ..., N_{el}$, to calculate $(\nabla \cdot \mathbf{u}_h^\epsilon, \nabla \cdot \mathbf{v}_h)$ and let $\omega_j^e > 0$ be the corresponding weights. Denote by $(\cdot, \cdot)_*$ the approximate L^2 inner product associated to this rule, i.e.,

$$(f_1, f_2)_* := \sum_{e=1}^{N_{el}} \sum_{j=1}^{N_{qp}} \omega_j^e f_1(\xi_j^e) f_2(\xi_j^e) \tag{3.36}$$

Instead of (3.35) we will consider:

$$a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) + \frac{1}{\epsilon}(\nabla \cdot \mathbf{u}_h^\epsilon, \nabla \cdot \mathbf{v}_h)_* = l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \tag{3.37}$$

Now, let Q_h be the pressure space defined by discontinuous piecewise polynomials N_{qp}–unisolvent (for example, in 2D, constants if $N_{qp} = 1$, linear if $N_{qp} = 3$, bilinear if $N_{qp} = 4$ and so on) and consider also the problem

$$a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) - (p_h^\epsilon, \nabla \cdot \mathbf{v}_h)_* = l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \tag{3.38}$$

$$\epsilon(p_h^\epsilon, q_h)_* + (q_h, \nabla \cdot \mathbf{u}_h^\epsilon)_* = 0 \qquad \forall q_h \in Q_h \tag{3.39}$$

From Eqn. (3.39) we have that

$$0 = (\epsilon p_h^\epsilon + \nabla \cdot \mathbf{u}_h^\epsilon, q_h)_* = \sum_{e=1}^{N_{el}} \sum_{j=1}^{N_{qp}} \omega_j^e \left[\epsilon p_h^\epsilon(\xi_j^e) + \nabla \cdot \mathbf{u}_h^\epsilon(\xi_j^e)\right] q_h(\xi_j^e)$$

for all $q_h \in Q_h$, from where it follows that

$$p_h^\epsilon(\xi_j^e) = -\frac{1}{\epsilon} \nabla \cdot \mathbf{u}_h^\epsilon(\xi_j^e) \tag{3.40}$$

for $j = 1, ..., N_{qp}$, $e = 1, ..., N_{el}$. Inserting this expression in Eqn. (3.38) we obtain (3.37). Therefore, we have proved that *problems (3.37) and (3.38)-(3.39) are equivalent*. As a trivial consequence, the following result follows from the comparison of (3.31) and (3.38)–(3.39): *If the numerical quadrature rule is such that*

$$(q_h, q_h')_* = (q_h, q_h') \qquad \text{and} \qquad (q_h, \nabla \cdot \mathbf{v}_h)_* = (q_h, \nabla \cdot \mathbf{v}_h) \tag{3.41}$$

for all q_h, $q_h' \in Q_h$ and $\mathbf{v}_h \in V_h$, then the weak penalty method (3.31) and the RIP method (3.37) are equivalent.

This equivalence theorem was first established by Malkus & Hughes [MH], although the proof presented here is closer to the one given by Oden [Od]. See also [ESG] and [Fo3] for an interesting discussion.

The definition of the pressure space Q_h is obvious once the numerical quadrature rule has been chosen. Of course the real problem is to check whether condition (3.41) is satisfied or not. For the Q_1/P_0 element, the equivalence happens to be exact for general distorted meshes [BFo], [JP], since it is easy to see that if a one-point quadrature rule is used to evaluate the volumetric term, condition (3.41) holds true. In Section 3.3 we will discuss what happens for quadratic quadrilateral elements.

Matrix formulation

The penalty methods discussed so far can be applied to the Navier-Stokes equations as well. Corresponding to problems (3.31) and (3.37) we will have, respectively,

Method 1 (weak penalization): *Find* $\mathbf{u}_h^\epsilon \in V_h$ *and* $p_h^\epsilon \in Q_{0,h}$ *such that*

$$\begin{aligned} c(\mathbf{u}_h^\epsilon, \mathbf{u}_h^\epsilon, \mathbf{v}_h) + a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) - b(p_h^\epsilon, \mathbf{v}_h) &= l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \\ \epsilon(p_h^\epsilon, q_h) + b(q_h, \mathbf{u}_h^\epsilon) &= 0 \qquad \forall q_h \in Q_h \end{aligned} \tag{3.43}$$

Method 2 (strong penalization): *Find* $\mathbf{u}_h^\epsilon \in V_h$ *such that*

$$c(\mathbf{u}_h^\epsilon, \mathbf{u}_h^\epsilon, \mathbf{v}_h) + a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) + \frac{1}{\epsilon} d(\mathbf{u}_h^\epsilon, \mathbf{v}_h) = l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \tag{3.44}$$

where the notation

$$d(\mathbf{u}_h, \mathbf{v}_h) := (\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h)_* \tag{3.45}$$

has been introduced. Define now the following integers associated to the finite element mesh:

$$\begin{aligned} &N_{fp} && : \text{ total number of free nodes,} \\ &N_{vu} := N_{fp} \times N_{sd} && : \text{ total number of velocity unknowns,} \\ &N_{pu} := N_{qp} \times N_{el} && : \text{ total number of pressure unknowns.} \end{aligned}$$

Each term in Eqns. (3.43) and (3.44) will yield a matrix or a vector once the finite element basis (shape functions) has been chosen. Let $\mathbf{U}$ and $\mathbf{P}$ be the vectors whose components are the velocity and pressure nodal unknowns, respectively, and denote by $\mathbf{U}^\epsilon$ and $\mathbf{P}^\epsilon$ their 'penalized' counterparts. The matrices that will appear in the final algebraic system of equations and the term from where they come are given in Box 3.2. The vector $\mathbf{F}_p$ accounts for possible non-homegeneous velocity boundary conditions.

Having introduced all these arrays, the matrix version of problems (3.43) and (3.44) will be

Method 1 (weak penalization): *Solve the nonlinear algebraic system*

$$\begin{aligned} \mathbf{K}_c(\mathbf{U}^\epsilon)\,\mathbf{U}^\epsilon + \mathbf{K}_d\mathbf{U}^\epsilon - \mathbf{G}\mathbf{P}^\epsilon &= \mathbf{F}_u \\ \mathbf{G}^T\mathbf{U}^\epsilon + \epsilon\mathbf{M}_p\mathbf{P}^\epsilon &= \mathbf{F}_p \end{aligned} \tag{3.46}$$

Method 2 (strong penalization): *Solve the nonlinear algebraic system*

$$\mathbf{K}_c(\mathbf{U}^\epsilon)\,\mathbf{U}^\epsilon + \mathbf{K}_d\mathbf{U}^\epsilon + \frac{1}{\epsilon}\mathbf{K}_v\mathbf{U}^\epsilon = \mathbf{F}_u \tag{3.47}$$

Although expressions (3.46) will be kept in what follows, their implementation uses the fact that the pressure interpolation is discontinuous. This allows to eliminate the pressure degrees of freedom, thus making the method much more efficient from the computational standpoint. For discontinuous pressures, the second equation in (3.46) holds for each element. If we denote by superscript e the element arrays, before imposing the boundary conditions we will have that

$$\mathbf{G}^{(e)^T}\mathbf{U}^{(e)^\epsilon} + \epsilon \mathbf{M}_p^{(e)}\mathbf{P}^{(e)^\epsilon} = \mathbf{0} \tag{3.48}$$

and hence

$$\mathbf{P}^{(e)^\epsilon} = -\frac{1}{\epsilon}\mathbf{M}_p^{(e)^{-1}}\mathbf{G}^{(e)^T}\mathbf{U}^{(e)^\epsilon} \tag{3.49}$$

Let $\mathcal{A}$ denote the standard finite element assembly operator. From (3.48) and (3.49) we have that:

$$\left[\mathbf{K}_c(\mathbf{U}^\epsilon) + \mathbf{K}_d + \frac{1}{\epsilon}\mathcal{A}_{e=1}^{N_{el}}\left(\mathbf{G}^{(e)}\mathbf{M}_p^{(e)^{-1}}\mathbf{G}^{(e)^T}\right)\right]\mathbf{U}^\epsilon = \mathbf{F}_u \tag{3.50}$$

Once $\mathbf{U}^{\epsilon(i)}$ is found by solving Eqn. (3.50), the pressure nodal values can be computed for each element from the expression (3.49). It should be remarked that the matrix of system (3.50) has to be factored only once for Newtonian Stokesian flows and that the inversion of $\mathbf{M}_p^{(e)}$ is trivial (it is a $(N_{sd}+1)\times(N_{sd}+1)$ matrix if linear pressures are used). Due to this simplification when the pressure is discontinuous, the weak penalty method is rarely applied when the pressure space consists of continuous functions.

Box 3.2 Matrices and vectors for the algebraic system

Matrix	comes from	and has dimensions
$\mathbf{K}_c(\mathbf{U})$	$c(\mathbf{u},\cdot,\cdot)$	$N_{vu}\times N_{vu}$
$\mathbf{K}_d$	$a(\cdot,\cdot)$	$N_{vu}\times N_{vu}$
$\mathbf{K}_v$	$d(\cdot,\cdot)$	$N_{vu}\times N_{vu}$
$\mathbf{G}$	$b(\cdot,\cdot)$	$N_{vu}\times N_{pu}$
$\mathbf{M}_p$	$(\cdot,\cdot)$	$N_{pu}\times N_{pu}$
$\mathbf{F}_u$	$l(\cdot)$	N_{vu}
$\mathbf{F}_p$	Bound. cond.	N_{pu}

Remark 3.1

From the preceeding discussion and comparing expressions (3.47) and (3.50), it is clear that the equality

$$\mathcal{A}_{e=1}^{N_{el}}\left(\mathbf{G}^{(e)}\mathbf{M}_p^{(e)^{-1}}\mathbf{G}^{(e)^T}\right) = \mathbf{K}_v \tag{3.51}$$

will hold if the quadrature rule $(\cdot,\cdot)_*$ verifies conditions (3.41). Checking this matrix relation for the particular case of the Q_1/P_0 element was the first step towards the understanding of the equivalence between strong and weak penalizations [Hu1]. □

Some bibliographical notes on the BB condition

The BB condition introduced earlier plays a fundamental role in the theory of mixed and penalty methods. Here we want to mention briefly the most important works that sketch its evolution towards the actual knowledge. The first fundamental paper to be referred is due to Babuška [Ba1], where a general variational problem is considered. He introduced two general conditions on the bilinear form that defines the problem that may be viewed as a generalization of the classical Lax-Milgram Lemma. The particular case of saddle point problems was considered by Brezzi [Br], who presented the condition under the form that has been used here, that is, condition (3.10). He proved a general existence and uniqueness theorem and the error estimate (3.34). It was early recognized that the main difficulty that arises when considering the discrete finite element problem is that, although the stability condition holds for the continuous problem, it is not automatically inherited by the discrete version. The stability of the continuous Stokes equations had already been proved by Ladyzhenskaya [La]. This is why the BB condition is sometimes called LBB (Ladyzhenskaya-Babuška-Brezzi) condition.

Later, this stability condition has been applied to a wide variety of mixed problems. Babuška *et al.* [BOP] proved convergence of several mixed methods using mesh dependent norms for different methods that were known to work well. However, most of the effort was placed on finding methods for effectively checking the stability condition. With regards to the incompressibility constraint, the first method that enjoyed widespread use is due to Fortin [Fo1], who showed that in some cases the BB condition follows if a continuous operator can be built from the space V to V_h. This method has been used, for example, to prove the stability of the so called *mini-element* [ABF]. Another important method was designed by Crouzieux & Raviart [CR], who also introduced a widely used stabilization technique based on the introduction of bubble functions. Bercovier & Pironneau [BP] first proved the stability of the Taylor-Hood elements [TH]. Later, their analysis was simplifyed by Verfürth [Ve1]. Recently, Brezzi & Falk have proved the stability of higher-order Taylor-Hood elements [BFa]. The stability of some low-order elements was also proved by Boland & Nicolaides [BN3], using the method they introduced in [BN1]. Let us also mention a promising technique due to Stenberg [St1], [St2] (see also [St3] for a self-contained presentation) that together with Fortin's method seems to be the simplest tool for proving div-stability of finite elements.

A review of mixed methods using the velocity-pressure and the tension-velocity approaches can be found in References [BB] and [Ar], respectively.

Finally, let us mention another technique introduced by Zienkiewicz *et al.* [ZQT] that *is not* sufficient to assert div-stability but that turns out to be of practical interest is most of the cases. It is based on a degrees-of-freedom counting, requiring solvability of the discrete problem (not stability!) for any 'patch' of elements. The idea underlying this method is the patch test introduced by Irons (see, e.g. [IL], [Ra], [TSZ]).

An engineering-oriented approach to the use of mixed finite element interpolations can be found in the books of Hughes [Hu2] and Zienkiewicz & Taylor [ZT].

3.3 An example of strong penalization: The biquadratic element for two–dimensional incompressible flows

The first tentative title for this section was *An example of failure: the ...* This already gives an idea of the conclusions that may be drawn from the results to be presented here.

At first glance, the use of a certain quadrature rule for the volumetric term (3.45) to emulate a pure mixed velocity-pressure interpolation seems to be quite attractive. Nevertheless, it turns out that problems arise when trying to do this. The purpose of this section is to convince the reader that the slightly higher computational effort associated to the weak penalization (3.43) method when it is compared to the strong penalization (3.44) is certainly worth affording.

Here, we will restrict ourselves to the Lagrangian biquadratic interpolation for the velocity. Several reduced integration rules for the volumetric term will be treated, trying to reproduce the Q_2/P_0, Q_2/Q_1 and Q_2/P_1 elements. In order to illustrate the exposition, the well known cavity flow test will be used to exemplify the application of the ideas of the text.

Since the strong penalization will not be used any more in this work, a simple although complete numerical algorithm will be presented for the Navier-Stokes equations, including the pressure calculation and the treatment of problems with a high cell Reynolds number using a Petrov-Galerkin technique.

3.3.1 Gauss–Legendre quadrature for the volumetric term

First we shall consider a one-point quadrature rule. Clearly, the pressure space Q_h in this case will consist of discontinuous piecewise constant functions. The element to be reproduced will be the Q_2/P_0 pair, which is div-stable. Let us introduce the notation

$$b_*(q_h, \mathbf{v}_h) := (q_h, \nabla \cdot \mathbf{v}_h)_*, \qquad q_h \in Q_h, \quad \mathbf{v}_h \in V_h \tag{3.52}$$

Since the divergence of biquadratic functions contains a complete second degree polynomial, the second condition in (3.41) will not be satisfied and the strong and weak penalizations will not be exactly equivalent. There is a consistency error due to the numerical quadrature rule that will be given by the power α in the estimate

$$|b(q_h, \mathbf{v}_h) - b_*(q_h, \mathbf{v}_h)| \leq Ch^\alpha \|q_h\|_Q \|\mathbf{v}_h\|_V, \qquad q_h \in Q_h, \quad \mathbf{v}_h \in V_h \tag{3.53}$$

Clearly, $\alpha > 0$ is needed if the mixed interpolation is to be approximated as $h \to 0$. For the particular element under consideration, the one-point rule integrates exactly bilinear functions. Therefore, $\alpha = 1$.

In Reference [Fo3], it is proved that condition (3.53) is sufficient to assert that the bilinear form $b_*(\cdot,\cdot)$ will satisfy the BB condition for h sufficiently small if $b(\cdot,\cdot)$ does. However, it can be proved that $b_*(\cdot,\cdot)$ is stable for *any* h using the following stronger result (cf. [BFo], Prop. II.2.19): *If there exists an homeomorphism* $\Pi_h : V_h \longrightarrow V_h$ *such that*

$$b(q_h, \mathbf{v}_h) = b_*(q_h, \Pi_h \mathbf{v}_h), \qquad q_h \in Q_h, \quad \mathbf{v}_h \in V_h \tag{3.54}$$

then $b_*(\cdot,\cdot)$ *satisfies the BB condition iff* $b(\cdot,\cdot)$ *does.*

In [BFo], Π_h is constructed explicitly for uniform meshes, thus proving that the resulting scheme will be stable in this case. The integration error (3.53) will nevertheless remain.

Consider now the Gauss-Legendre 2×2 integration rule for the volumetric term. The associated pressure space will consist of piecewise bilinear pressures. Since this rule can integrate exactly bicubic polynomials, conditions (3.41) will be verified for uniform meshes (although a small quadrature error will remain if they are distorted) and the Q_2/Q_1 pair will be exactly reproduced. However, it has already been mentioned that this element suffers from having a small stability constant K_b, which tends to zero as $h \to 0$ and also from the fact that spurious pressure modes may appear.

The 3×3 quadrature rule is inadequate, since it leads to the Q_2/Q_2 element, known to be completely unstable and to yield meaningless answers (locking phenomenon).

We have tested the Q_2/P_0 and Q_2/Q_1 elements for the driven cavity flow problem. Results are shown in Figure 3.1 (Stokes problem). In all the cases, we have taken $\mu = 1$ and a uniform mesh of 21×21 nodal points (10×10 biquadratic elements) to discretize the domain $[0,1] \times [0,1]$. Body forces are zero. The penalty parameter has been taken $\epsilon = 10^{-8}$. Homogeneous Dirichlet conditions have been prescribed everywhere on the boundary except in the upper edge, where two types of boundary conditions have been tested:

$$A - \textit{Leaky lid cavity}: \mathbf{u} = (1,0) \quad \text{for} \quad y = 1, \quad 0 \leq x \leq 1$$

$$B - \textit{Ramp condition}: \mathbf{u} = (1,0) \quad \text{for} \quad y = 1, \quad 0 < x < 1$$

It is known that the singularity on the boundary conditions of problem B is more difficult to reproduce by numerical schemes than the one of problem A.

Figures 3.1.(1) and 3.1.(2) show the velocity vectors obtained for problem A using the one-point and 2×2 rules, respectively. Results are good in both cases, although the former seems to yield overdiffusive answers. This is known to happen for the pure Q_2/P_0 interpolation and it is even more accentuated now due to the integration error (3.53). For problem B, the one-point rule still gives a stable approximation (Figure 3.1.(3)), but oscillations appear when the 2×2 rule is employed (Figure 3.1.(4)). From these facts it may be concluded that neither the one-point nor the 2×2 rules are especially robust and accurate. Accuracy is the problem associated to the former, whereas the latter shows a lack of stability.

3.3.2 About the (im)possibility of emulating the Q_2/P_1 element

In this section we shall prove that the Q_2/P_1 *cannot* be reproduced using the strong penalty method with a reduced integration rule for the volumetric term. First, a three point quadrature rule will be presented, showing also that it is optimal. Using this rule, it will be proved that the consistency error (3.53) is bounded below by a positive constant *independent of* h and that tends to zero when the three quadrature points collapse in a single one. Therefore, for $\epsilon \to 0$ and $h \to 0$ the numerical solution ***will not*** converge to the exact incompressible solution and the error will rely on the position of the integration points.

Some numerical experiments for the driven cavity flow have been conducted to show the disastrous behavior of the algorithm for different positions of the quadrature points.

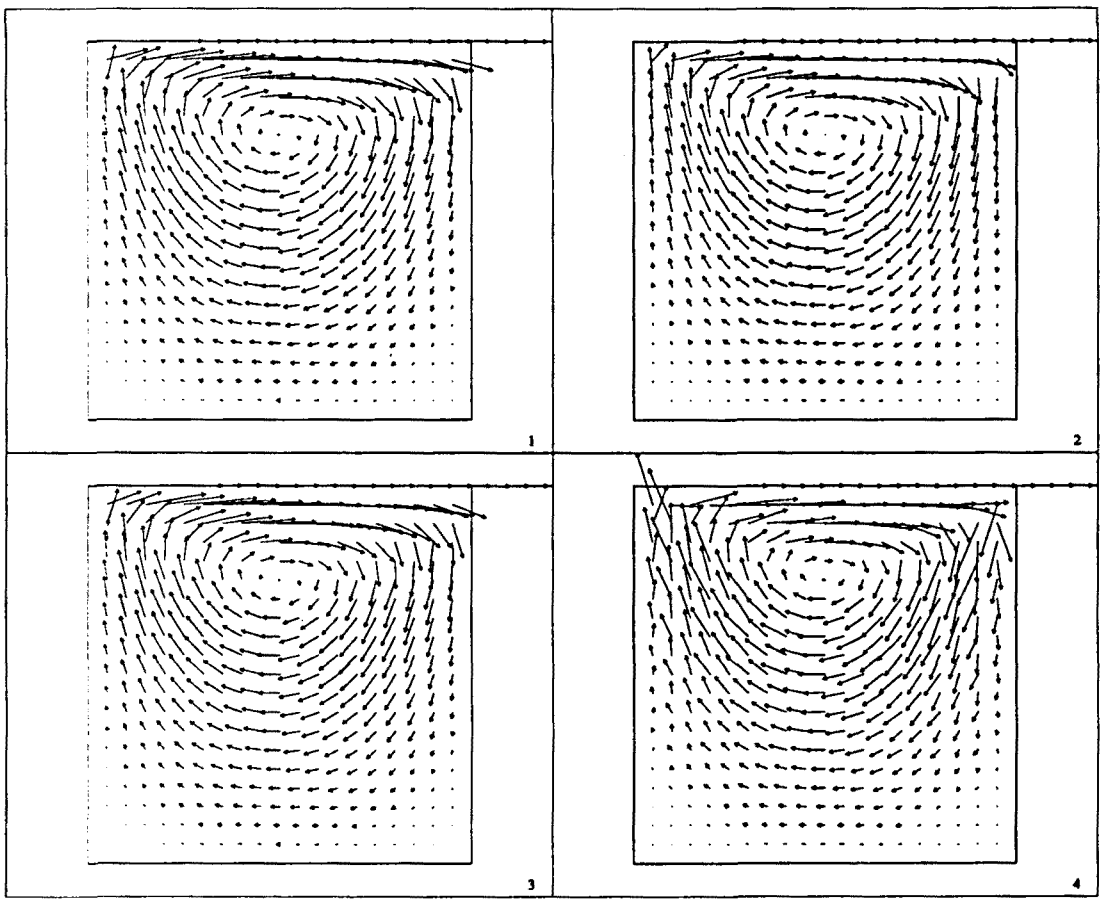

Figure 3.1 Results for the driven cavity flow problem. Stokes flow. (1): one-point rule, problem A; (2): 2 × 2 rule, problem A; (3): one-point rule, problem B; (4): 2 × 2 rule, problem B

A three point quadrature rule

Let $\Omega_0 := [-1,1] \times [-1,1]$ be the parent domain where the numerical integration is to be carried out. As usual, every subdomain Ω^e of the finite element partition will be mapped to Ω_0 using the standard isoparametric mapping. We denote by P_m the set of polynomials of degree m and by Q_m the set of tensor product polynomials of degree m in each Cartesian direction x and y.

Proposition 3.1 *The numerical quadrature rule*

$$\int_{\Omega_0} f(x,y)dxdy \approx \frac{4}{3}\sum_{i=1}^{3} f(x_i,y_i) \tag{3.55}$$

with

$$\begin{aligned}(x_1,y_1) &= (0,r),\\ (x_2,y_2) &= (-r\cos(\pi/6), -r\sin(\pi/6)),\\ (x_3,y_3) &= (r\cos(\pi/6), -r\sin(\pi/6)),\end{aligned} \tag{3.56}$$

is exact for any $f \in Q_1(\Omega_0)$ *and for all* $r \in [0,1]$. *For* $r = \sqrt{2/3}$, *it is also exact for any* $f \in P_2(\Omega_0)$. *Moreover, it is optimal, in the sense that polynomials of the form* $f(x,y) = p(x,y) + axy^2 + bx^2y$, *with* $p \in P_2(\Omega_0)$ *and* a *and* b *constants, cannot be exactly integrated using three points.*

Proof: Let $f(x,y) = p(x,y) + axy^2 + bx^2y$, with $p \in P_2(\Omega_0)$. Denote by (x_i, y_i), $i = 1,2,3$ the coordinates of the quadrature points and by ω_i their weights. Imposing that

$$\int_{\Omega_0} f(x,y)dxdy = \sum_{i=1}^{3} \omega_i f(x_i, y_i)$$

the following relations are found:

$$\begin{aligned}
\omega_1 + \omega_2 + \omega_3 &= 4 && (a)\\
x_1\omega_1 + x_2\omega_2 + x_3\omega_3 &= 0 && (b)\\
y_1\omega_1 + y_2\omega_2 + y_3\omega_3 &= 0 && (c)\\
x_1y_1\omega_1 + x_2y_2\omega_2 + x_3y_3\omega_3 &= 0 && (d)\\
x_1^2\omega_1 + x_2^2\omega_2 + x_3^2\omega_3 &= 4/3 && (e)\\
y_1^2\omega_1 + y_2^2\omega_2 + y_3^2\omega_3 &= 4/3 && (f)\\
x_1y_1^2\omega_1 + x_2y_2^2\omega_2 + x_3y_3^2\omega_3 &= 0 && (g)\\
x_1^2y_1\omega_1 + x_2^2y_2\omega_2 + x_3^2y_3\omega_3 &= 0 && (h)
\end{aligned}$$

Let us first show that it is impossible to fulfil all these conditions with three different points.

Assume $\omega_i \neq 0$, $i = 1,2,3$. If $x_j \neq 0$ for some j, requiring non-trivial solvability from conditions (b), (d) and (g) it is found that

$$\omega_1\omega_2\omega_3(y_2 - y_1)(y_3 - y_1)(y_3 - y_2) = 0$$

Assume, without loss of generality, that $y_2 = y_3 \neq y_1$. If $y_j \neq 0$ for some j, conditions (c), (d) and (h) imply

$$\omega_1\omega_2\omega_3(x_2 - x_1)(x_3 - x_1)(x_3 - x_2) = 0$$

Suppose that $x_1 = x_2 \neq x_3$ ($x_2 = x_3$ would mean that points 2 and 3 collapse). Since $\omega_i \neq 0$, from conditions (d), (g) and (h) it follows that

$$x_1^2x_3y_1y_2^2(y_2 - y_1)(x_3 - x_1) = 0$$

If $y_1 = y_2$ or $x_1 = x_3$ two points would collapse. Assume that $x_1 = x_2 = 0$, $x_3 \neq 0$. From Eqns. (g) and (h) it is found that $y_2 = y_3 = 0$, and from (c) $y_1 = 0$. Hence, points 1 and 2 have inevitably collapsed.

Once it is known that two points must coincide, it is easy to see that system $(a) - (h)$ does not have any solution. For example, take $\omega_3 = 0$. From (d) and (g) it follows that $y_1 = y_2$ and from (d) and (h) $x_1 = x_2$. With a single point only bilinear polynomials can be exactly integrated.

The Proposition follows by checking that conditions $(a) - (d)$ hold for $\omega_1 = \omega_2 = \omega_3 = 4/3$ and (x_i, y_i) given by (3.56). If $r = \sqrt{2/3}$, also (e) and (f) hold. □

Remark 3.2

The quadrature rule defined by (3.55)–(3.56) seems to have been used before (cf. [Ki]), although the author was not aware of this fact at the moment of undertaking this work. □

Inconsistency of $b_*(\cdot,\cdot)$

If the volumetric term is integrated using a three-point quadrature rule, the pressure space will consist of piecewise linear polynomials. For $r = \sqrt{2/3}$, the first condition (3.41) will be fulfilled for uniform meshes. However, the bilinear form $b_*(\cdot,\cdot)$ cannot approximate the bilinear form $b(\cdot,\cdot)$ for a fixed $r > 0$. This is what the following result states.

Proposition 3.2 *Assume that the quadrature rule of Proposition 3.1 is used for the volumetric term. Then, there is a constant* $C(r)$ *independent of* h, *with* $C(r) \to 0$ *as* $r \to 0$, *such that*

$$\sup_{q_h,\mathbf{v}_h} \frac{|b(q_h,\mathbf{v}_h) - b_*(q_h,\mathbf{v}_h)|}{\|q_h\|_Q \|\mathbf{v}_h\|_V} \geq C(r),$$

where the supremum is taken over all the $q_h \in Q_h - \{0\}$ *and* $\mathbf{v}_h \in V_h - \{0\}$.

Proof: Since the functions in Q_h are discontinuous, we can take q_h^0 constant on an element Ω^e and zero everywhere else. Take also $\mathbf{v}_h^0 = N_9\mathbf{a}$, where $N_9(x,y)$ is the shape function associated to the central node of the element and $\mathbf{a} \neq \mathbf{0}$ is a constant vector. Without loss of generality, we shall assume Ω^e to be affine with the parent domain Ω_0, where natural coordinates ξ, η are taken. We will have that

$$b(q_h^0,\mathbf{v}_h^0) = q_h^0 \int_{\Omega^e} \nabla\cdot(N_9\mathbf{a})d\Omega = q_h^0 \int_{\partial\Omega^e} \mathbf{n}\cdot(N_9\mathbf{a})d\Gamma = 0$$

and

$$b_*(q_h^0,\mathbf{v}_h^0) = C_1\ q_h^0\ \frac{1}{3}\mathrm{meas}(\Omega^e)\ |\sum_{i=1}^{3} \nabla N_9(\xi_i,\eta_i)||\hat{\mathbf{a}}|\cos\beta,$$

where the constant C_1 takes into account the angles of the edges in Ω^e, $|\hat{\mathbf{a}}|$ is the Euclidian norm of the vector transformed of $\mathbf{a}$ to the parent domain and β is the angle between $\hat{\mathbf{a}}$ and $\sum_{i=1}^{3} \nabla N_9(\xi_i,\eta_i)$. Since

$$\nabla N_9(\xi,\eta) = \left(-2\xi(1-\eta^2), -2\eta(1-\xi^2)\right),$$

it is found that

$$|\sum_{i=1}^{3} \nabla N_9(\xi_i,\eta_i)| = \frac{3}{2}r$$

Considering that

$$\|q_h^0\|_Q = |q_h^0|\ \mathrm{meas}(\Omega^e)^{\frac{1}{2}}, \qquad \|\mathbf{v}_h^0\|_V = C_2\ |\hat{\mathbf{a}}|\ \mathrm{meas}(\Omega^e)^{\frac{1}{2}},$$

with C_2 depending on β and the $V-$norm of N_9 in Ω_0, we will have that

$$\sup_{q_h,\mathbf{v}_h} \frac{|b(q_h,\mathbf{v}_h) - b_*(q_h,\mathbf{v}_h)|}{\|q_h\|_Q\|\mathbf{v}_h\|_V} \geq \sup_{\beta} \frac{|b_*(q_h^0,\mathbf{v}_h^0)|}{\|q_h^0\|_Q\|\mathbf{v}_h^0\|_V} = Cr$$

for a certain constant C. □

It is clear from this result that the higher r is (with $0 \leq r \leq 1$), the more inaccurate will be approximation to the incompressibility condition. To verify this numerically, the driven cavity flow using the introduced method has been solved. Results are shown in Figure 3.2. In Figures 3.2.(1) and 3.2.(2), $r = \sqrt{2/3}$ has been used. The first case corresponds to the leaky lid boundary conditions and the second to the ramp condition. Although no oscillations are apparent in any of the two cases, it is obvious that the incompressibility constraint has been excessively relaxed. The central vortex has moved down with respect to the reference results of Figure 3.1. To check the effect of the value of r, the problem with the ramp condition has been run again with $r = 0.5$ and $r = 0.9$ (Figures 3.2.(3) and 3.2.(4), respectively). As expected, the zero divergence condition is totally violated in the latter case, where results are meaningless. Looking at Figures 3.2.(3), 3.2.(2) and 3.2.(4) we see how increasing value of r (0.5, $\sqrt{2/3}$ and 0.9, respectively) this constraint is more and more relaxed.

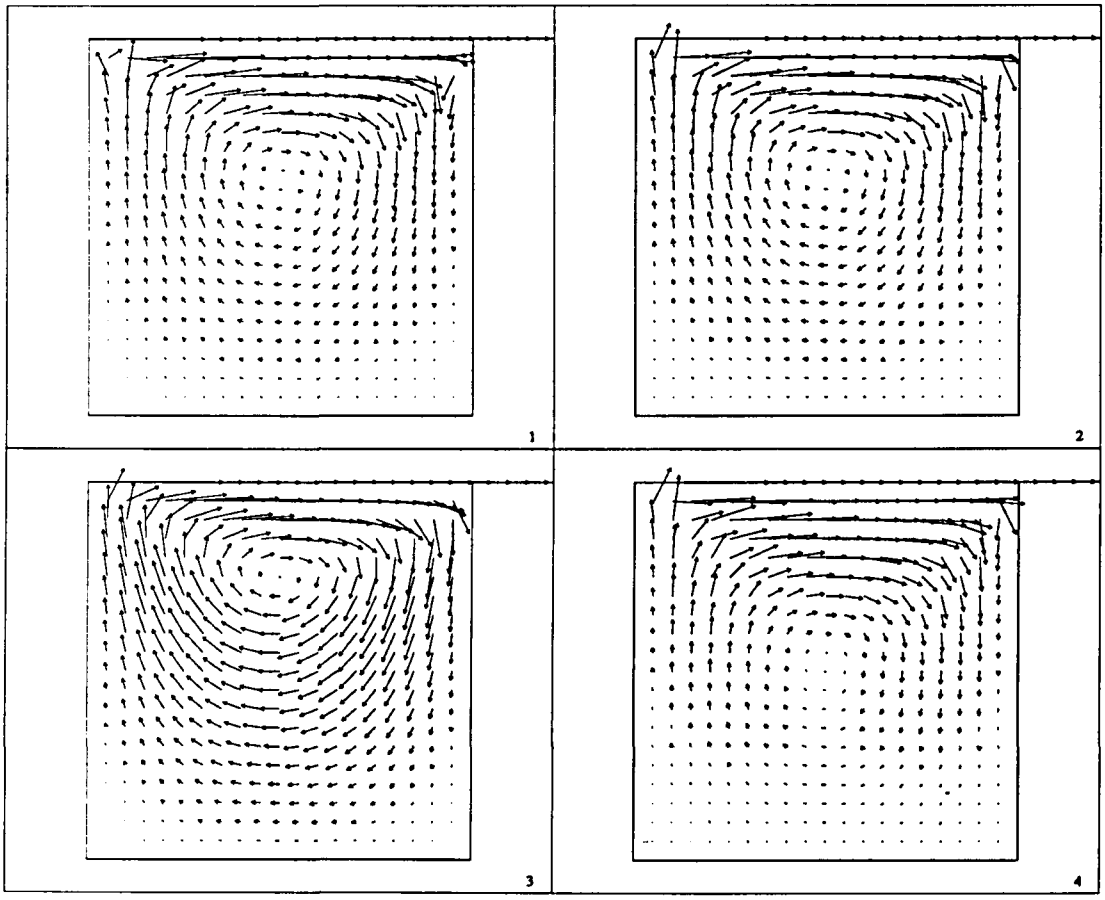

Figure 3.2 Results for the driven cavity flow problem. Stokes flow. Three-point quadrature rule for the volumetric term. (1): $r = \sqrt{2/3}$, leaky lid condition; (2): $r = \sqrt{2/3}$, ramp condition; (2): $r = 0.5$, ramp condition; (2): $r = 0.9$, ramp condition.

3.3.3 Pressure calculation

Once the velocity has been calculated by solving problem (3.47), relation (3.40) gives the values of the pressure at the reduced integration points. There are two questions to be considered. First, the pressures will be discontinuous across interelement boundaries and thus a smoothing facility has to be introduced, mainly for plotting purposes. The second aspect is that some of the elements that can be emulated through underintegration of the volumetric term exhibit spurious pressure modes (cf. Box 3.1). Sometimes, the smoothing technique utilized may be enough to remove these modes, but this in general will not be true, especially near the boundaries [HLB]. On the other hand, the penalty approach automatically precludes the appearence of these spurious pressures. Nevertheless, pressure convergence cannot be ensured [JP], [OKS] and the pressure space has to be redefined in order to obtain this convergence (under stricter regularity assumptions than usual for the exact solution [JP]).

Here, both problems will be treated and applied to the Q_2/Q_1 and Q_2/P_0 elements. Also, the possibility of solving a Poisson equation for the pressure will be discussed. As far as we are aware, this method has only been applied to the Q_1/P_0 element by Sohn & Heinrich [SH]. Since it is impossible for this element to approximate the second derivatives for the velocity field that are required for this method, they had to calculate them via the interpolation of the nodal *first* derivatives and using a least squares technique. This problem is circumvented if quadratic elements are used.

For a detailed and up-to-date discussion of the Poisson problem for the pressure the reader is referred to the paper of Gresho & Sani [GS], where both theoretical and practical questions are treated. Of special interest is the discussion on the boundary conditions to be imposed to the pressure.

Pressure filter and Least-squares smoothing (LSS)

Consider first the Q_2/Q_1 element and assume that the velocity field $\mathbf{u}_h$ is known in the parent domain Ω_0, where the numerical integration has been performed. Eqn. (3.40) will give the value of the pressure at the quadrature points. As explained earlier, spurious modes may be present for the non-penalized problem and pollute the penalized solution. In order to remove them, what we do is to project the space of piecewise bilinear pressures onto the space of piecewise *linear* pressures, since it is known that no spurious modes appear for the Q_2/P_1 element. We have successfully tested the following method.

Let p_i^e, $i = 1, 2, 3, 4$, be the pressure values in the parent domain corresponding to element e computed using (3.40). Let also $\gamma = \sqrt{3}/2$ and

$$(\xi_1, \eta_1) = (-\gamma, -\gamma), \qquad (\xi_2, \eta_2) = (\gamma, -\gamma),$$
$$(\xi_3, \eta_3) = (-\gamma, \gamma), \qquad (\xi_4, \eta_4) = (\gamma, \gamma)$$

be the coordinates of the quadrature rule-points. From p_i^e we construct the following piecewise linear pressure:

$$p_l^e(\xi, \eta) = a_0 + a_1\xi + a_2\eta \tag{3.57}$$

with

$$a_0 := \frac{1}{4}\sum_{i=1}^{4} p_i^e, \qquad \text{(mean value)}$$

$$a_1 := \frac{1}{2}\left(\frac{p_2^e - p_1^e}{2\gamma} + \frac{p_4^e - p_3^e}{2\gamma}\right) \qquad (\xi\text{–derivative})$$

$$a_2 := \frac{1}{2}\left(\frac{p_3^e - p_1^e}{2\gamma} + \frac{p_4^e - p_2^e}{2\gamma}\right) \qquad (\eta\text{–derivative})$$

Now we have a discontinuous piecewise linear pressure given by (3.57), for which no spurious modes are expected. In order to obtain a continuous pressure field $p_s(x,y)$, we use a standard Least-squares method [Hu2], [ZT], described in more detail in Chapter 4. The pressure $p_s(x,y)$ is interpolated using the biquadratic element. If $N^{(i)}(x,y)$ denotes the shape function of the ith node of the finite element mesh, the vector of smoothed pressure nodal unknowns, say $\mathbf{P}_s$, will be found by solving the algebraic system

$$\mathbf{M}\mathbf{P}_s = \mathbf{R}_s, \tag{3.58}$$

where the symmetric and positive-definite matrix $\mathbf{M}$ has components $M_{ij} = \int_\Omega N^{(i)}N^{(j)}d\Omega$ and

$$\begin{aligned} R_{s,j} &= \mathcal{A}_{e=1}^{N_{el}} \int_{\Omega^e} N^{(j)}(x,y)\; p_l^e(x,y)d\Omega \\ &= \mathcal{A}_{e=1}^{N_{el}} \int_{\Omega_0} N^{(j)}(\xi,\eta)\; p_l^e(\xi,\eta)\; |J(\xi,\eta)|d\Omega, \end{aligned}$$

$J(\xi,\eta)$ being the Jacobian determinant of the isoparametric mapping and $p_l^e(\xi,\eta)$ being given by (3.57).

For the Q_2/P_0 element, this smoothing is also applied, although now the pressure is constant and given directly by (3.40).

Pressure Poisson equation

Taking the divergence of the momentum equation in (3.1) yields

$$\Delta p = \nabla\cdot[\rho\mathbf{f} + \mu\Delta\mathbf{u} - \rho(\mathbf{u}\cdot\nabla)\mathbf{u}] \qquad \text{in } \Omega \tag{3.59}$$

i.e., a Poisson equation for the pressure. If we assume the velocity field to be solenoidal and sufficiently smooth, we would have that $\nabla\cdot(\mu\Delta\mathbf{u}) = \mu\Delta(\nabla\cdot\mathbf{u}) = 0$. However, we will keep the term $\mu\Delta\mathbf{u}$ in (3.59) for the moment.

In Reference [GS], it is concluded that the correct boundary condition for (3.59) is the imposition of the conservation of the normal component of the momentum. Multiplying the first equation in (3.1) by the unit outward normal $\mathbf{n}$ and evaluating on the boundary (understanding this process as a limit) we are led to

$$\frac{\partial p}{\partial n} = \mathbf{n}\cdot[\rho\mathbf{f} + \mu\Delta\mathbf{u} - \rho(\mathbf{u}\cdot\nabla)\mathbf{u}] \qquad \text{on } \partial\Omega \tag{3.60}$$

The solution of the Neumann problem (3.59)–(3.60) is unique up to an additive constant that can be fixed by specifying the value of the pressure at a certain point.

Let $\hat{Q} = H^1(\Omega)$ and $\hat{Q}_0 = H^1(\Omega)/\mathbb{R}$. The weak form of problem (3.59)–(3.60) is: Find $p \in \hat{Q}_0$ such that

$$\int_\Omega \nabla q \cdot \nabla p \, d\Omega = \int_{\partial\Omega} q \left[\frac{\partial p}{\partial n} - \mathbf{n} \cdot \mathcal{F}_m\right] d\Gamma + \int_\Omega \nabla q \cdot \mathcal{F}_m \, d\Omega \qquad \forall q \in \hat{Q} \qquad (3.61)$$

where

$$\mathcal{F}_m(\mathbf{u}) := \rho \mathbf{f} + \mu \Delta \mathbf{u} - \rho(\mathbf{u} \cdot \nabla)\mathbf{u}$$

is the vector whose divergence appears in the right-hand-side of (3.59). Since (3.60) establishes that $\partial p/\partial n = \mathbf{n} \cdot \mathcal{F}_m$, the boundary term in (3.61) vanishes.

Construct now conforming finite element spaces $\hat{Q}_{0,h} \subset \hat{Q}_0$ and $\hat{Q}_h \subset \hat{Q}$ using the biquadratic element. The discrete counterpart of problem (3.61) is to find $p_h \in \hat{Q}_{0,h}$ such that

$$(\nabla q_h, \nabla p_h) = (\nabla q_h, \mathcal{F}_m(\mathbf{u}_h)) \qquad \forall q_h \in \hat{Q}_h \qquad (3.62)$$

Let us discuss now the approximation properties that can be expected from the solution of problem (3.62). It is well established (see, e.g. [Ci]) that the error $\|p - p_h\|_0$ is of order $O(h^3)$ for a given function $\mathcal{F}_m(\mathbf{u}_h)$. However, the real problem is the approximation of $\mathcal{F}_m(\mathbf{u}_h)$ to $\mathcal{F}_m(\mathbf{u})$, with $\mathbf{u}_h \in V_h$. If we assume that the inverse estimate [Ci]

$$\|\mathbf{v}_h\|_{s,\Omega^e} \leq C \, h^{-1} \|\mathbf{v}_h\|_{s-1,\Omega^e}$$

holds for any $\mathbf{v}_h \in V_h$, we can roughly expect that

$$\|\mathcal{F}_m(\mathbf{u}) - \mathcal{F}_m(\mathbf{u}_h)\|_0 \sim h^{-2} \, \|\mathbf{u} - \mathbf{u}_h\|_0$$

since $\mathcal{F}_m(\mathbf{u}_h)$ involves second derivatives of $\mathbf{u}_h$. Therefore, the error in the pressure associated to problem (3.62) will be driven by the approximation of $\mathcal{F}_m(\mathbf{u}_h)$ to $\mathcal{F}_m(\mathbf{u})$.

For the Q_2/Q_1 element, the best we can hope for is the interpolation error $\|\mathbf{u} - \mathbf{u}_h\|_0 = O(h^3)$, if no instabilities are present. Therefore, $\|\mathcal{F}_m(\mathbf{u}) - \mathcal{F}_m(\mathbf{u}_h)\|_0 = O(h)$. However, for the Q_2/P_0 element the error will be $\|\mathbf{u} - \mathbf{u}_h\|_0 = O(h^2)$, due to the poor pressure approximation. In this case, $\|\mathcal{F}_m(\mathbf{u}) - \mathcal{F}_m(\mathbf{u}_h)\|_0 = O(1)$, i.e., no convergence can be expected of $\mathcal{F}_m(\mathbf{u}_h)$ to $\mathcal{F}_m(\mathbf{u})$.

Numerical experiments show that this quasi-heuristic considerations are pessimistic, at least for the Stokes problem.

Remarks 3.3

(1) If we would have taken $\mathcal{F}_m(\mathbf{u}) = \rho \mathbf{f} - \rho(\mathbf{u} \cdot \nabla)\mathbf{u}$, without the viscous term, second derivatives should be calculated not in Ω but on $\partial\Omega$, since the boundary term in (3.61) would be

$$\int_{\partial\Omega} q \, \mu \, \mathbf{n} \cdot \Delta \mathbf{u} \, d\Gamma$$

The situation now is even worse than before, since roughly speaking approximations on the boundary have a gap 1/2 with respect to approximations in the interior of the domain (cf. Remark 1.4.(3)). In fact, in Reference [SH] it is concluded that the method we have considered yields better numerical results than this one.

(2) Although we have considered the velocity $\mathbf{u}_h$ known, the way the Poisson equation for the pressure is usually used is to guess a velocity field and solve (3.62). The pressure so calculated is used to recompute the velocity. This procedure is

repeated until convergence is achieved. This is a common way to uncouple the velocity and pressure calculations. □

Numerical performance

We have computed the pressure for the driven cavity flow with leaky lid boundary conditions using the two methods just described. Figures 3.3.(1) and 3.3.(2) show the results obtained using the pressure filtering and the least-squares smoothing for the one-point and the 2 × 2 integration rules, respectively. As it was already observed for the velocity field, results are overdiffusive for the first option. Figures 3.3.(3) and 3.3.(4) correspond to the same cases using the pressure Poisson equation. Results are 'smoother' than before and qualitatively similar. It must be pointed out that for the one-point rule pressure convergence cannot be guaranteed and for the 2 × 2 rule it is reduced to $O(h)$. Nevertheless, results seem to be fairly good.

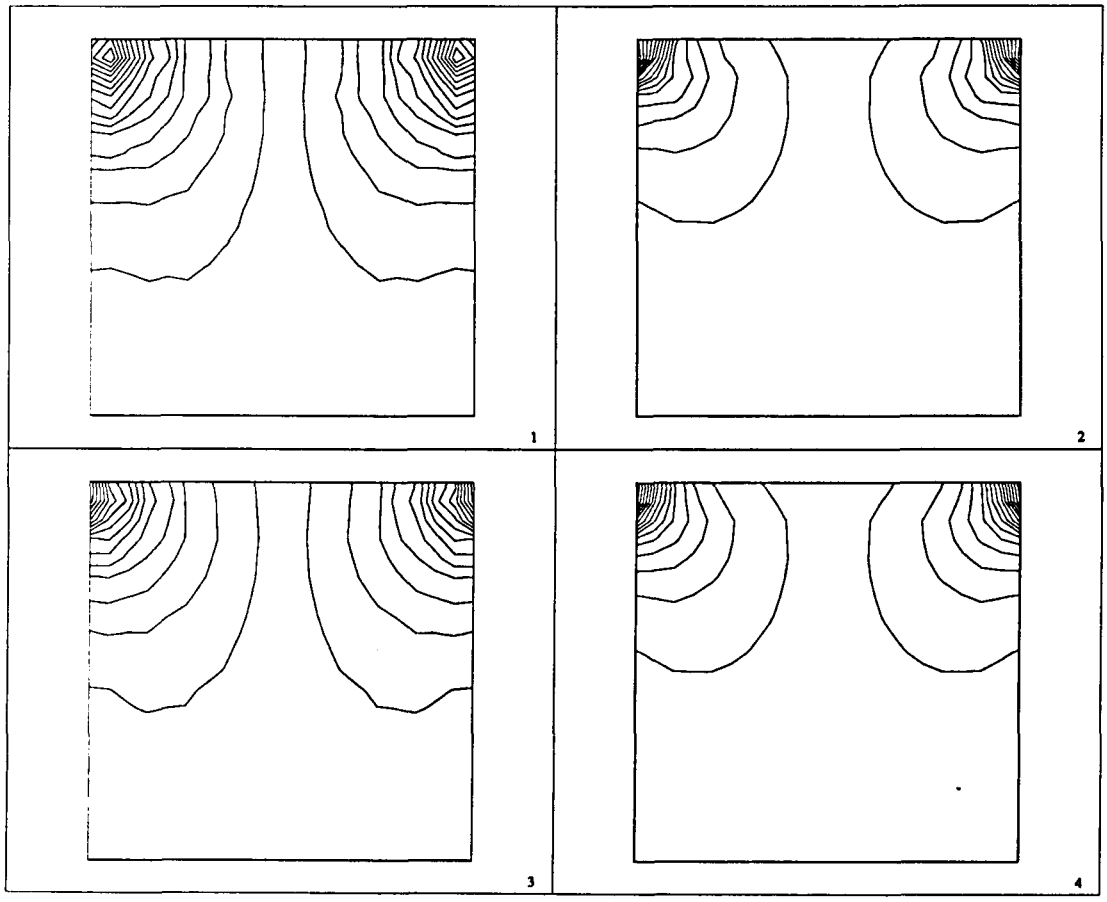

Figure 3.3 Pressures for the driven cavity flow problem (biquadratic element). Stokes flow. (1): one-point rule, least-squares smoothing; (2): 2 × 2 rule, least-squares smoothing; (3): one-point rule, pressure Poisson equation; (4): 2 × 2 rule, pressure Poisson equation.

3.3.4 Numerical procedures for the Navier–Stokes equations

To close this incursion into the use of the strong penalty method, we shall briefly describe the numerical techniques we have employed when dealing with the Navier-Stokes equations (3.1). The matrix notation introduced at the end of Section 3.2.3 will be kept.

Linearization technique

In order to solve the nonlinear algebraic system (3.47), the simplest linearization method has been employed:

Given $\mathbf{U}^{\epsilon(0)}$, *for* $i = 1, 2, \ldots$ *solve for* $\mathbf{U}^{\epsilon(i)}$:

$$\mathbf{K}_d\mathbf{U}^{\epsilon(i)} + \frac{1}{\epsilon}\mathbf{K}_v\mathbf{U}^{\epsilon(i)} = \mathbf{F}_u - \mathbf{K}_c\left(\mathbf{U}^{\epsilon(i-1)}\right)\mathbf{U}^{\epsilon(i-1)} \tag{3.63}$$

This method has the important advantadge that the matrix of the algebraic system to be solved at each iteration, viz. $\mathbf{K}_d+(1/\epsilon)\mathbf{K}_v$, is symmetric and positive-definite. Thus, it has to be factored only once and the computational effort of each iteration is reduced to a forward and a backward substitution. In our computations, we have used the frontal algorithm for symmetric matrices due to Irons [Ir] and described in [HO].

Convergence has been checked using the criterion

$$|\mathbf{U}^{\epsilon(i)} - \mathbf{U}^{\epsilon(i-1)}| \leq TOL|\mathbf{U}^{\epsilon(i)}| \tag{3.64}$$

where TOL is a given tolerance and $|\cdot|$ denotes the Euclidian norm of a vector.

Continuation method

The main drawback of algorithm (3.63) is that it only converges (and linearly) if the initial guess $\mathbf{U}^{\epsilon(0)}$ is close enough to the final solution. We take $\mathbf{U}^{\epsilon(0)} = \mathbf{0}$ and thus the effective initial guess is the Stokesian solution $\mathbf{U}^{\epsilon(1)}$. Convergence will only be possible for moderate values of the Reynolds number Re. It turns out that in practice the algorithm only converges for very small values of Re. In order to 'push' the scheme to higher values, we have employed a very simple continuation method, based on incrementing Re up to the final value in several steps. The density ρ has been used as the incrementing factor.

A detailed analysis of different iterative techniques and continuation methods for the strong penalty approach can be found in the papers by Carey & Krishnan [CK3], [CK4], where several numerical results are presented.

Petrov-Galerkin weighting

The extension of the Streamline Diffusion (SD) method described in Chapter 1 to the Navier-Stokes equations is by no means unique, in the sense that several schemes have been proposed in the recent literature. We will come back to this point later in Chapter 4.

Although the iterative method described earlier does not allow for solving highly convective flows, it may be possible that the mesh diameter be large and therefore the *cell Reynolds number*

$$(Re)^e := \rho\frac{|\mathbf{u}^e|h^e}{2\mu} \tag{3.65}$$

be also large. As usual, superscript e denotes characteristic values for an element. From the discussion of Chapter 1, for values $(Re)^e > 2$ numerical oscillations may be expected when quadratic elements are used.

The SD method that has been implemented is based on the perturbation of the test function $\mathbf{v}_h \in V_h$ to

$$\mathbf{v}_h + \tau(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h, \tag{3.66}$$

the second term only affecting the element interiors. The parameter τ in (3.66) is the intrinsic time defined by Eqn. (1.36) and the upwind functions involved are calculated as explained in Chapter 1, replacing the Péclet number γ by the cell Reynolds number given by (3.65). Problem (3.44) will be modified as follows: Find $\mathbf{u}_h \in V_h$ such that

$$\begin{aligned} c(\mathbf{u}_h^\epsilon, \mathbf{u}_h^\epsilon, \mathbf{v}_h) + a(\mathbf{u}_h^\epsilon, \mathbf{v}_h) + \frac{1}{\epsilon} d(\mathbf{u}_h^\epsilon, \mathbf{v}_h) \\ + \sum_{e=1}^{N_{el}} \mathcal{S}_e(\mathbf{u}_h^\epsilon, \mathbf{v}_h) = l(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h \end{aligned} \tag{3.67}$$

where $\mathcal{S}_e$ is the nonlinear functional

$$\begin{aligned} \mathcal{S}_e(\mathbf{u}_h, \mathbf{v}_h) := (\tau^e(\mathbf{u}_h^\epsilon \cdot \nabla)\mathbf{v}_h, \rho(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \mu\Delta\mathbf{u}_h - \rho\mathbf{f})_{\Omega^e} \\ + \left(\tau^e(\mathbf{u}_h^\epsilon \cdot \nabla)\mathbf{v}_h, -\frac{1}{\epsilon}\nabla(\nabla \cdot \mathbf{u}_h)\right)_{*,\Omega^e} \end{aligned} \tag{3.68}$$

Observe that the dependence of τ on $\mathbf{u}_h$, who is now the unknown of the problem, will be complicated.

Denote by $\mathbf{S}(\mathbf{U}) \cdot \mathbf{V}$ the matrix version of the functional $\sum_{e=1}^{N_{el}} \mathcal{S}_e(\mathbf{u}_h, \mathbf{v}_h)$ once the finite element discretization has been performed. In order to preserve the advantatges of scheme (3.63) when the SD method is used, it has been modified as follows:

Given $\mathbf{U}^{\epsilon(0)}$, *for* $i = 1, 2, \ldots$ *solve for* $\mathbf{U}^{\epsilon(i)}$:

$$\mathbf{K}_d\mathbf{U}^{\epsilon(i)} + \frac{1}{\epsilon}\mathbf{K}_v\mathbf{U}^{\epsilon(i)} = \mathbf{F}_u - \mathbf{K}_c\left(\mathbf{U}^{\epsilon(i-1)}\right)\mathbf{U}^{\epsilon(i-1)} - \mathbf{S}\left(\mathbf{U}^{\epsilon(i-1)}\right) \tag{3.69}$$

We have found numerically that the convergence rate of the initial scheme is not deteriorated because of the SD term, but rather improved.

Final algorithm

Having in mind all the techniques discussed so far, the final algorithm will read as indicated in Box 3.3. We have used the following notation: $\delta\rho$ is the density increment of each continuation step, $\mathbf{L}$ is the discrete Laplacian matrix, $\mathbf{P}_p$ are the nodal values of the pressure computed using the Poisson equation (3.62) and $\mathbf{R}_p$ is the right-hand-side arising from the discretization of this equation. Terms between parenthesis and italic characters denote logical variables.

Box 3.3 Algorithm for the Navier–Stokes equations

- Compute $\mathbf{K}_d$ using full integration
- Compute $\mathbf{K}_v$ using reduced integration
- Factorise and store $\mathbf{A} := \mathbf{K}_d + (1/\epsilon)\mathbf{K}_v$
- Set $\rho_c := 0$ and $\mathbf{U}^{\epsilon(0)} := \mathbf{0}$
- WHILE $\rho_c < \rho$ DO:
  - $i := 0$
  - $\rho_c \leftarrow \rho_c + \delta\rho$
  - WHILE *(not converged)* DO:
    - $i \leftarrow i + 1$
    - Compute $\mathbf{R}^{(i-1)} := \mathbf{F}_u - \mathbf{K}_c\left(\mathbf{U}^{\epsilon(i-1)}\right)\mathbf{U}^{\epsilon(i-1)} - \mathbf{S}\left(\mathbf{U}^{\epsilon(i-1)}\right)$
    - Solve $\mathbf{A}\mathbf{U}^{\epsilon(i)} = \mathbf{R}^{(i-1)}$
    - If $|\mathbf{U}^{\epsilon(i)} - \mathbf{U}^{\epsilon(i-1)}| \leq TOL|\mathbf{U}^{\epsilon(i)}|$ then *(converged)*

    END while *(not converged)*
  - New initial guess: $\mathbf{U}^{\epsilon(0)} \leftarrow \mathbf{U}^{\epsilon(i)}$

  END while $\rho_c < \rho$
- Compute the pressure:
  - IF *(LSS)* then
    - Solve $\mathbf{M}\mathbf{P}_s = \mathbf{R}_s$
  - ELSE if *(Poisson)* then
    - Solve $\mathbf{L}\mathbf{P}_p = \mathbf{R}_p$

    END

  END

Numerical experiments

Some very simple numerical experiments have been conducted for the driven cavity flow with leaky lid boundary conditions. The pressure has been computed solving the Poisson equation (3.62). The Reynolds number based on the length of the square (= 1) and the prescribed velocity in the upper edge (= (1, 0)) is equal to the density ρ for $\mu = 1$. In all the cases, $TOL = 0.001$ has been chosen. The penalty parameter is $\epsilon = 10^{-8}$.

First we solve the Navier-Stokes problem for $Re = 200$ and using the mesh of 21×21 nodal points employed before (using biquadratic elements). Ten continuation steps and two iterations per step have been required for convergence. In this case, no oscillations appear when using the standard Galerkin formulation. Figures 3.4.(1) and 3.4.(2) show the velocity vectors using the one-point and the 2×2 quadrature rules. The first case, as it has been already observed in former examples, yields overdiffusive and inaccurate results. In Figures 3.4.(3) and 3.4.(4) the pressure contours have been plotted. Observe that results are bad for the one-point rule and seem to be correct for the 2×2 rule. Recall that for the former case, no pressure convergence can be guaranteed, whereas for the latter the best one can hope for is an $O(h)$ approximation. These results seem to confirm the discussion of Section 3.3.3.

Next, a $Re = 250$ problem has been run on a new uniform mesh of 11×11 nodal points (5×5 biquadratic elements). Only the 2×2 point rule has been used. Figures 3.5.(1) and 3.5.(3) show the velocity vectors and the pressure contours obtained using

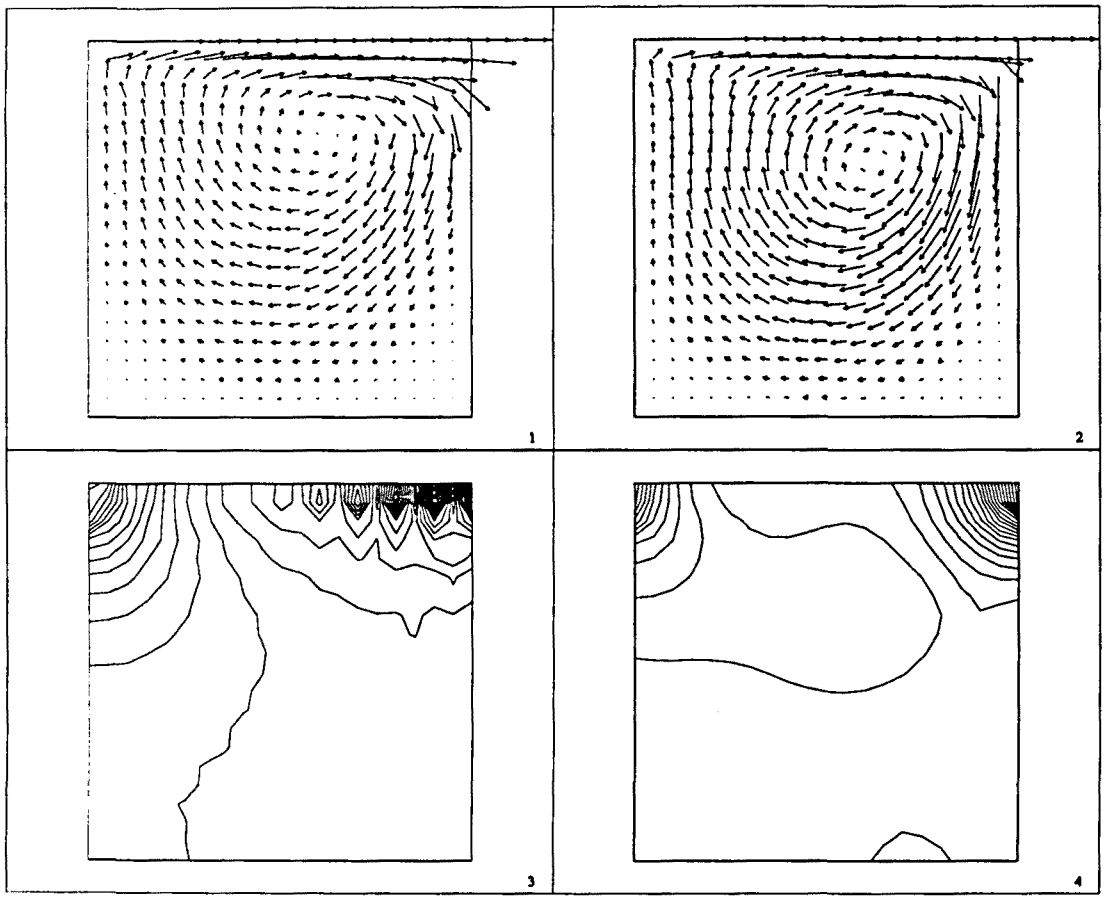

Figure 3.4 Navier-Stokes results using the 21 × 21 mesh, $Re = 200$. The pressure has been computed solving a Poisson equation. (1): Velocities using the one-point rule; (2): Velocities using the 2 × 2 point rule; (3): Pressure contours using the one-point rule; (4): Pressure contours using the 2 × 2 point rule.

the Galerkin formulation. In this case, the cell Reynolds number exceeds two for some elements in the upper right corner, where oscillations can be observed. These are removed if the SD method is used (Figures 3.5.(2) and 3.5.(4)). The upwind functions have been determined according to the methodology proposed in Section 1.4.2.

3.4 Weak penalization: analysis of an iterative penalty method

Because of the problems encountered when the strong penalty method is used, we shall move now to the weak penalization approach (hereafter referred to just as penalty method). It is not our purpose to investigate its behavior, which will be appreciated from the numerical examples in this and the following chapters, but rather to present

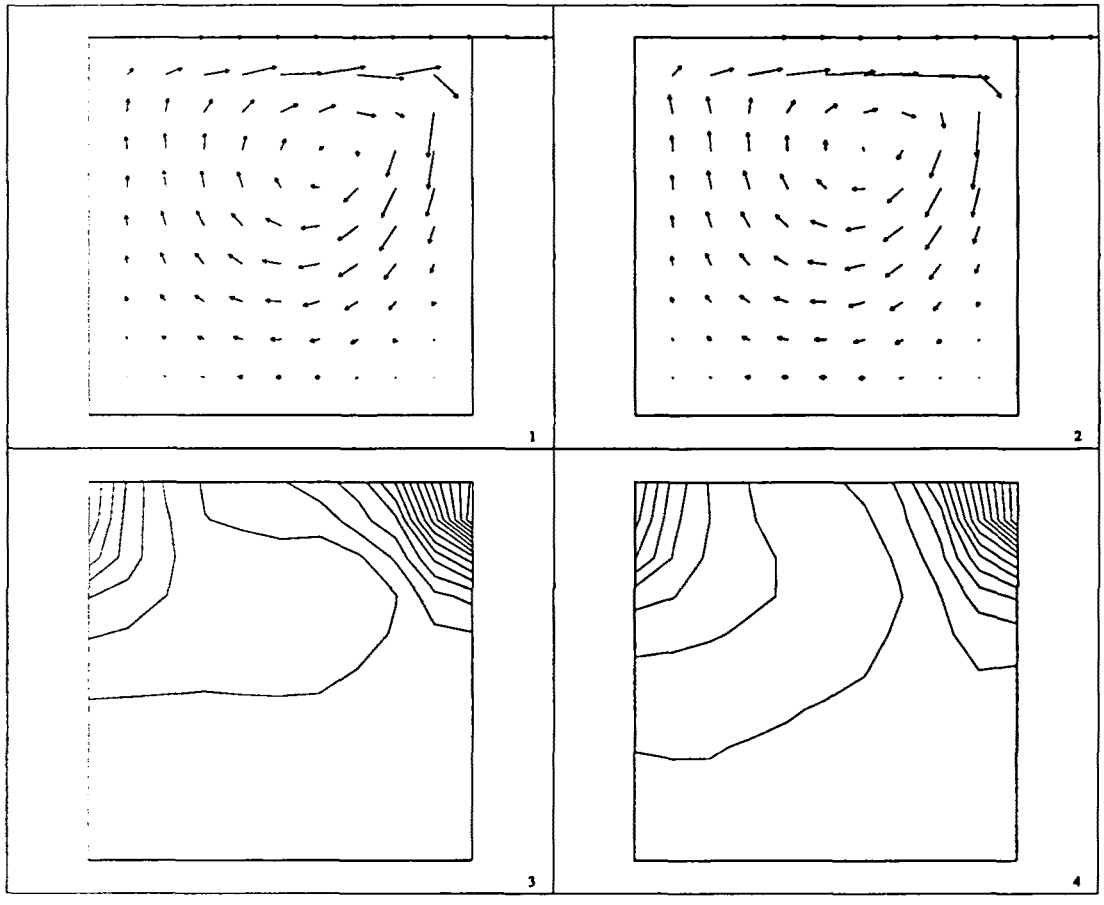

Figure 3.5 Navier-Stokes results using the 11 × 11 mesh, $Re = 250$. The pressure has been computed solving a Poisson equation and the 2 × 2 quadrature rule has been employed. (1): Velocities using the Galerkin formulation; (2): Velocities using the SD method; (3): Pressure contours using the Galerkin formulation; (4): Pressure contours using the SD method.

and analyse an *iterative* version that will be used throughout this work. The method described by Eqns. (3.43) will be called *classical penalty method.*

The main drawback of the penalty method is the ill-conditioning of the stiffness matrix when the penalty parameter is very small. For Newtonian flows, a fairly wide range of values of this parameter are known to yield good results (that is, the incompressibility equation is sufficiently well approximated) and to be easily handled if direct solvers are employed. Experience shows that the range $\epsilon = 10^{-6}\mu^{-1}$ to $\epsilon = 10^{-9}\mu^{-1}$ is recommended [HLB]. Two questions arise. The first is what happens for non-Newtonian flows. In this case, the viscosity may vary several orders of magnitude in the fluid domain, especially if the physical properties of the material are considered to be thermally sensitive. The above rule for choosing the penalty parameter has to be applied using the smallest value of the viscosity (in order to avoid ill-conditioning), which is unknown before the calculation. Besides that, the incompressibility constraint will be excessively relaxed in the high viscosity zones. Another question to be considered is whether it-

erative solvers can be safely used or not. Usually, their convergence is very sensitive to the condition number of the stiffness matrix, which grows as the penalty parameter decreases.

The objective of this section is to present an iterative penalty finite element method whose basic motivation is alleviating the problems mentioned above. The incompressibility equation is penalized in each iteration but the residual of the previous iterate is added as a forcing term. The interesting issue is what happens when the iterative penalization is coupled with the iterative procedure due to the nonlinearity of the problem. As in Reference [Co], we analyse here what happens when the Picard and the Newton-Raphson algorithms are employed.

We will use the notation of the continuous problem. All the results also apply for its discrete finite element approximation.

3.4.1 Iterative penalization for the Stokes problem

The problem we consider in this section is (3.8) with $c = 0$, i.e., to find $\mathbf{u} \in V$ and $p \in Q_0$ such that

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) - b(p,\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ b(q,\mathbf{u}) &= 0 \qquad \forall q \in Q \end{aligned} \tag{3.70}$$

The iterative penalty method that will be analysed is particularly simple to introduce for this linear problem.

Motivation and statement of the algorithm

If the penalty method is applied to solve (3.70), one has to find $\mathbf{u}^{\epsilon(1)} \in V$ and $p^{\epsilon(1)} \in Q$ such that

$$\begin{aligned} a(\mathbf{u}^{\epsilon(1)},\mathbf{v}) - b(p^{\epsilon(1)},\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(1)},q) + b(q,\mathbf{u}^{\epsilon(1)}) &= 0 \qquad \forall q \in Q \end{aligned} \tag{3.71}$$

Once $\mathbf{u}^{\epsilon(1)}$ and $p^{\epsilon(1)}$ are found, define $\delta\mathbf{u}$ and δp such that $\mathbf{u} = \mathbf{u}^{\epsilon(1)} + \delta\mathbf{u}$ and $p = p^{\epsilon(1)} + \delta p$. Then, $\delta\mathbf{u}$ and δp will be the solution of

$$\begin{aligned} a(\delta\mathbf{u},\mathbf{v}) - b(\delta p,\mathbf{v}) &= 0 \qquad \forall \mathbf{v} \in V \\ b(q,\delta\mathbf{u}) &= \epsilon(p^{\epsilon(1)},q) \qquad \forall q \in Q \end{aligned}$$

Now, this problem can also be solved using the penalty method. If $\delta\mathbf{u}^{\epsilon}, \delta p^{\epsilon}$ is the penalized solution and we define $\mathbf{u}^{\epsilon(2)} = \mathbf{u}^{\epsilon(1)} + \delta\mathbf{u}^{\epsilon}$, $p^{\epsilon(2)} = p^{\epsilon(1)} + \delta p^{\epsilon}$, we will have that

$$\begin{aligned} a(\mathbf{u}^{\epsilon(2)},\mathbf{v}) - b(p^{\epsilon(2)},\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(2)},q) + b(q,\mathbf{u}^{\epsilon(2)}) &= \epsilon(p^{\epsilon(1)},q) \qquad \forall q \in Q \end{aligned} \tag{3.72}$$

The argument used to arrive at problem (3.72) may be applied iteratively. This leads to the following algorithm:

Given $p^{\epsilon(0)} \in Q_0$, for $i = 1, 2, \ldots$ find $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ such that

$$\begin{aligned} a(\mathbf{u}^{\epsilon(i)},\mathbf{v}) - b(p^{\epsilon(i)},\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(i)},q) + b(q,\mathbf{u}^{\epsilon(i)}) &= \epsilon(p^{\epsilon(i-1)},q) \qquad \forall q \in Q \end{aligned} \tag{3.73}$$

Existence and uniqueness of solution follows considering the problem in the space $V \times Q$ and applying Lax-Milgram's Lemma. This algorithm will be analysed below and extended to the Navier-Stokes equations. Before that, some other existing methods will be discussed.

Some remarks on related methods

Algorithm (3.73) may be viewed as a variant of the Augmented Lagrangian method (see, e.g. References [Gl], [Te]) provided that Uzawa's algorithm is used to update the pressure. This was implicitly done in early works whose basic motivation was also the computational problems encountered when small penalties are used (for example, the method was applied to linear incompressible Elasticity in Reference [Sa] and the discrete algebraic system was considered in Reference [Fe]. See also Reference [ZVT]). An important difference between our approach and the Augmented Lagrangian method is that, as will be seen below, we *will not* require the bilinear form $a(\cdot,\cdot)$ to be symmetric (although it certainly is, for the problem we consider) and thus an associated minimization problem is not needed for deriving (3.73).

The residual out-of-balance argument used to arrive at algorithm (3.73) is completely general and has physical meaning for nonlinear cases, either if the nonlinearity comes from the momentum equations (Navier-Stokes problem) or from the constitutive law of the material. The step from problem (3.72) to problem (3.73) may be applied to nonlinear problems as well, although in these cases $\delta \mathbf{u}$ and δp will be the solution of a nonlinear problem that in turn has to be linearized. Our leading idea is trying to converge in this process to the true incompressible solution. See the Appendix of Reference [CCO].

There are possibly many ways for 're-discovering' algorithm (3.73). Another one is to introduce a false transient for the pressure (not for the momentum equation) assuming the fluid to be slightly compressible and then to discretize the temporal derivative using the backward Euler scheme. Except for this discretization, this is nothing but the artificial compressibility method introduced by Chorin [Ch]. The penalty parameter ϵ in this case would be the inverse of $c^2 \Delta t$, where c is the speed of sound in the fluid and Δt the time step (see Reference [HD] for further discussion). This approach makes sense for algorithm (3.73) and for the algorithm considered in Section 3.4.4. However, it ceases to be valid when the second equation in (3.73) is coupled with a linearized form of the Navier-Stokes equations, whereas the residual argument used above can be easily extended to these cases.

Convergence of the algorithm

Before studying the convergence of the iterates of (3.73), let us state two simple results:

Lemma 3.1 *If* $q \in Q_0$ *then* $\|q\|_{Q/Z} = \|q\|$.

Proof: It follows directly from the definition of $\|\cdot\|_{Q/Z}$ and the fact that, in our case, $Z = \mathbb{R}$:

$$\begin{aligned} \|q\|_{Q/Z} &= \inf_{c \in \mathbb{R}} \|q + c\| \\ &= \inf_{c \in \mathbb{R}} \{\|q\|^2 + \|c\|^2 + 2(q,c)\}^{\frac{1}{2}} \end{aligned}$$

$$= \inf_{c \in \mathbb{R}} \{\|q\|^2 + \|c\|^2\}^{\frac{1}{2}}$$
$$= \|q\|. \qquad \square$$

This lemma will allow us to omit the subscript Q/Z when using condition (3.10). We will also need the following *a priori* estimates:

Lemma 3.2 *Let* $\mathbf{u}$ *and* p *be the solution of the Stokes problem (3.70). Then*

$$\|\mathbf{u}\| \leq \frac{N_l}{K_a} \tag{3.74}$$

$$\|p\| \leq \frac{N_l}{K_b}\left(1 + \frac{N_a}{K_a}\right) \tag{3.75}$$

Proof: Taking $\mathbf{v} = \mathbf{u}$ in (3.70) we get

$$K_a\|\mathbf{u}\|^2 \leq a(\mathbf{u},\mathbf{u}) = l(\mathbf{u}) \leq N_l\|\mathbf{u}\|$$

and (3.74) follows. On the other hand, condition (3.10) implies that there exists $\mathbf{v} \in V - \{0\}$ such that

$$\begin{aligned} K_b\|p\|\|\mathbf{v}\| &\leq b(p,\mathbf{v}) \\ &= a(\mathbf{u},\mathbf{v}) - l(\mathbf{v}) \\ &\leq (N_l + N_a\|\mathbf{u}\|)\,\|\mathbf{v}\| \end{aligned}$$

and using (3.74) we obtain (3.75). $\square$

We now proceed to the main result of this subsection.

Theorem 3.1 *Let* $(\mathbf{u},p) \in V \times Q_0$ *be the solution of the Stokes problem (3.70) and* $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ *the solution of (3.73). Define*

$$\bar{\epsilon} := \epsilon\frac{N_a^2}{K_a K_b^2}$$

If $\bar{\epsilon} < 1$ *then*

$$\lim_{i\to\infty} \|p - p^{\epsilon(i)}\| = 0, \qquad \lim_{i\to\infty} \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| = 0$$

Moreover, convergence is linear with $\bar{\epsilon}$*:*

$$\|p - p^{\epsilon(i)}\| \leq \bar{\epsilon}\,\|p - p^{\epsilon(i-1)}\| \tag{3.76}$$

$$\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \leq \bar{\epsilon}\frac{K_b}{N_a}\|p - p^{\epsilon(i-1)}\| \tag{3.77}$$

Proof: Substracting the equations of (3.70) and (3.73) one finds:

$$\begin{aligned} a(\mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{v}) - b(p - p^{\epsilon(i)}, \mathbf{v}) &= 0 \quad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(i-1)} - p^{\epsilon(i)}, q) + b(q, \mathbf{u} - \mathbf{u}^{\epsilon(i)}) &= 0 \quad \forall q \in Q \end{aligned} \tag{3.78}$$

On the other hand, we have that:

$$\begin{aligned} 0 &\leq (p - p^{\epsilon(i)}, p - p^{\epsilon(i)}) \\ &= (p^{\epsilon(i-1)} - p^{\epsilon(i)}, p - p^{\epsilon(i)}) + (p - p^{\epsilon(i-1)}, p - p^{\epsilon(i)}) \end{aligned}$$

and then

$$(p^{\epsilon(i)} - p^{\epsilon(i-1)}, p - p^{\epsilon(i)}) \leq (p - p^{\epsilon(i-1)}, p - p^{\epsilon(i)}) \tag{3.79}$$

This inequality will be used several times. Taking $\mathbf{v} = \mathbf{u} - \mathbf{u}^{\epsilon(i)}$ and $q = p - p^{\epsilon(i)}$ in (3.78) and using (3.79) we get:

$$\begin{aligned} K_a \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|^2 &\leq a(\mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{u} - \mathbf{u}^{\epsilon(i)}) \\ &= b(p - p^{\epsilon(i)}, \mathbf{u} - \mathbf{u}^{\epsilon(i)}) \\ &= \epsilon(p^{\epsilon(i)} - p^{\epsilon(i-1)}, p - p^{\epsilon(i)}) \\ &\leq \epsilon(p - p^{\epsilon(i-1)}, p - p^{\epsilon(i)}) \\ &\leq \epsilon \|p - p^{\epsilon(i-1)}\| \, \|p - p^{\epsilon(i)}\| \end{aligned} \tag{3.80}$$

Using the BB condition, there exists $\mathbf{v} \in V - \{\mathbf{0}\}$ such that

$$\begin{aligned} K_b \|p - p^{\epsilon(i)}\| \, \|\mathbf{v}\| &\leq b(p - p^{\epsilon(i)}, \mathbf{v}) \\ &= a(\mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{v}) \\ &\leq N_a \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \, \|\mathbf{v}\| \end{aligned}$$

and hence

$$\|p - p^{\epsilon(i)}\| \leq \frac{N_a}{K_b} \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \tag{3.81}$$

Combining (3.80) and (3.81) relations (3.76) and (3.77) are found. Applying inductively this inequalities, we get

$$\begin{aligned} \|p - p^{\epsilon(i)}\| &\leq \bar{\epsilon}^i \|p - p^{\epsilon(0)}\| \\ \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| &\leq \bar{\epsilon}^i \frac{K_b}{N_a} \|p - p^{\epsilon(0)}\| \end{aligned}$$

The theorem follows from the fact that $\bar{\epsilon} < 1$. □

If we take $p^{\epsilon(0)} = 0$ and apply Lemma 3.2, we see that

$$\begin{aligned} \|p - p^{\epsilon(i)}\| &\leq C_1 \bar{\epsilon}^i \\ \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| &\leq C_2 \bar{\epsilon}^i \end{aligned} \tag{3.82}$$

where the constants C_1 and C_2 are

$$\begin{aligned} C_1 &:= \frac{N_l}{K_b} \left(1 + \frac{N_a}{K_a}\right) \\ C_2 &:= \frac{N_l}{N_a} \left(1 + \frac{N_a}{K_a}\right) \end{aligned}$$

The rates of convergence (3.82) will be checked numerically in Section 3.5.

The result stated by Theorem 3.1 can also be proved for the fully discrete system and in matrix form, only requiring simple linear algebra concepts. This has been done in Reference [CCO], where it is also explained why the penalty parameter must be taken proportional to μ^{-1}.

3.4.2 A Picard-based iterative algorithm for the Navier-Stokes equations

The Stokes problem is linear and to iterate is a price to be paid if one wants to satisfy (weakly) the constraint $\nabla \cdot \mathbf{u} = 0$ up to a certain tolerance with a given penalty parameter ϵ. However, an iterative algorithm is needed for the Navier-Stokes equations and if the iteration loop could be coupled with the iterative penalization, we would have satisfied the incompressibility constraint at a low computational cost. The purpose of this section is to investigate whether this is possible or not when the Picard scheme is used to deal with the nonlinear term. Theorem 3.2 below gives sufficient conditions under which the final algorithm is convergent.

The problem to be considered now is (3.8) with c given by (3.3) (or its skew-symmetric form):

Find $(\mathbf{u}, p) \in V \times Q_0$ *such that*

$$\begin{aligned} c(\mathbf{u},\mathbf{u},\mathbf{v}) + a(\mathbf{u},\mathbf{v}) - b(p,\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ b(q,\mathbf{u}) &= 0 \qquad \forall q \in Q \end{aligned} \tag{3.83}$$

The Picard or successive substitution algorithm for this problem is:

Given $\mathbf{u}^{(0)} \in V$, *for* $i = 1, 2, \ldots$ *find* $(\mathbf{u}^{(i)}, p^{(i)}) \in V \times Q_0$ *such that*

$$\begin{aligned} c(\mathbf{u}^{(i-1)},\mathbf{u}^{(i)},\mathbf{v}) + a(\mathbf{u}^{(i)},\mathbf{v}) - b(p^{(i)},\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ b(q,\mathbf{u}^{(i)}) &= 0 \qquad \forall q \in Q \end{aligned} \tag{3.84}$$

We note that sometimes the name *Picard algorithm* is reserved for the case when all arguments of the nonlinear term are evaluated in the previous iteration (see Eqn. (3.62)). Convergence of algorithm (3.84) for any initial guess $\mathbf{u}^{(0)}$ assuming that condition (3.12) holds is a well known result (see, e.g. [CK4] for the case of strong penalty methods). The scheme we propose and analyse is the following:

Given $(\mathbf{u}^{\epsilon(0)}, p^{\epsilon(0)}) \in V \times Q_0$, *for* $i = 1, 2, \ldots$ *find* $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ *such that*

$$\begin{aligned} c(\mathbf{u}^{\epsilon(i-1)},\mathbf{u}^{\epsilon(i)},\mathbf{v}) + a(\mathbf{u}^{\epsilon(i)},\mathbf{v}) - b(p^{\epsilon(i)},\mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(i)},q) + b(q,\mathbf{u}^{\epsilon(i)}) &= \epsilon(p^{\epsilon(i-1)},q) \qquad \forall q \in Q \end{aligned} \tag{3.85}$$

Once again, existence and uniqueness of solution for (3.85) follows considering the problem in the space $V \times Q$ and applying Lax-Milgram's Lemma, since coercivity of the associated bilinear form is a consequence of (3.9) and (3.11). Before proving convergence in norm of the iterates of (3.85) to the solution of (3.83) we state the following *a priori* estimates to be used later:

Lemma 3.3 *Let* $(\mathbf{u}, p)$ *and* $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)})$ *be the solutions of (3.83) and (3.85), respectively. Then*

$$\|\mathbf{u}\| \leq \frac{N_l}{K_a} \tag{3.86}$$

$$\|\mathbf{u}^{\epsilon(i)}\| \leq \frac{N_l}{K_a} + \sqrt{\frac{\epsilon}{K_a}} \left(\|p - p^{\epsilon(i-1)}\| + \|p\| \right) \tag{3.87}$$

Proof: Estimate (3.86) is obtained exactly as (3.74) noting (3.11). To prove (3.87), take $\mathbf{v} = \mathbf{u}^{\epsilon(i)}$ and $q = p^{\epsilon(i)}$ in (3.85). We get:

$$\begin{aligned} l(\mathbf{u}^{\epsilon(i)}) &= a(\mathbf{u}^{\epsilon(i)}, \mathbf{u}^{\epsilon(i)}) + \epsilon(p^{\epsilon(i)} - p^{\epsilon(i-1)}, p^{\epsilon(i)}) \\ &= a(\mathbf{u}^{\epsilon(i)}, \mathbf{u}^{\epsilon(i)}) + \frac{\epsilon}{2}\|p^{\epsilon(i)} - p^{\epsilon(i-1)}\|^2 + \frac{\epsilon}{2}\left(\|p^{\epsilon(i)}\|^2 - \|p^{\epsilon(i-1)}\|^2\right) \\ &\geq a(\mathbf{u}^{\epsilon(i)}, \mathbf{u}^{\epsilon(i)}) - \frac{\epsilon}{2}\|p^{\epsilon(i-1)}\|^2 \end{aligned}$$

and hence

$$\begin{aligned} 2K_a\|\mathbf{u}^{\epsilon(i)}\|^2 &\leq 2N_l\|\mathbf{u}^{\epsilon(i)}\| + \epsilon\|p^{\epsilon(i-1)}\|^2 \\ &\leq \frac{N_l^2}{K_a} + K_a\|\mathbf{u}^{\epsilon(i)}\|^2 + \epsilon\|p^{\epsilon(i-1)}\|^2 \end{aligned}$$

and (3.87) follows easily applying the triangle inequality to $\|p^{\epsilon(i-1)} - p + p\|$. □

We next establish convergence of algorithm (3.85).

Theorem 3.2 *Let $(\mathbf{u}, p) \in V \times Q_0$ and $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ be the solutions of (3.83) and (3.85) respectively. Assume that*

$$\epsilon < \frac{K_a K_b^2}{N_c^2}$$

and for any $\alpha \geq 2$ define the following constants:

$$M := \left[\frac{N_l}{K_a} + \sqrt{\frac{\epsilon}{K_a}}\left(\alpha\frac{N_a}{K_b}\|\mathbf{u} - \mathbf{u}^{\epsilon(0)}\| + \|p - p^{\epsilon(0)}\| + \|p\|\right)\right]\left[1 - \sqrt{\frac{\epsilon}{K_a}}\frac{N_c}{K_b}\right]^{-1}$$

$$C_\alpha := \frac{1}{K_b}(N_c M + \alpha N_a)$$

$$\beta := \frac{1}{2} + \frac{1}{2}\left(1 + \frac{2}{\alpha}\right)^{\frac{1}{2}}$$

$$\bar{\epsilon} := \epsilon\frac{1}{K_a}C_\alpha^2$$

$$\bar{\chi} := \beta(\chi + \bar{\epsilon})$$

with χ defined in (3.12). Suppose that $\bar{\chi} < 1$ and that the initial guess for the velocity satisfies $\|\mathbf{u}^{\epsilon(0)}\| \leq M$. Then, the following holds:

$$\|\mathbf{u}^{\epsilon(i)}\| \leq M, \quad i = 1, 2, \ldots \tag{3.88}$$

$$\lim_{i\to\infty}\|p - p^{\epsilon(i)}\| = 0, \quad \lim_{i\to\infty}\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| = 0 \tag{3.89}$$

Moreover, convergence is linear with $\bar{\chi}$, that is, there exist constants C and C' such that, for i=1,2,...

$$\begin{aligned} \|p - p^{\epsilon(i)}\| &\leq C\bar{\chi}^i \\ \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| &\leq C'\bar{\chi}^i \end{aligned} \tag{3.90}$$

Proof: Only (3.90) has to be proved, since (3.89) follows from the fact that $\bar{\chi} < 1$. We proceed by induction. By hypothesis, (3.88) holds for $i = 0$. Assume that it is true up

to $i-1$, with $i \geq 1$ fixed. Substracting the equations of (3.85) from (3.83) and using the fact that for a bilinear form g we have $g(u_1, v_1) - g(u_2, v_2) = g(u_1 - u_2, v_1) + g(u_2, v_1 - v_2)$, we get:

$$\begin{aligned} c(\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}, \mathbf{u}, \mathbf{v}) + c(\mathbf{u}^{\epsilon(i-1)}, \mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{v}) & \\ + a(\mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{v}) - b(p - p^{\epsilon(i)}, \mathbf{v}) = 0 & \qquad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(i-1)} - p^{\epsilon(i)}, q) + b(q, \mathbf{u} - \mathbf{u}^{\epsilon(i)}) = 0 & \qquad \forall q \in Q \end{aligned}$$

Taking $\mathbf{v} = \mathbf{u} - \mathbf{u}^{\epsilon(i)}$ and $q = p - p^{\epsilon(i)}$ we obtain

$$a(\mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{u} - \mathbf{u}^{\epsilon(i)}) = \epsilon(p^{\epsilon(i)} - p^{\epsilon(i-1)}, p - p^{\epsilon(i)}) - c(\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}, \mathbf{u}, \mathbf{u} - \mathbf{u}^{\epsilon(i)})$$

and using the coercivity of a, (3.12) and Lemma 3.3:

$$\begin{aligned} K_a \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|^2 &\leq N_c \|\mathbf{u}\| \|\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}\| \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + \epsilon \|p - p^{\epsilon(i-1)}\| \|p - p^{\epsilon(i)}\| \\ &\leq K_a \chi \|\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}\| \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + \epsilon \|p - p^{\epsilon(i-1)}\| \|p - p^{\epsilon(i)}\| \end{aligned} \tag{3.91}$$

The BB condition implies that there exists $\mathbf{v} \in V - \{\mathbf{0}\}$ such that

$$\begin{aligned} K_b \|p - p^{\epsilon(i)}\| \, \|\mathbf{v}\| &\leq b(p - p^{\epsilon(i)}, \mathbf{v}) \\ &= a(\mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{v}) + c(\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}, \mathbf{u}, \mathbf{v}) + c(\mathbf{u}^{\epsilon(i-1)}, \mathbf{u} - \mathbf{u}^{\epsilon(i)}, \mathbf{v}) \\ &\leq \left(N_c \|\mathbf{u}\| \|\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}\| + N_c \|\mathbf{u}^{\epsilon(i-1)}\| \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + N_a \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \right) \|\mathbf{v}\| \end{aligned}$$

Using again Lemma 3.3 and (3.88) for $i-1$:

$$\|p - p^{\epsilon(i)}\| \leq \frac{K_a}{K_b} \chi \|\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}\| + \frac{N_c M}{K_b} \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + \frac{N_a}{K_b} \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \tag{3.92}$$

Inequalities (3.91) and (3.92) can be written as

$$U^{(i)^2} \leq \chi U^{(i-1)} U^{(i)} + \epsilon \frac{1}{K_a} P^{(i-1)} P^{(i)} \tag{3.93}$$

$$P^{(i)} \leq \frac{K_a}{K_b} \chi U^{(i-1)} + \left(\frac{N_c M}{K_b} + \frac{N_a}{K_b} \right) U^{(i)} \tag{3.94}$$

where we have defined

$$U^{(j)} := \|\mathbf{u} - \mathbf{u}^{\epsilon(j)}\|, \quad P^{(j)} := \|p - p^{\epsilon(j)}\|$$

for $j = i$ or $i-1$. Using (3.94) in (3.93) we get:

$$U^{(i)^2} \leq A_1 U^{(i)} + A_2 \tag{3.95}$$

where

$$\begin{aligned} A_1 &:= \chi U^{(i-1)} + \epsilon \frac{1}{K_a} P^{(i-1)} \left(\frac{N_c M}{K_b} + \frac{N_a}{K_b} \right) \\ A_2 &:= \epsilon \frac{1}{K_a} P^{(i-1)} \chi \frac{K_a}{K_b} U^{(i-1)} \end{aligned}$$

Since

$$A_1 \leq A_0 := \chi U^{(i-1)} + \epsilon \frac{1}{K_a} P^{(i-1)} C_\alpha$$

we see from (3.95) that

$$U^{(i)^2} \leq A_0 U^{(i)} + A_2 \tag{3.96}$$

A_0^2 contains the term

$$2\chi U^{(i-1)} \epsilon \frac{1}{K_a} P^{(i-1)} C_\alpha$$

and, since $K_a \leq N_a$ we get

$$\begin{aligned} A_2 &\leq \frac{\epsilon}{K_a} P^{(i-1)} \chi \frac{N_a}{K_b} U^{(i-1)} \\ &\leq \frac{\epsilon}{K_a} P^{(i-1)} \chi \frac{1}{\alpha} C_\alpha U^{(i-1)} \end{aligned}$$

Hence $A_0^2 \geq 2\alpha A_2$ and from (3.96) we obtain:

$$\begin{aligned} U^{(i)} &\leq \frac{1}{2} A_0 + \frac{1}{2} \left(A_0^2 + 4A_2\right)^{\frac{1}{2}} \\ &\leq \left[\frac{1}{2} + \frac{1}{2}\left(1 + \frac{2}{\alpha}\right)^{\frac{1}{2}}\right] A_0 \\ &= \beta A_0 \end{aligned}$$

Substitution of this inequality in (3.94) and writing the expression of A_0 leads to

$$\begin{aligned} U^{(i)} &\leq \beta\chi U^{(i-1)} + \beta\epsilon \frac{1}{K_a} C_\alpha P^{(i-1)} \\ &= \beta\chi U^{(i-1)} + \beta\bar{\epsilon} C_\alpha^{-1} P^{(i-1)} \\ P^{(i)} &\leq \frac{K_a}{K_b} \chi U^{(i-1)} + \beta\chi C_1 U^{(i-1)} + \beta\epsilon \frac{1}{K_a} C_1 C_\alpha P^{(i-1)} \\ &\leq \beta\chi C_2 U^{(i-1)} + \beta\epsilon \frac{1}{K_a} C_\alpha C_\alpha P^{(i-1)} \\ &\leq \beta\chi C_\alpha U^{(i-1)} + \beta\bar{\epsilon} P^{(i-1)} \end{aligned}$$

From this and assuming that (3.90) holds up to $i-1$ one easily gets that (3.90) is also true for the given i with the constants appearing in (3.90) $C = C_\alpha \|\mathbf{u} - \mathbf{u}^{\epsilon(0)}\| + \|p - p^{\epsilon(0)}\|$, $C' = \|\mathbf{u} - \mathbf{u}^{\epsilon(0)}\| + C_\alpha^{-1} \|p - p^{\epsilon(0)}\|$. It only remains to show that (3.88) holds for this iteration. Applying Lemma 3.3 and the fact that $P^{(i-1)} \leq C_\alpha U^{(0)} + P^{(0)}$ we obtain:

$$\begin{aligned} \|\mathbf{u}^{\epsilon(i)}\| &\leq \frac{N_l}{K_a} + \sqrt{\frac{\epsilon}{K_a}} \left[C_\alpha \|\mathbf{u} - \mathbf{u}^{\epsilon(0)}\| + \|p - p^{\epsilon(0)}\| + \|p\|\right] \\ &= \frac{N_l}{K_a} + \sqrt{\frac{\epsilon}{K_a}} \left[\left(\frac{N_c}{K_b} M + \alpha \frac{N_a}{K_b}\right) \|\mathbf{u} - \mathbf{u}^{\epsilon(0)}\| + \|p - p^{\epsilon(0)}\| + \|p\|\right] \\ &= M \end{aligned}$$

and the induction is complete. □

It is interesting to compare the result stated by this theorem with what one obtains for the Picard algorithm (3.84). First of all, in this case convergence is achieved regardless of the initial guess $\mathbf{u}^{(0)}$, whereas for (3.85) we have seen that $\mathbf{u}^{\epsilon(0)}$ has to have a norm bounded by the constant M. For practical purposes, this does not present any trouble, since one usually starts taking $\mathbf{u}^{\epsilon(0)} = \mathbf{0}$.

For (3.84) it is known that (3.90) holds with χ instead of $\bar{\chi}$. However, from the definition of C_α, β and $\bar{\epsilon}$ we see that if α is taken of order ϵ^{-q}, with $q < \frac{1}{2}$, then $\bar{\epsilon} \to 0$ and $\beta \to 1$ as $\epsilon \to 0$, that is, $\bar{\chi} \to \chi$. So, for ϵ small the convergence of algorithm (3.85) is the same as that of (3.84). One can obtain α such that $\bar{\chi}$ be minimized (under the restriction $\alpha \geq 2$). For example, taking the norms and the coercivity constants of the forms involved in the problem equal to unity, one obtains that the optimal condition for achieving convergence is $\chi < 0.7511$ for $\epsilon = 10^{-1}$ and $\chi < 0.9744$ for $\epsilon = 10^{-4}$. The fact that, in any case, $\bar{\chi} \geq \chi$ is the cost of converging to the true incompressible solution using a penalized scheme.

3.4.3 A Newton-Raphson-based iterative algorithm for the Navier-Stokes equations

The objective of this section is to analyse the algorithm obtained when the Newton-Raphson scheme is coupled with the iterative penalization in the sense used previously for the Picard scheme. Once again, we will see that the usual convergence requirements of the Newton-Raphson method have to be slightly restricted. In this case, there is another important issue to be considered. It is well known that convergence is quadratic for Newton-Raphson's iterates. The question is whether this rate of convergence will be inherited by the scheme we propose. The answer is that this is true up to a certain iteration. From there onwards, the convergence rate will only be linear. However, the numerical experiments we have performed, some of which are presented in Section 3.5, indicate that the situation is not so bad as it might seem. For small penalty parameters, usually much larger than those used in classical penalty methods, convergence is achieved before its rate turns from quadratic to linear.

The Newton-Raphson algorithm applied to problem (3.83) reads as follows:

Given $\mathbf{u}^{(0)} \in V$, *for* $i = 1, 2, \ldots$ *find* $(\mathbf{u}^{(i)}, p^{(i)}) \in V \times Q_0$ *such that*

$$\begin{aligned} c(\mathbf{u}^{(i-1)}, \mathbf{u}^{(i)}, \mathbf{v}) + c(\mathbf{u}^{(i)}, \mathbf{u}^{(i-1)}, \mathbf{v}) + a(\mathbf{u}^{(i)}, \mathbf{v}) - b(p^{(i)}, \mathbf{v}) & \\ = c(\mathbf{u}^{(i-1)}, \mathbf{u}^{(i-1)}, \mathbf{v}) + l(\mathbf{v}) & \quad \forall \mathbf{v} \in V \\ b(q, \mathbf{u}^{(i)}) = 0 & \quad \forall q \in Q \end{aligned} \tag{3.97}$$

That this algorithm is convergent if the initial guess is sufficiently close to the exact solution and that convergence is quadratic is a well known result. In [GR], this is proved when the solution of (3.84) belongs to a nonsingular branch. The modified algorithm we will consider is the following:

Given $(\mathbf{u}^{\epsilon(0)}, p^{\epsilon(0)}) \in V \times Q_0$, *for* $i = 1, 2, \ldots$ *find* $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ *such that*

$$\begin{aligned} c(\mathbf{u}^{\epsilon(i-1)}, \mathbf{u}^{\epsilon(i)}, \mathbf{v}) + c(\mathbf{u}^{\epsilon(i)}, \mathbf{u}^{\epsilon(i-1)}, \mathbf{v}) + a(\mathbf{u}^{\epsilon(i)}, \mathbf{v}) - b(p^{\epsilon(i)}, \mathbf{v}) & \\ = c(\mathbf{u}^{\epsilon(i-1)}, \mathbf{u}^{\epsilon(i-1)}, \mathbf{v}) + l(\mathbf{v}) & \quad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(i)}, q) + b(q, \mathbf{u}^{\epsilon(i)}) = \epsilon(p^{\epsilon(i-1)}, q) & \quad \forall q \in Q \end{aligned} \tag{3.98}$$

Our analysis will be based on the assumption that (3.12) holds.

Theorem 3.3 *Let* $(\mathbf{u},p) \in V \times Q_0$ *and* $(\mathbf{u}^{\epsilon(i)},p^{\epsilon(i)}) \in V \times Q_0$ *be the solutions of (3.83) and (3.98) respectively. Let* $\alpha \geq 2$ *be given and define the following constants:*

$$C := \frac{2N_c}{K_a(1-\chi)}$$
$$C_\alpha := \frac{K_a}{2K_b}(1+3\chi) + \alpha\frac{N_a}{K_b}$$
$$\beta := \frac{1}{2} + \frac{1}{2}\left(1+\frac{2}{\alpha}\right)^{\frac{1}{2}}$$
$$\bar{\epsilon} := \epsilon\frac{CC_\alpha^2}{N_c}$$

Assume that the following conditions are satisfied:

$$(H1) \qquad \|\mathbf{u}-\mathbf{u}^{\epsilon(0)}\| < \frac{1}{C\beta}\frac{\sigma}{\gamma}, \quad \text{with} \quad \sigma<1, \ \gamma>1$$

$$(H2) \qquad \bar{\epsilon} < \frac{\gamma-1}{\beta\gamma}\sigma$$

$$(H3) \qquad \bar{\epsilon}\|p-p^{\epsilon(0)}\| < \frac{\gamma-1}{\beta^2\gamma^2}\frac{C_\alpha}{C}\sigma^2 \quad \text{or} \quad \|p-p^{\epsilon(0)}\| < \frac{C_\alpha}{C\beta}\frac{\sigma}{\gamma}$$

Under these conditions, we have that, for $i = 1,2,...$

$$(T1) \qquad \|\mathbf{u}-\mathbf{u}^{\epsilon(i)}\| < \frac{1}{C\beta}\frac{\sigma}{\gamma}, \quad \|p-p^{\epsilon(i)}\| < \frac{C_\alpha}{C\beta}\frac{\sigma}{\gamma}$$

$$(T2) \qquad K_a - N_c\|\mathbf{u}^{\epsilon(i)}\| > \frac{K_a}{2}(1-\chi)$$

$$(T3) \qquad \lim_{i\to\infty}\|p-p^{\epsilon(i)}\| = 0, \quad \lim_{i\to\infty}\|\mathbf{u}-\mathbf{u}^{\epsilon(i)}\| = 0$$

Moreover, if for a certain I

$$(H4) \qquad \bar{\epsilon} < \frac{\gamma-1}{\beta\gamma}\sigma^{2^I-1}$$

then convergence is quadratic up to this I*:*

$$(T4) \qquad \|\mathbf{u}-\mathbf{u}^{\epsilon(i)}\| < \frac{1}{C\beta}\frac{\sigma^{2^i}}{\gamma}, \quad \|p-p^{\epsilon(i)}\| < \frac{C_\alpha}{C\beta}\frac{\sigma^{2^i}}{\gamma}, \quad 1 \leq i \leq I$$

Proof: We proceed by induction. For $i = 0$, $(T1)$ is precisely $(H1)$ and $(H3)$ if the second option in this hypothesis is taken. In fact, for $i = 1,2,...$ we will see that any of the two possibilities in $(H3)$ are sufficient for proving $(T1)$. On the other hand, $(T2)$ for $i = 0$ follows from $(H1)$ and (3.86):

$$\begin{aligned} K_a - N_c\|\mathbf{u}^{\epsilon(0)}\| &\geq K_a - N_c\|\mathbf{u}-\mathbf{u}^{\epsilon(0)}\| - N_c\|\mathbf{u}\| \\ &> K_a - \frac{N_c}{\beta}\frac{K_a(1-\chi)}{2N_c} - K_a\chi \\ &> \frac{1}{2}K_a(1-\chi) \end{aligned} \tag{3.99}$$

since $\beta > 1$. Now, let i be fixed and assume $(T1)$ and $(T2)$ hold up to $i-1$. In order to prove existence and uniqueness of solution for (3.98), let us write this problem as follows: Find $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ such that

$$\mathcal{B}_{i-1}(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}; \mathbf{v}, q) = \mathcal{L}_{i-1}(\mathbf{v}, q) \quad \forall (\mathbf{v}, q) \in V \times Q$$

where

$$\begin{aligned}\mathcal{B}_{i-1}(\mathbf{u}, p; \mathbf{v}, q) :&= c(\mathbf{u}^{\epsilon(i-1)}, \mathbf{u}, \mathbf{v}) + c(\mathbf{u}, \mathbf{u}^{\epsilon(i-1)}, \mathbf{v}) + a(\mathbf{u}, \mathbf{v}) \\ &\quad - b(p, \mathbf{v}) + \epsilon(p, q) + b(q, \mathbf{u}) \\ \mathcal{L}_{i-1}(\mathbf{v}, q) &:= l(\mathbf{v}) + c(\mathbf{u}^{\epsilon(i-1)}, \mathbf{u}^{\epsilon(i-1)}, \mathbf{v}) + \epsilon(p^{\epsilon(i-1)}, q)\end{aligned}$$

If we prove that $\mathcal{B}_{i-1}$ is coercive in $V \times Q$, existence and uniqueness will follow from Lax-Milgram's Lemma. We see that

$$\begin{aligned}\mathcal{B}_{i-1}(\mathbf{v}, q; \mathbf{v}, q) &= c(\mathbf{v}, \mathbf{u}^{\epsilon(i-1)}, \mathbf{v}) + a(\mathbf{v}, \mathbf{v}) + \epsilon(q, q) \\ &\geq (K_a - N_c\|\mathbf{u}^{\epsilon(i-1)}\|)\|\mathbf{v}\|^2 + \epsilon\|q\|^2 \\ &\geq \min\{K_a - N_c\|\mathbf{u}^{\epsilon(i-1)}\|, \epsilon\}(\|\mathbf{v}\|^2 + \|q\|^2)\end{aligned}$$

and the fact that $K_a - N_c\|\mathbf{u}^{\epsilon(i-1)}\| > 0$ is a consequence of $(T2)$ used inductively.

Applying the same arguments as in Theorem 3.2 to arrive at (3.91) and (3.92) we now obtain:

$$\begin{aligned}K_a\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|^2 &\leq N_c\|\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}\|^2\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \\ &\quad + N_c\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|^2\|\mathbf{u}^{\epsilon(i-1)}\| + \epsilon\|p - p^{\epsilon(i-1)}\|\|p - p^{\epsilon(i)}\| \qquad (3.100) \\ K_b\|p - p^{\epsilon(i)}\| &\leq N_c\|\mathbf{u}\|\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + N_c\|\mathbf{u} - \mathbf{u}^{\epsilon(i-1)}\|^2 \\ &\quad + N_c\|\mathbf{u}^{\epsilon(i-1)}\|\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + N_a\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \qquad (3.101)\end{aligned}$$

If we call

$$\begin{aligned}&U^{(j)} := \|\mathbf{u} - \mathbf{u}^{\epsilon(j)}\|, \quad P^{(j)} := \|p - p^{\epsilon(j)}\|, \quad j = i \text{ or } i-1 \\ &A_1 := K_a - N_c\|\mathbf{u}^{\epsilon(i-1)}\| \\ &A_2 := N_c U^{(i-1)^2} + \epsilon P^{(i-1)}\left(\frac{N_c}{K_b}\|\mathbf{u}\| + \frac{N_c}{K_b}\|\mathbf{u}^{\epsilon(i-1)}\| + \frac{N_a}{K_b}\right) \\ &A_3 := \epsilon P^{(i-1)}\frac{N_c}{K_b}U^{(i-1)^2}\end{aligned}$$

we find from the previous inequalities that

$$A_1 U^{(i)^2} \leq A_2 U^{(i)} + A_3 \qquad (3.102)$$

From $(T2)$ it is $A_1 > 0$. Note now that, from $(T1)$ and (3.86):

$$\begin{aligned}\frac{N_c}{K_b}\|\mathbf{u}\| + \frac{N_c}{K_b}\|\mathbf{u}^{\epsilon(i-1)}\| + \frac{N_a}{K_b} &< \frac{N_c}{K_b}\|\mathbf{u}\| + \frac{N_c}{K_b}\frac{1}{C} + \frac{N_c}{K_b}\|\mathbf{u}\| + \frac{N_a}{K_b} \\ &= \frac{2K_a}{K_b}\chi + \frac{N_a}{K_b} + \frac{K_a(1-\chi)}{2K_b} \\ &< C_a\end{aligned}$$

and so we have

$$A_2 < A_0 := N_c U^{(i-1)^2} + \epsilon P^{(i-1)} C_\alpha$$

A_0^2 contains the term

$$2N_c U^{(i-1)^2} \epsilon P^{(i-1)} \alpha \frac{N_a}{K_b}$$

Since $K_a \leq N_a$ we have that

$$2A_1 A_3 < 2K_a \epsilon P^{(i-1)} \frac{N_c}{K_b} U^{(i-1)^2} < \frac{1}{\alpha} A_0^2$$

and hence, from (3.102) and using that $A_2 < A_0$ and $(T2)$:

$$\begin{aligned} U^{(i)} &< \frac{1}{2}\frac{A_0}{A_1} + \frac{1}{2A_1}\left(A_0^2 + 4A_1A_3\right)^{\frac{1}{2}} \\ &< \frac{1}{2}\frac{A_0}{A_1} + \frac{1}{2A_1}\left(1+\frac{2}{\alpha}\right)^{\frac{1}{2}} A_0 \\ &= \beta \frac{A_0}{A_1} \\ &< \frac{2\beta}{K_a(1-\chi)}\left(N_c U^{(i-1)^2} + \epsilon P^{(i-1)} C_\alpha\right) \\ &= C\beta U^{(i-1)^2} + \epsilon\beta \frac{C}{N_c} C_\alpha P^{(i-1)} \end{aligned}$$

On the other hand, for the pressures we obtain from (3.101) that:

$$\begin{aligned} P^{(i)} &\leq \frac{N_c}{K_b} U^{(i-1)^2} + \left(\frac{N_c}{K_b}\|\mathbf{u}^{\epsilon(i-1)}\| + \frac{N_c}{K_b}\|\mathbf{u}\| + \frac{N_a}{K_b}\right) U^{(i)} \\ &< \frac{N_c}{K_b} U^{(i-1)^2} + C_1 U^{(i)} \\ &< \frac{N_c}{K_b} U^{(i-1)^2} + C_1 C\beta U^{(i-1)^2} + \epsilon\beta C_\alpha \frac{C_1}{N_c} P^{(i-1)} \end{aligned}$$

Noting that

$$\frac{N_a}{K_b}\beta C = \beta \frac{N_a}{K_b}\frac{2N_c}{K_a(1-\chi)} > \frac{N_c}{K_b}$$

we finally obtain that (3.100) and (3.101) imply:

$$U^{(i)} < C\beta U^{(i-1)^2} + \bar{\epsilon}\beta C_\alpha^{-1} P^{(i-1)} \tag{3.103}$$

$$P^{(i)} < CC_\alpha \beta U^{(i-1)^2} + \bar{\epsilon}\beta P^{(i-1)} \tag{3.104}$$

Now, using $(H2)$ and $(T1)$ for $i-1$ one gets:

$$\begin{aligned} U^{(i)} &< C\beta \frac{1}{C^2\beta^2}\frac{\sigma^2}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma \frac{C_\alpha^{-1} C_\alpha}{C\beta}\frac{\sigma}{\gamma} \\ &= \frac{1}{C\beta}\frac{\sigma^2}{\gamma} \\ P^{(i)} &< CC_\alpha\beta \frac{1}{C^2\beta^2}\frac{\sigma^2}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma \frac{C_\alpha}{C\beta}\frac{\sigma}{\gamma} \\ &= \frac{C_\alpha}{C\beta}\frac{\sigma^2}{\gamma} \end{aligned}$$

Since $\sigma < 1$, the induction for $(T1)$ is closed. Observe that either of the two possibilities in $(H3)$ suffices for proving this part of the thesis. $(T2)$ is obtained from $(T1)$ using the same steps as in (3.99). In order to prove $(T3)$, from (3.103) and (3.104) we see that the required condition is:

$$\theta^{(i)} := C\beta U^{(i)} + \bar{\epsilon}\beta < 1$$

From $(T1)$ and $(H2)$ it follows that $\theta^{(i)} < \sigma < 1$, so convergence of the algorithm is ensured. Inequalities (3.103) and (3.104) show that the rate of convergence will be at least linear.

Now, suppose that $(H4)$ holds. For $1 \leq i \leq I$ we obtain from (3.103) and (3.104) and assuming $(T4)$ to be true up to $i-1$, that:

$$\begin{aligned} U^{(i)} &< C\beta \frac{1}{C^2\beta^2}\frac{\sigma^{2^i}}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma^{2^{i-1}}\frac{C_\alpha^{-1}C_\alpha}{C\beta}\frac{\sigma^{2^{i-1}}}{\gamma} \\ &= \frac{1}{C\beta}\frac{\sigma^{2^i}}{\gamma} \\ P^{(i)} &< CC_\alpha\beta\frac{1}{C^2\beta^2}\frac{\sigma^{2^i}}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma^{2^{i-1}}\frac{C_\alpha}{C\beta}\frac{\sigma^{2^{i-1}}}{\gamma} \\ &= \frac{C_\alpha}{C\beta}\frac{\sigma^{2^i}}{\gamma} \end{aligned}$$

This proves $(T4)$ and completes the proof of the theorem. □

This theorem states that if the initial guess is close enough to the final solution and ϵ is sufficiently small, algorithm (3.98) will converge. The situation is similar for the standard Newton-Raphson scheme (3.97). The only difference in the requirement for the initial velocity guess is that $(H1)$ has to hold but with $\beta = 1$. Nevertheless, the same remarks as in Theorem 3.2 apply and in our case α can be taken such that $\beta \to 1$ when $\epsilon \to 0$. Another observation is that in $(H3)$ we can choose either having a 'good' initial pressure guess or limiting the value of ϵ.

3.4.4 Uncoupling of the iterative penalization

In Sections 3.4.2 and 3.4.3 we have seen that if the iterative penalization is coupled with the iterative scheme used to deal with the convective term of the Navier-Stokes equations, the conditions under which this scheme converges have to be restricted. There is also the possibility of uncoupling the iteration due to the nonlinearity of the equations and the imposition on the incompressibility constraint. The purpose of this section is the analysis of the following algorithm:

Given $p^{\epsilon(0)} \in Q_0$, *for* $i = 1, 2, \ldots$ *find* $(\mathbf{u}^{\epsilon(i)}, p^{\epsilon(i)}) \in V \times Q_0$ *such that*

$$\begin{aligned} c(\mathbf{u}^{\epsilon(i)}, \mathbf{u}^{\epsilon(i)}, \mathbf{v}) + a(\mathbf{u}^{\epsilon(i)}, \mathbf{v}) - b(p^{\epsilon(i)}, \mathbf{v}) &= l(\mathbf{v}) \qquad \forall \mathbf{v} \in V \\ \epsilon(p^{\epsilon(i)}, q) + b(q, \mathbf{u}^{\epsilon(i)}) &= \epsilon(p^{\epsilon(i-1)}, q) \qquad \forall q \in Q \end{aligned} \tag{3.105}$$

For a given i, existence and uniqueness of solution is a consequence of assumption (3.12) [GR]. For each iteration of this algorithm a nonlinear problem has to be solved.

We will assume that the solution found in this process is exact. In this case, we obtain a result similar to the one encountered for the Stokes problem in Section 3.4.1:

Theorem 3.4 *Let* $(\mathbf{u},p) \in V \times Q_0$ *and* $(\mathbf{u}^{\epsilon(i)},p^{\epsilon(i)}) \in V \times Q_0$ *be the solutions of (3.83) and (3.105), respectively. Define the following constants:*

$$
\begin{aligned}
M &:= \frac{N_l}{K_a} + \sqrt{\frac{\epsilon}{K_a}}(\|p - p^{\epsilon(0)}\| + \|p\|) \\
C_1 &:= K_a(1-\chi) \\
C_2 &:= \frac{K_a}{K_b}\chi + \frac{N_c}{K_b}M + \frac{N_a}{K_b} \\
\bar{\epsilon} &:= \frac{\epsilon}{C_1}C_2^2
\end{aligned}
$$

If $\bar{\epsilon} < 1$ *then:*

$$\|\mathbf{u}^{\epsilon(i)}\| \leq M, \quad i = 1,2,... \tag{3.106}$$

$$\lim_{i\to\infty} \|p - p^{\epsilon(i)}\| = 0, \quad \lim_{i\to\infty} \|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| = 0 \tag{3.107}$$

Moreover, convergence is linear with $\bar{\epsilon}$*:*

$$
\begin{aligned}
\|p - p^{\epsilon(i)}\| &\leq \bar{\epsilon}^i \|p - p^{\epsilon(0)}\| \\
\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| &\leq C_2^{-1}\bar{\epsilon}^i \|p - p^{\epsilon(0)}\|
\end{aligned}
\tag{3.108}
$$

Proof: First observe that (3.87) holds for the solution $\|\mathbf{u}^{\epsilon(i)}\|$ of (3.105). This can be proved exactly as in Lemma 3.3. Thus, (3.106) is verified for $i = 1$. Let $i > 1$ be given and assume this is true up to this iteration. If the ideas used to arrive at (3.91) and (3.92) in Theorem 3.2 are now applied, one finds:

$$K_a\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|^2 \leq N_c\|\mathbf{u}\|\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|^2 + \epsilon\|p - p^{\epsilon(i-1)}\|\|p - p^{\epsilon(i)}\| \tag{3.109}$$

$$
\begin{aligned}
K_b\|p - p^{\epsilon(i)}\| \leq\ & N_c\|\mathbf{u}\|\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| + N_c\|\mathbf{u}^{\epsilon(i)}\|\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| \\
& + N_a\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\|
\end{aligned}
\tag{3.110}
$$

Combining inequalities (3.109) and (3.110), using the estimates (3.86) for $\mathbf{u}$ and (3.106) for $\mathbf{u}^{\epsilon(i)}$ and considering the definition of the constants in the statement of the theorem, we arrive at:

$$
\begin{aligned}
\|p - p^{\epsilon(i)}\| &\leq \bar{\epsilon}\|p - p^{\epsilon(i-1)}\| \\
\|\mathbf{u} - \mathbf{u}^{\epsilon(i)}\| &\leq \bar{\epsilon}C_2^{-1}\|p - p^{\epsilon(i-1)}\|
\end{aligned}
$$

from where (3.108) follows. Since $\bar{\epsilon} < 1$ we have that $\|p - p^{\epsilon(i)}\| < \|p - p^{\epsilon(0)}\|$ and we obtain from (3.87) that (3.106) holds for $i + 1$. This closes the induction. Finally, (3.108) implies (3.107) for $\bar{\epsilon} < 1$. □

3.5 Numerical examples

The three examples presented in this section concern the numerical behavior of the iterative penalty method analysed heretofore. The implementation of the scheme is treated in detail in Chapter 4, where the calculation of the pressure, the streamfunction and the vorticity $\boldsymbol{\omega} := \nabla \times \mathbf{u}$ is described in detail. Only results using the Q_2/P_1 will be shown. Similar answers are obtained using other elements that satisfy the BB condition (see Chapter 4).

Convergence has been checked only in velocities, using the norm of the residual over the norm of the last iterate as the parameter to decide whether this convergence has been achieved or not (cf. Eqn. (3.64)). In the figures presented thereafter, this parameter in % is what is called *residual.* Since we are mainly interested in the satisfaction of the incompressibility constraint, also the L^2 norm of the discrete divergence of the velocity has been computed.

All the calculations have been carried out on a CONVEX-C120 computer using double arithmetic precision.

Example 3.1 *Two-dimensional driven cavity flow*

In this example, the Stokes problem with $\mu = 1$ in the unit square $[0,1] \times [0,1]$ has been solved. The boundary conditions have been taken as $\mathbf{u} = (1,0)$ for $y = 1, 0 \leq x \leq 1$ (x, y being the Cartesian coordinates) and $\mathbf{u} = (0,0)$ on the rest of the boundary (leaky-lid conditions). External body forces have been taken zero. The domain has been discretized using a uniform mesh of 25×25 nodal points (12×12 elements). The velocity vectors, streamlines, pressure contours and vorticity contours for this problem have been plotted in Figure 3.6. Figure 3.7 shows the convergence history for the values of the penalty parameter $\epsilon = 10^{-1}$, 10^{-2}, 10^{-3} and 10^{-4}. Observe that the difference in the slope of the curves agrees with the theoretical prediction (3.82).

Once the finite element discretization has been performed, the term $\nabla \cdot \mathbf{u}$ leads to $\mathbf{BU}$, where $\mathbf{B} = \mathbf{G}^T$ is the discrete divergence matrix (cf. Box 3.2). In order to study the convergence of the iterates to the incompressible solution, the norm of $\mathbf{BU}^{\epsilon(i)}$ has been computed. Figure 3.8 shows the results obtained for different values of the penalty parameter. The curves correspond to 1, 2 and 3 iterations in the algorithm (3.73). Once again, their relative slope agrees with what (3.82) predicts. Observe that $\|\mathbf{BU}^{\epsilon(i)}\|$ will be bounded by $\epsilon G \|p^{\epsilon(i)} - p^{\epsilon(i-1)}\|$, where G is the norm of the Gramm matrix $\mathbf{M}_p$ given in Box 3.2 whose components are the scalar products of the basis functions for the pressure (see (3.73)).

Example 3.2 *Three-dimensional driven cavity flow*

The numerical simulation of three dimensional flows is a challenge, even for simple problems, mainly because of the large amount of computer memory required. One may afford long CPU times in research environments, but the real problem is to make the problem fit in the limits of the available computer memory.

Iterative solvers for algebraic linear systems have the important feature of being much less memory demanding than direct methods. On the other hand, they are very sensitive to the condition number of the system matrix. The number of iterations to be performed to achieve convergence is highly increased when this condition number grows.

In this example we present some preliminary results obtained for the Stokes flow in a 3D cavity using the iterative penalty method and solving the algebraic system

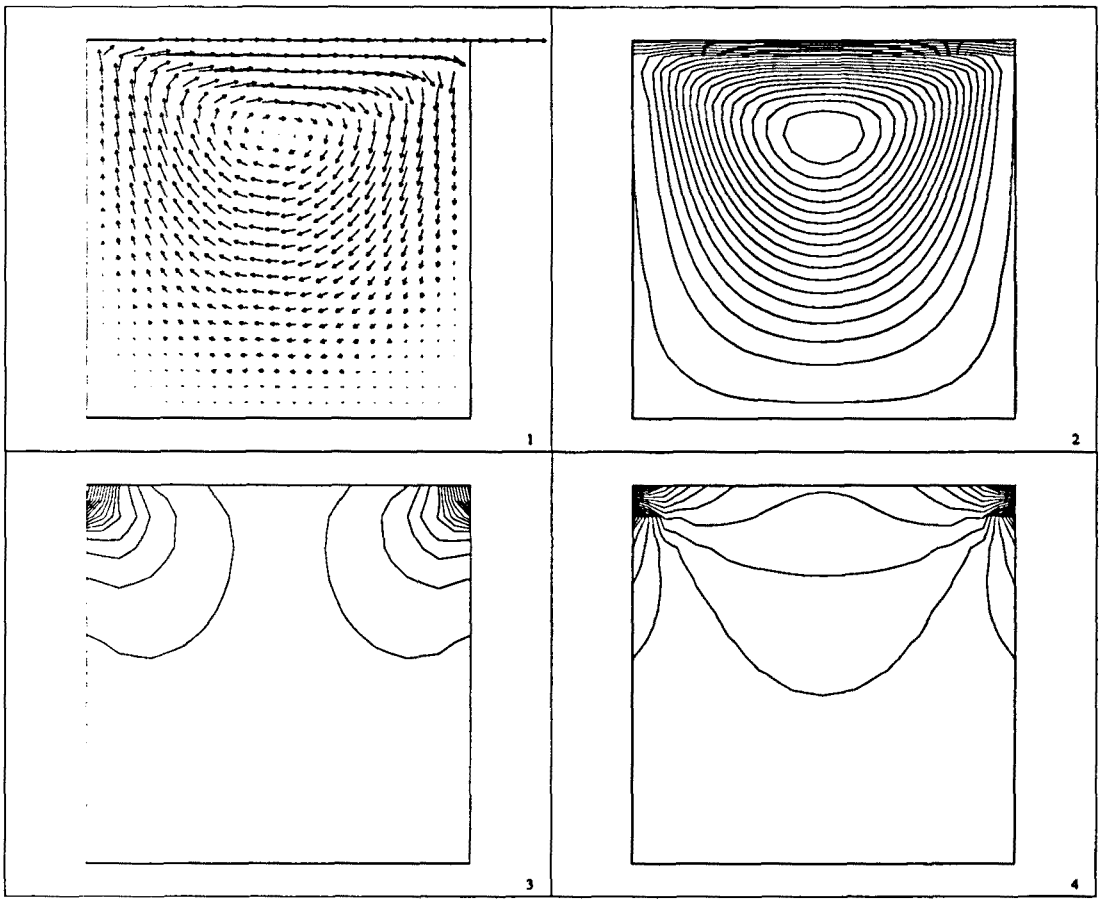

Figure 3.6 Numerical solution of the two-dimensional cavity flow problem. (1): Velocity vectors; (2): Streamfunction contours; (3): Pressure contours; (4): Vorticity contours.

using the conjugate gradient algorithm (see Reference [RTF] for similar experiments). The domain is the unit cube $[0,1]^3$ discretized using a uniform mesh of $21 \times 21 \times 21$ nodal points ($10 \times 10 \times 10$ Q_2/P_1 elements). The boundary conditions are $\mathbf{u} = (0,1,0)$ for $y = 1$, $0 \leq x \leq 1$, $0 \leq z \leq 1$ and $\mathbf{u} = \mathbf{0}$ on the rest of the boundary. The viscosity has been taken $\mu = 1$ and body forces are zero. Pressure contours are shown in Figure 3.9.

The total memory required for this problem has been 81.27 Mb (Mega-bytes), a considerable figure if we consider that it has been run on a computer with 128 Mb of central memory. Most of this memory (78.91 Mb) has been needed to allocate the element arrays (shape functions, derivatives, element matrices, etc.). The memory required for the conjugate gradient solver has been only 149 Kb (Kilo-bytes).

It is clear that the smaller ϵ be, the larger will be the condition number of the stiffness matrix and therefore less congugate gradient (CG) iterations will be needed, but more iterative penalization (IP) iterations will be required to reach a prescribed convergence tolerance. Here, the tolerance for the CG algorithm has been taken as

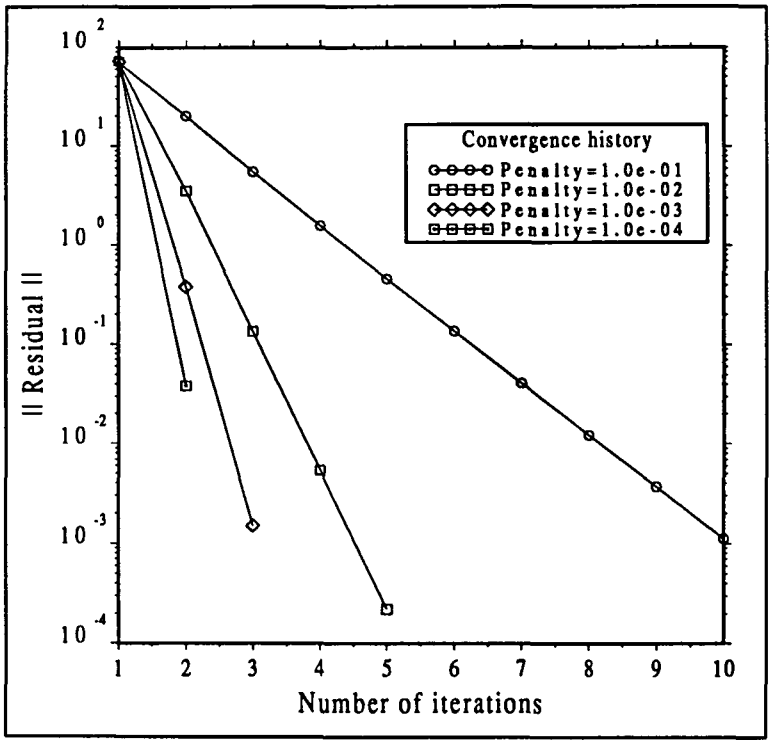

Figure 3.7 Convergence history for the two-dimensional cavity flow example using different penalty parameters

10^{-6} and the convergence tolerance as 10^{-3} %. The results obtained are the following:

Penalty	CG iterations	IP iterations	CPU (seconds)
10^{-2}	418	6	8778
10^{-4}	2913	3	25924
10^{-6}	6303	2	36594

The CPU times are only referred to the solver algorithm. From these results it is apparent that it is worth using a 'large' penalty parameter, since the final CPU time is smaller, thanks to the small number of CG iterations. This compensates the fact that more iterations have to be performed to satisfy the incompressibility constraint up to the prescribed tolerance.

The number of CG and IP iterations and the CPU time have been plotted in Figure 3.10, as well as the convergence history for the values of ϵ considered.

Example 3.3 *Flow over a backward-facing step*

The purpose of this example is to present some numerical results concerning the algorithms studied in Sections 3.4.2 and 3.4.3 for the incompressible Navier-Stokes equations. We have chosen this well known benchmark problem because a large number of numerical results are available. The computational domain we have taken is the rectangle $[0, 22] \times [0, 1.5]$ with a step of length 3 and height 0.5 placed in the lower left corner. A detail of the mesh used in the calculation is shown in Figure 3.11.(1). This

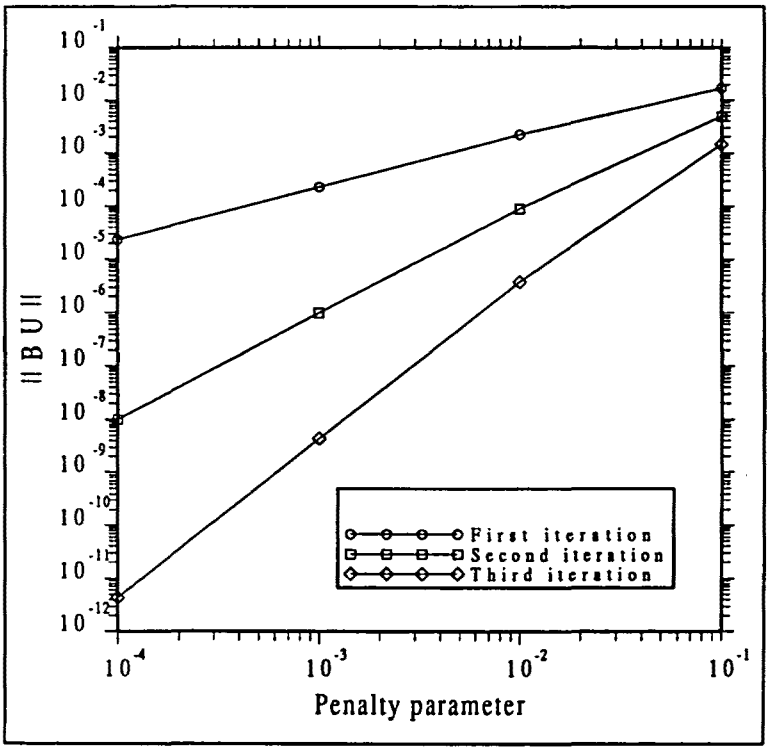

Figure 3.8 Norm of the discrete divergence for the 2D cavity flow example for different number of iterations.

mesh is composed of 408 biquadratic elements (for the velocity interpolation) and 1721 nodal points.

On the left boundary $x = 0$ a parabolic velocity profile with maximum value $(1, 0)$ has been prescribed. The viscosity has been taken as $\mu = 0.005$ and the density $\rho = 1$. Thus, the Reynolds number based on the inflow profile and the step height is $Re = 100$. The outflow boundary $x = 22$, $0 < y < 1.5$ has been left free. We have employed the expression (3.4) for the viscous term. In this case, the associated natural boundary condition is zero traction. On the rest of the boundary, the no-slip condition $\mathbf{u} = \mathbf{0}$ has been imposed. External body forces are zero.

The computed pressure contours and a detail of the streamlines are plotted in Figures 3.11.(2) to 3.11.(4). These results have been obtained using a penalty parameter $\epsilon = 10^{-4}$ and with a tolerance of $10^{-4}\%$. The iterative scheme employed has been (3.98). Now we discuss the performance of the algorithms (3.85) and (3.98) for this problem when the classical penalty method and the iterative penalization proposed in this chapter are used. In the former case, the right-hand-side in the second equation of both (3.85) and (3.98) is zero.

Consider first algorithm (3.85). Figure 3.12.(1) shows the convergence history in the discrete L^2 norm when both the classical and the iterative penalty methods are used. No difference can be observed in the plot even though the parameter that defines convergence in the former case is χ whereas in the latter it is $\bar{\chi} > \chi$ (see (3.90)). The values of the relative norm of the residual in iteration number 18 are 0.87215×10^{-2} for the classical penalty method and 0.87108×10^{-2} for the iterative penalization. However, the important issue is the evolution of $\|\mathbf{B}\mathbf{U}^{\epsilon(i)}\|$ shown in Fig-

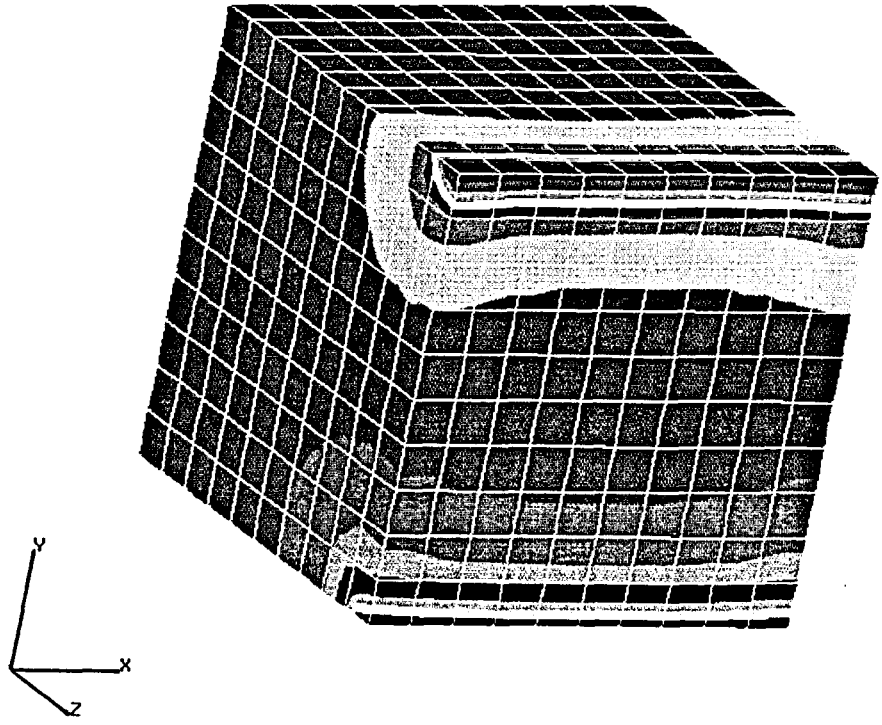

Figure 3.9 Pressure contours for the 3D cavity flow problem.

ure 3.12.(2). For the classical penalty method this norm remains constant (and, as expected, of order ϵ). On the other hand, the velocity solution of algorithm (3.85) converges linearly to a (weakly) solenoidal field. The same experiments have been performed using the Newton-Raphson-based algorithm (3.98). If the initial guess is taken as $\mathbf{u}^{\epsilon(0)} = 0$, $p^{\epsilon(0)} = 0$ the scheme does not converge. In order to obtain a good initial guess, at least two iterations of algorithm (3.85) have to be performed, both for the classical and the iterative penalty methods. For $\epsilon = 10^{-4}$, the convergence history of the two methods shown in Figure 3.12.(3) is the same. The relative norm of the residual reaches the value 0.9986×10^{-8} at iteration number 7 for the classical penalty method and 0.9781×10^{-8} if (3.98) is used. The evolution of the norm of the discrete divergence (Figure 3.12.(4)) is certainly very different for the two methods. Whereas the penalty method yields a constant value, the iterative penalization converges quadratically to a zero divergence velocity field.

Similar results are obtained when the penalty parameter is $\epsilon = 10^{-1}$. Figures 3.13.(1) and 3.13.(2) show the convergence history and the evolution of $\|\mathbf{B}\mathbf{U}^{\epsilon(i)}\|$ for the Picard-based algorithm (3.85). It is interesting to observe that for this large penalty number the residual norm using (3.85) is only slightly larger than using the classical penalty method. For the Newton-Raphson-based method, results are shown in Figures 3.13.(3) and 3.13.(4) and are especially interesting. The convergence rate for the iterates of (3.98) (see Figure 3.13.(3)) is quadratic up to iteration number 6 (except for the two first iterations, in which scheme (3.85) has been used). From there on, this rate turns to be linear. This possibility was already predicted in Theorem 3.3. The classical penalty method has a global quadratic convergence rate, but $\|\mathbf{B}\mathbf{U}^{\epsilon(i)}\|$ keeps constant in the iterative process (Figure 3.13.(4)) at an unacceptable value.

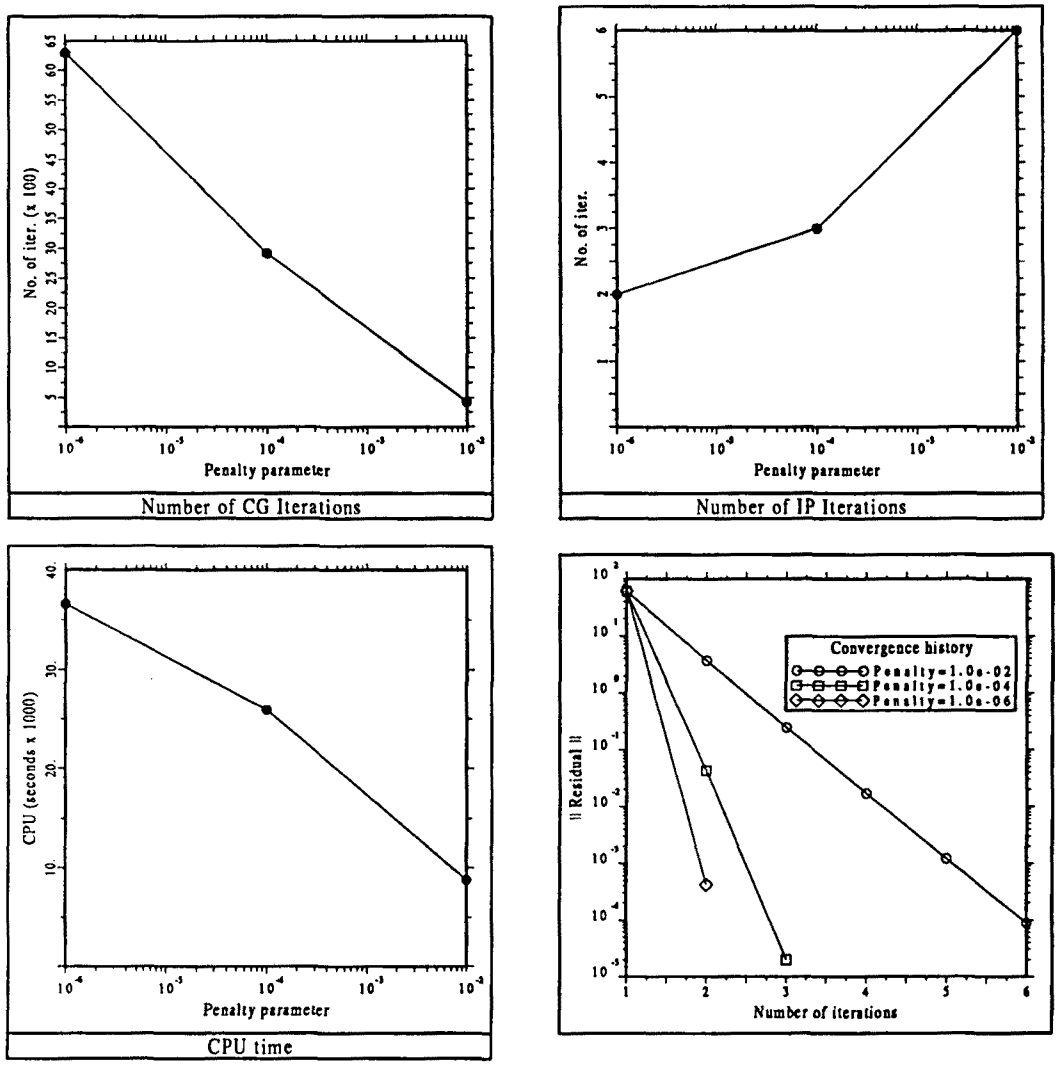

Figure 3.10 Results for the 3D cavity flow problem.

3.6 Summary and conclusions

The interest of this chapter has been focussed on the use of penalty methods to deal with the incompressibility constraint of the Navier-Stokes equations, with special reference to their computational behavior. We have started considering the strong penalty method and the related selective underintegration (RIP method), since this approach seems simpler than the weak penalization. *A priori*, there are no reasons to reject it. Nevertheless, the main conclusion of Section 3.3 is that the weak penalty method is to be preferred.

The new results and methods that have been introduced in this direction are:

- *Stability of some 2D elements.* Numerical experiments have demonstrated that the Q_2/Q_1 element suffers from having a small stability constant. The Q_2/P_0 element is robust, although very inaccurate (overdiffusive). Furthermore, it cannot be exactly reproduced using a RIP method. The integration error may be

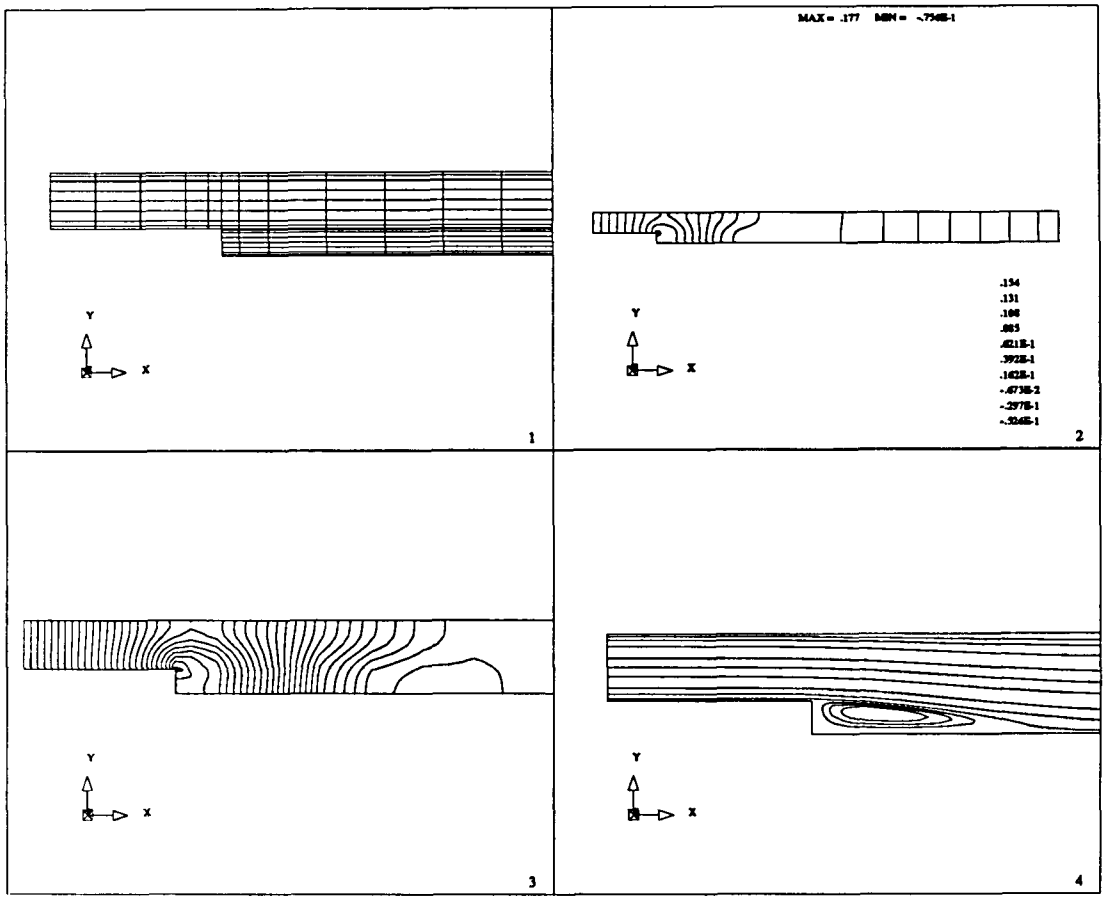

Figure 3.11 Numerical solution of the backward-facing step problem. (1): Detail of the mesh; (2): Pressure contours; (3): Detail of pressure contours; (4): Detail of streamlines.

responsible in part for this wrong behavior. It has also been proved theoretically and confirmed through numerical experiments that the Q_2/P_1 element cannot be emulated using the strong penalty method.

- *Pressure calculation.* A filtering technique has been proposed to compute the pressure for the Q_2/Q_1 element that has proved to be effective. Also, the possibility of solving a Poisson equation for the pressure after the velocity is known has been studied.
- *Petrov-Galerkin weighting.* Based on the results of Chapter 1, a Streamline Diffusion method has been introduced for the Navier-Stokes equations when the strong penalization is used.

It is important to point out that the numerical experiments presented in Section 3.3 are only those that have been considered *representative*, but many other tests have also been conducted that, for brevity, have not been included here.

However, the most important results are those concerning the iterative (weak) penalty method proposed and analysed in Section 3.4. We believe that this method has very interesting features. The penalty method for the incompressible Navier-Stokes

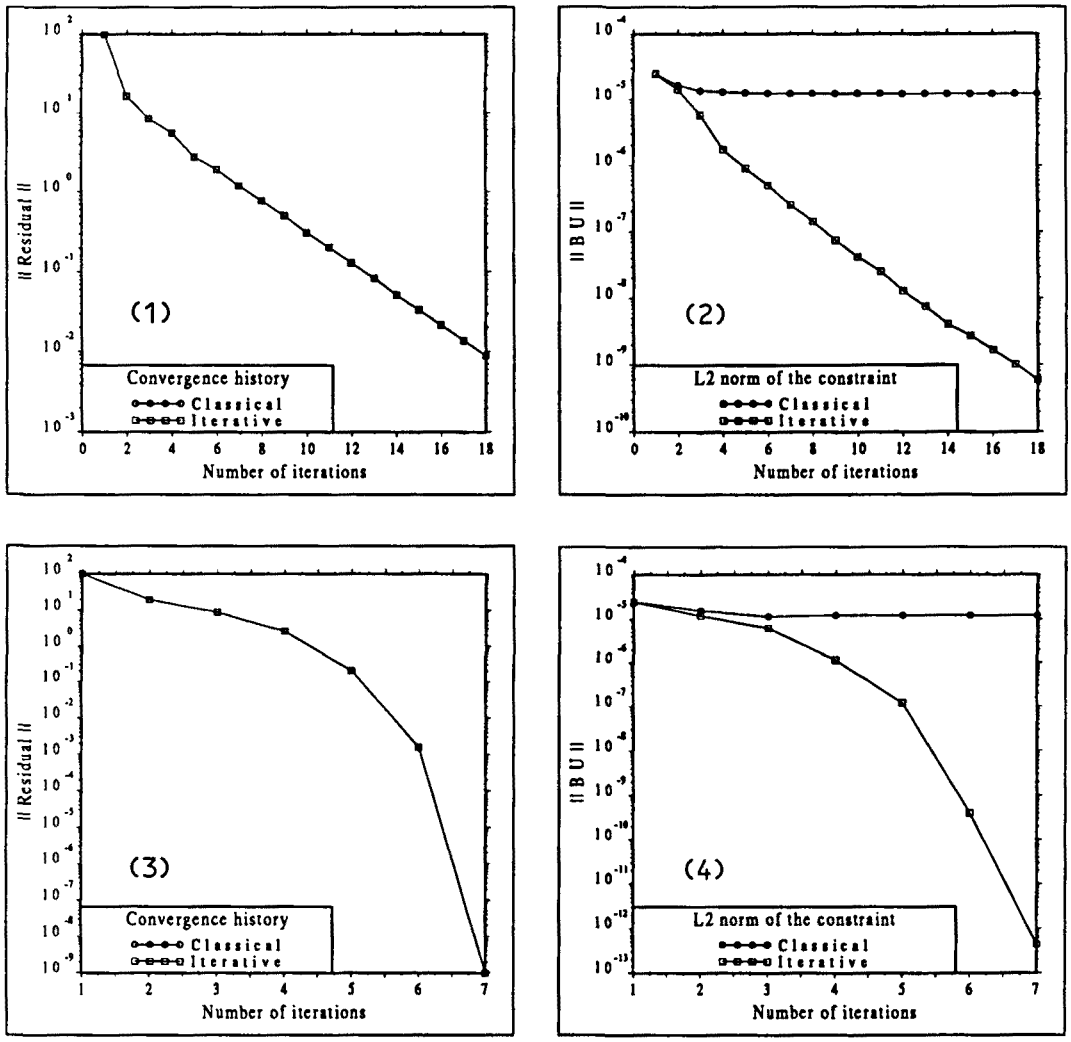

Figure 3.12 Comparison of the classical and iterative penalty methods for the backward-facing step problem with $\epsilon = 10^{-4}$. (1): Convergence history for the Picard algorithm; (2): Norm of the constraint for the Picard algorithm; (3): Convergence history for the Newton-Raphson algorithm; (4): Norm of the constraint for the Newton-Raphson algorithm.

equations in its classical form is attractive. It reduces the number of nodal unknowns and yields good results. This is a very important attribute if three-dimensional problems have to be solved on medium-size computers. However, small penalty parameters lead to ill-conditioned stiffness matrices. Usually, this ill-conditioning is not a trouble if direct solvers are used. But, still thinking in the numerical simulation of 3-D flows, iterative solvers are almost imperative when a real problem has to be faced. These solvers are very sensitive to the condition number of the stiffness matrix and this seriously limits the feasibility of the classical penalty method. The iterative penalization presented here tries to circumvent, at least in part, this inconvenience. It allows the use of much larger penalty parameters, thus yielding matrices whose condition numbers are much smaller. Whether this will be enough for using iterative solvers or not

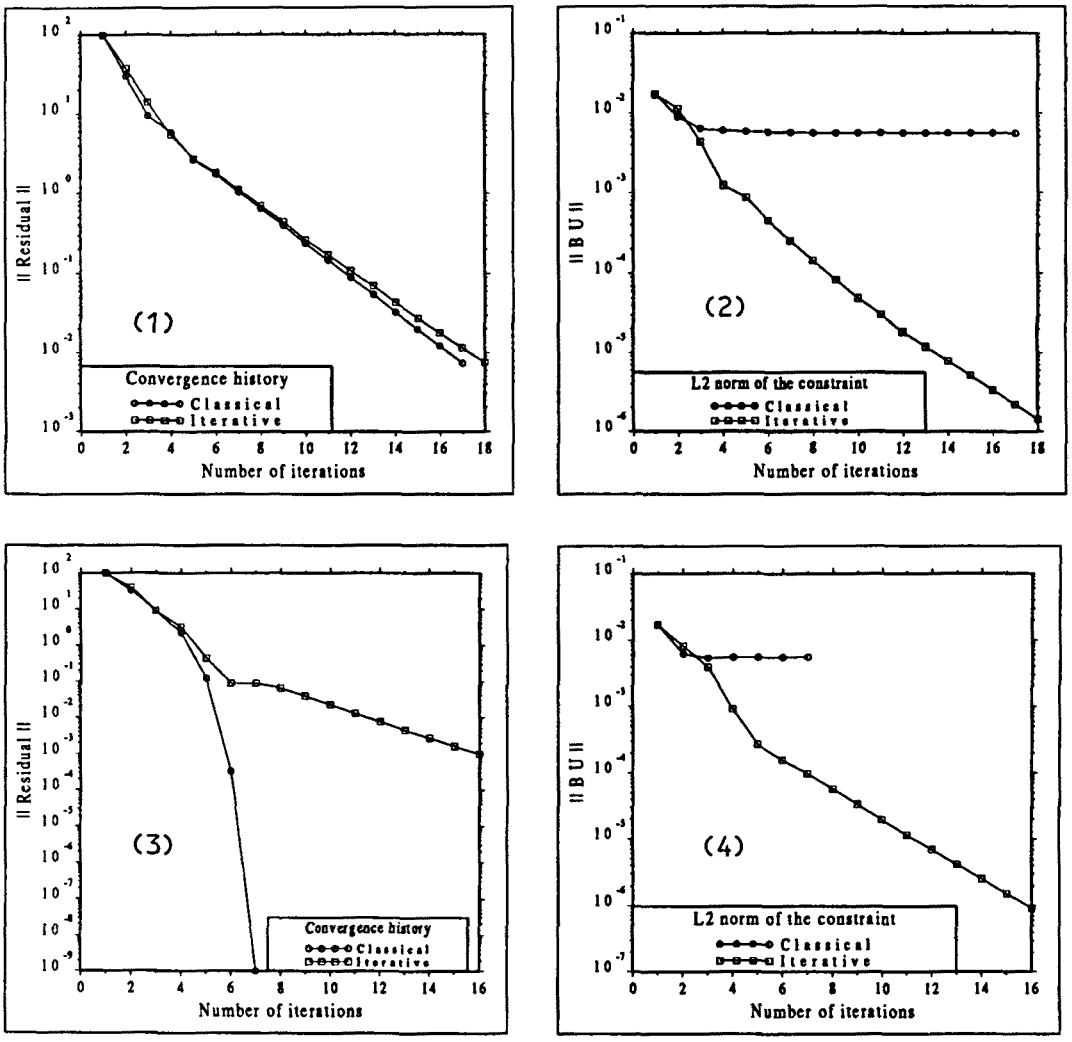

Figure 3.13 Comparison of the classical and iterative penalty methods for the backward-facing step problem with $\epsilon = 10^{-1}$. (1): Convergence history for the Picard algorithm; (2): Norm of the constraint for the Picard algorithm; (3): Convergence history for the Newton-Raphson algorithm; (4): Norm of the constraint for the Newton-Raphson algorithm.

is something that experience has to provide. Results for the 3D cavity flow example presented in Section 3.5 (for the Stokes problem) using the conjugate gradient method are certainly encouraging.

The iterative penalization presented here may be obtained from different approaches conceptually different. For the Stokes problem, it reduces to the Augmented Lagrangian method combined with the Uzawa algorithm to uncouple the pressure. It can also be interpreted as the introduction of an artificial compressibility and a false transient only for the pressure whenever the temporal derivative is discretized using the backward Euler scheme. However, we prefer the residual argument described in Section 3.4.1 since it is still valid when the iterative equation for the pressure is coupled with a linearized form of the momentum equations in the Navier-Stokes problem. The

convergence analysis of the iterative penalty method coupled with this linearization of the convective term has shown that:

- *Picard-based algorithm.* The analysis of the algorithm (Theorem 3.2) reveals that the rate of convergence is smaller than for the classical penalty method.
- *Newton-Raphson based algorithm.* The attraction ball of the exact solution happens to be smaller than for the classical penalization. Quadratic convergence can only be ensured up to a certain iteration (Theorem 3.3).

However, numerical experiments indicate that these effects are only apparent when the penalty parameter is 'very large', compared with the standards of the classical approach. Anyway, if needed, there is also the possibility of uncoupling the iterative penalization (Theorem 3.4) and obtain a rate of convergence that only depends on the penalty parameter.

In practice, it is common to use penalties of order $10^{-6}\mu^{-1}$ to $10^{-9}\mu^{-1}$. We have already said that these values can be easily handled using direct solvers. However, there are some practical cases in which the viscosity varies several orders of magnitude in the fluid domain, as in quasi-Newtonian fluids with thermal dependent physical properties. In these cases, the above rule has to be applied using the smallest value of the viscosity, thus relaxing in excess the incompressibility constraint in the high viscosity zones. We will start up again with this argument in Chapter 5.

References

[Ar] D.N. Arnold. Mixed finite element methods for elliptic problems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 82 (1990), 281–300

[ABF] D.N. Arnold., F. Brezzi and M. Fortin. A stable finite element for the Stokes equations. *Calcolo*, vol. 21 (1984), 337–344

[Ba1] I. Babuška. Error bounds for finite element method. *Numer. Math.*, vol. 16 (1971), 322–333

[Ba2] I. Babuška. The finite element method with Lagrangian multipliers. *Numer. Math.*, vol. 20 (1973), 179–192

[BOP] I. Babuška, J. Osborn and J. Pitkäranta. Analysis of mixed methods using mesh dependent norms *Math. Comp.*, vol. 35 (1980), 1039–1062

[Be] M. Bercovier. Perturbation of a mixed variational problem, application to mixed finite element methods. *RAIRO Anal. Numer.*, vol. 12 (1978), 211–236

[BP] M. Bercovier and O. Pironneau. Error estimates for finite element method solution of the Stokes problem in the primitive variables. *Numer. Math.*, vol. 33 (1979), 211–224

[BN1] J.M. Boland and R.A. Nicolaides. Stability of finite elements under divergence constraints. *SIAM J. Numer. Anal.*, vol. 20 (1983), 722–731

[BN2] J.M. Boland and R.A. Nicolaides. On the stability of bilinear-constant velocity-pressure finite elements. *Numer. Math.*, vol. 44 (1984), 219–222

[BN3] J.M. Boland and R.A. Nicolaides. Stable and semistable low order finite elements for viscous flows. *SIAM J. Numer. Anal.*, vol. 22 (1985), 474–492

[Br] F. Brezzi. On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers. *RAIRO Anal. Numer.*, vol. 8

(1974), 129–151

[BB] F. Brezzi and K.J. Bathe. A discourse on the stability conditions for mixed finite element formulations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 82 (1990), 27–57

[BD] F. Brezzi and J. Douglas. Stabilized mixed methods for the Stokes problem. *Numer. Math.*, vol. 53 (1988), 225–235

[BFa] F. Brezzi and R. Falk. Stability of higher-order Taylor-Hood elements. *SIAM J. Numer. Anal.*, vol. 28 (1991), 581–590

[BFo] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods* (Springer-Verlag, 1991).

[BR1] F. Brezzi, J. Rappaz and P.A. Raviart. Finite dimensional approximation of nonlinear problems. Part I: Branches of non-singular solutions. *Numer. Math.*, vol. 36 (1981), 1–25

[BR2] F. Brezzi, J. Rappaz and P.A. Raviart. Finite dimensional approximation of nonlinear problems. Part II: Limit points. *Numer. Math.*, vol. 37 (1981), 1–28

[BR3] F. Brezzi, J. Rappaz and P.A. Raviart. Finite dimensional approximation of nonlinear problems. Part III: Simple bifurcation points. *Numer. Math.*, vol. 38 (1981), 1–30

[CK1] G.F. Carey and R. Krishnan. Penalty approximation of Stokes flow. *Comput. Meths. Appl. Mech. Engrg.*, vol. 35 (1982), 169–206

[CK2] G.F. Carey and R. Krishnan. Penalty finite element method for the Navier-Stokes equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 42 (1984), 183–224

[CK3] G.F. Carey and R. Krishnan. Continuation techniques for a penalty approximation of the Navier-Stokes equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 48 (1985), 265–282

[CK4] G.F. Carey and R. Krishnan. Convergence of iterative methods in penalty finite element approximations of the Navier-Stokes equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 60 (1987), 1–29

[CO1] G.F. Carey and J.T. Oden. *Finite Elements: A second course.* The Texas Finite Element Series, vol. II (Prentice Hall, 1983)

[CO2] G.F. Carey and J.T. Oden. *Finite Elements: Fluid Mechanics.* The Texas Finite Element Series, vol. VI (Prentice Hall, 1986)

[Ch] A.J. Chorin. A numerical method for solving incompressible viscous flow problems. *J. Comput. Phys.*, vol. 2 (1967), 12–26

[Ci] P.G. Ciarlet. *The finite element method for elliptic problems.* (North-Holland, 1978)

[Co] R. Codina. An iterative penalty method for the finite element solution of the stationary Navier-Stokes equations. *CIMNE Report Num. 12* (1991) (Submitted to *Comput. Meths. Appl. Mech. Engrg.*)

[CCO] R. Codina, M. Cervera and E. Oñate. A penalty finite element method for non-Newtonian creeping flows. *CIMNE Report Num. 13* (1991) (Submitted to *Int. J. Numer. Meth. Engrg.*)

[CR] M. Crouzieux and P.A. Raviart. Conforming and non-conforming finite element methods for the stationary Stokes equations. *RAIRO Anal. Numer.*, vol. 7 (1973), 33–76

[CSS] C. Cuvelier, A. Segal and A. van Steenhoven. *Finite element methods and Navier-Stokes equations.* (Reidel, 1986)

[DJM] J. Donea, S. Giulinani, K. Morgan and L. Quartapelle. The significance of chequerboarding in a Galerkin finite element solution of the Navier-Stokes equations. *Int. J. Numer. Meth. Engrg.*, vol. 17 (1981), 790–795

[ESG] M.S. Engelman, R.L. Sani, P.M. Gresho and M. Bercovier. Consistent vs reduced integration penalty methods for incompressible media using several old and new elements. *Int. J. Numer. Meth. Fluids*, vol. 2 (1983), 25–42

[Fe] C.A. Felippa. Iterative procedures for improving penalty function solutions of algebraic systems. *Int. J. Numer. Meth. Engrg.*, vol. 12 (1978), 821–836

[Fo1] M. Fortin. Analysis of the convergence of mixed finite element methods. *RAIRO Anal. Numer.*, vol. 11 (1977), 341–354

[Fo2] M. Fortin. Old and new finite elements for incompressible flows. *Int. J. Numer. Meth. Fluids*, vol. 1 (1981), 347–364

[Fo3] M. Fortin. Two comments on: Consistent vs reduced integration penalty methods for incompressible media using several old and new elements. *Int. J. Numer. Meth. Fluids*, vol. 3 (1983), 93–98

[FB] M. Fortin and S. Boivin. Iterative stabilization of the bilinear velocity-constant pressure element. *Int. J. Numer. Meth. Fluids*, vol. 10 (1990), 125–140

[FF] M. Fortin and A. Fortin. Experiments with several elements for viscous incompressible flows. *Int. J. Numer. Meth. Fluids*, vol. 5 (1985), 911–928

[FH] L. Franca and T.J.R. Hughes. Two classes of mixed finite element methods. *Comput. Meths. Appl. Mech. Engrg.*, vol. 69 (1988), 89–129

[FS] L. Franca and R. Stenberg. Error analysis of some Galerkin least-squares methods for the elasticity equations. *INRIA, Rapports de Recherche* (1989)

[FHL] L. Franca, T.J.R. Hughes, A.F.D. Loula and I. Miranda. A new family of stable elements for nearly incompressible elasticity based on a mixed Petrov-Galerkin finite element formulation. *Numer. Math.*, vol. 53 (1988), 123–141

[Fr] I. Fried. Finite element analysis of incompressible material by residual energy balancing. *Int. J. of Solids and Struct.*, vol. 10 (1974), 993–1002

[GR] V. Girault and P.A. Raviart. *Finite element methods for Navier-Stokes equations* (Springer-Verlag, 1986).

[Gl] R. Glowinski. *Numerical methods for nonlinear variational problems.* (North-Holland, 1984)

[GP] R. Glowinski and O. Pironneau. On a mixed finite element approximation of the Stokes problem. *Numer. Math.*, vol. 33 (1979), 397–424

[GS] P.M. Gresho and R.L. Sani. On pressure boundary conditions for the incompressible Navier- Stokes equations. In: *Finite elements in fluids*, vol. 7, R.H. Gallagher, R. Glowinski, P.M. Gresho, J.T. Oden and O.C. Zienkiewicz (eds.) (John Wiley & Sons Ltd., 1987).

[Gu] M. Gunzburger. *Finite element methods for viscous incompressible flows* (Academic Press, 1989).

[HS] P. Hansbo and A. Szepessy. A velocity-pressure streamline diffusion finite element method for the incompressible Navier-Stokes equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 84 (1990), 175–192

[HD] J.C. Heinrich and B.R. Dyne. On the penalty method for incompressible fluids. In: *Finite elements in the 90's.* E. Oñate, J. Periaux and A. Samuelson (eds.) (Springer-Verlag/CIMNE, 1991).

[HO] E. Hinton and D.R.J. Owen. *Finite element programming.* (Academic Press, 1977)

[Hu1] T.J.R. Hughes. Equivalence of finite elements for nearly-incompressible elasticity. *J. Appl. Mech.*, vol. 44 (1977), 181–183

[Hu2] T.J.R. Hughes. *The finite element method. Linear static and dynamic analysis.* (Prentice-Hall, 1987)

[HA] T.J.R. Hughes and H. Allik. Finite elements for compressible and incompressible continua. In: *Proceedings of the Symposium on civil Engineering*, Vandervilt Univ., Nashville Tenn. (1969)

[HF] T.J.R. Hughes and L.P. Franca. A new finite element formulation for computational fluid dynamics: VII. Ths Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Meths. Appl. Mech. Engrg.*, vol. 65 (1987), 85–96

[HFB] T.J.R. Hughes, L.P. Franca and M. Balestra. A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska-Brezzi condition: a stable Petrov-Galerkin formulation for the Stokes problem accommodating equal-order interpolations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 59 (1986), 85–99

[HLB] T.J.R. Hughes, W.K. Liu and A. Brooks. Finite Element Analysis of Incompressible Viscous Flows by the Penalty Function Formulation. *J. Comput. Phys.*, vol. 30 (1979), 1–60

[Ir] B. Irons. A frontal solution program. *Int. J. Numer. Meth. Engrg.*, vol. 2 (1970), 5–32

[IL] B. Irons and M. Loikkanen. An engineers' defence of the patch test. *Int. J. Numer. Meth. Engrg.*, vol. 19 (1983), 1391–1401

[JP] C. Johnson and J. Pitkäranta. Analysis of some mixed finite element methods related to reduced integration. *Math. Comp.*, vol. 38 (1982), 375–400

[KOS] N. Kikuchi, J.T. Oden and Y.J. Song. Convergence of modified penalty methods and smoothing schemes of pressure for Stokes' flow problems. In: *Finite elements in fluids*, vol. 5, R.H. Gallagher, J.T. Oden, O.C. Zienkiewicz, T. Kawai and M. Kawahara (eds.) (John Wiley & Sons Ltd., 1984).

[Ki] S.W. Kim. A finite element computational method for high Reynolds number laminar flows. NASA CR-179135 (1987).

[La] O. Ladyzhenskaya. *The mathematical theory of viscous incompressible flow.* (Gordon-Breach, 1963)

[MH] D.S. Malkus and T.J.R. Hughes. Mixed finite element methods-reduced and selective integration techniques: a unification of concepts. *Comput. Meths. Appl. Mech. Engrg.*, vol. 15 (1978), 63–81

[NPR] J.C. Nagtegaal, D.M. Parks and J.R. Rice. On numerically accurate finite element solutions in the fully plastic range. *Comput. Meths. Appl. Mech. Engrg.*, vol. 4 (1974), 153–177

[Ne1] J.C. Nedelec. Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, vol. 35 (1980), 315–341

[Ne2] J.C. Nedelec. Elements finis mixtes incompressibles pour l'equation de Stokes dans $\mathbb{R}^3$. *Numer. Math.*, vol. 39 (1982), 97–112

[Od] J.T. Oden. RIP-methods for Stokesian flows. In: *Finite elements in fluids*, vol. 4, R.H. Gallagher, D.H. Norrie, J.T. Oden and O.C. Zienkiewicz (eds.) (John Wiley & Sons Ltd., 1982)

[OC] J.T. Oden and G.F. Carey. *Finite Elements: Mathematical aspects.* The Texas Finite Element Series, vol. IV (Prentice Hall, 1983)

[OJ1] J.T. Oden and O.P. Jacquotte. Stability of some mixed finite element methods

for Stokesian flows. *Comput. Meths. Appl. Mech. Engrg.*, vol. 43 (1984), 231–247

[OJ2] J.T. Oden and O.P. Jacquotte. Stable and unstable Rip/Perturbed Lagrangian methods for two-dimensional viscous flow problems. In: *Finite elements in fluids*, vol. 5, R.H. Gallagher, J.T. Oden, O.C. Zienkiewicz, T. Kawai and M. Kawahara (eds.) (John Wiley & Sons Ltd., 1984)

[OKS] J.T. Oden, N. Kikuchi and Y.J. Song. Penalty finite element methods for the analysis of Stokesian flows. *Comput. Meths. Appl. Mech. Engrg.*, vol. 31 (1982), 297–329

[Pi] O. Pironneau. *Finite element methods for fluid flow.* (John Wiley & Sons, 1989)

[RT] P.A. Raviart and J.M. Thomas. A mixed finite element method for second order elliptic problems. In: *Mathematical aspects of the finite element method*, I. Galligani, E. Magues (eds.). Lecture notes in Mathematics, 606 (Springer, 1977)

[Ra] A. Razzaque. The patch test for elements. *Int. J. Numer. Meth. Engrg.*, vol. 22 (1986), 63–71

[RTF] M.P. Robichaud, Ph. Tanguy and M. Fortin. An iterative implementation of the Uzawa algorithm for 3-D fluid flow problems. *Int. J. Numer. Meth. Fluids*, vol. 10 (1990), 429–442

[Sa] E.M. Salonen. An iterative penalty function method in structural analysis. *Int. J. Numer. Meth. Engrg.*, vol. 10 (1976), 413–421

[SG1] R.L. Sani, P.M. Gresho, R.L. Lee and D.F. Griffiths. The cause and cure (?) of the spurious pressures generated by certain FEM solutions of the incompressible Navier-Stokes equations. Part 1. *Int. J. Numer. Meth. Fluids*, vol. 1 (1981), 17–43

[SG2] R.L. Sani, P.M. Gresho, R.L. Lee and D.F. Griffiths. The cause and cure (!) of the spurious pressures generated by certain FEM solutions of the incompressible Navier-Stokes equations. Part 2. *Int. J. Numer. Meth. Fluids*, vol. 1 (1981), 171–204

[SH] J.L. Sohn and J.C. Heinrich. Pressure calculations in penalty finite element approximations to the Navier-Stokes equations. *Int. J. Numer. Meth. Engrg.*, vol. 30 (1990), 349–361

[St1] R. Stenberg. Analysis of mixed finite element methods for the Stokes problem: a unified approach. *Math. Comp.*, vol. 42 (1984), 9–23

[St2] R. Stenberg. Error analysis of some finite element methods for the Stokes problem. *Math. Comp.*, vol. 54 (1990), 495–508

[St3] R. Stenberg. A technique for analysing finite element methods for viscous incompressible flow. *Int. J. Numer. Meth. Fluids*, vol. 11 (1990), 935–948

[TR] Ph. Le Tallec and V. Ruas. On the convergence of the bilinear-velocity constant-pressure finite element method in viscous flow. *Comput. Meths. Appl. Mech. Engrg.*, vol. 54 (1986), 235–243

[TH] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element method. *Comp. & Fluids*, vol. 1 (1973), 73–100

[TSZ] R.L. Taylor, J.C. Simo, O.C. Zienkiewicz and C.H. Chan. The patch test – A condition for assessing FEM convergence. *Int. J. Numer. Meth. Engrg.*, vol. 22 (1986), 39–62

[Te] R. Temam. *Navier-Stokes equations.* (North-Holland, 1984)

[Ve1] R. Verfürth. Error estimates for a mixed finite element interpolations of Stokes problem. *RAIRO Anal. Numer.*, vol. 18 (1984), 175–182

[Ve2] R. Verfürth. Finite element approximation of incompressible Navier-Stokes equations with slip boundary conditions. *Numer. Math.*, vol. 50 (1987), 697–721

[ZQT] O.C. Zienkiewicz, S. Qu, R.L. Taylor and S. Nakazawa. The patch test for mixed formulations. *Int. J. Numer. Meth. Engrg.*, vol. 23 (1986), 1873–1883

[ZT] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method*, Fourth Edition, Vol. 1 (McGraw-Hill, 1989)

[ZTT] O.C. Zienkiewicz, R.L. Taylor and J.M. Too. Reduced integration technique in general analysis of plates and shells. *Int. J. Numer. Meth. Engrg.*, vol. 3 (1971), 275–290

[ZVT] O.C. Zienkiewicz, J.P. Vilotte, S. Toyoshima and S. Nakazawa. Iterative method for constraint and mixed approximation; an inexpensive improvement of F.E.M. performance. *Comput. Meths. Appl. Mech. Engrg.*, vol. 51 (1985), 3–29