INTERNATIONAL CENTER FOR NUMERICAL METHODS IN ENGINEERING

A Finite Element Formulation for the Numerical Solution of the Convection-Diffusion Equation

R. Codina

Monograph CIMNE Nº-14, January 1993

A Finite Element Formulation for the Numerical Solution of the Convection-Diffusion Equation

R. Codina

Monograph CIMNE Nº-14, January 1993

International Center for Numerical Methods in Engineering
Gran Capitán s/n, 08034 Barcelona, Spain

INTERNACIONAL CENTER FOR NUMERICAL METHODS IN ENGINEERING
Edificio C1, Campus Norte UPC
Gran Capitán s/n
08034 Barcelona, Spain
www.cimne.upc.es

First edition: January 1993

A FINITE ELEMENT FORMULATION FOR THE NUMERICAL SOLUTION OF THE CONVECTION-DIFFUSION EQUATION
Monograph CIMNE M14

ISBN: 84-87867-17-0

CONTENTS

CHAPTER 2

TRANSIENT ALGORITHMS—STABILITY ANALYSIS OF AN EXPLICIT SCHEME

CHAPTER 3

A DISCONTINUITY-CAPTURING CROSSWIND-DISSIPATION FOR THE STEADY-STATE PROBLEM

FOREWORD

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.

One of the difficulties associated to the numerical solution of the flow equations is the presence of the convective term when an Eulerian frame of reference is adopted. Using finite difference methods, centered approximations fail to give realistic solutions when this term becomes important, a problem that was very soon recognized and for which early remedies go back to the fifties. Finite element methods have also the same problems when the standard Galerkin formulation is employed. Flows for which convective effects are more important than diffusive ones are often referred to as convection-dominated flows.

In order to isolate possible sources of problems, the linear convection-diffusion equation is commonly employed as a model to study convection-dominated flows. Numerical techniques are first devised for this simple equation and then they are extended to more complex flow situations—an extension which is rarely straightforward, by the way.

In this work we present several finite element techniques to solve the convection-diffusion equation when the Péclet number is high, that is, when diffusion is very small. The problem is this case becomes a singularly perturbed one, since its nature changes when the zero diffusion case is considered. Physically, this is reflected by the appearence of very narrow zones with steep gradients of the solution, either to accomodate the boundary conditions (boundary layers) or to advect discontinuous profiles into the computational domain (internal layers).

There are two levels of difficulty in the numerical solution of the convection-diffusion equation. First, a numerical method must be designed in order to avoid the instability of the standard Galerkin approach. Accurate methods accomplishing this objective usually yield small oscillations in the vicinity of sharp layers. Sometimes, even this low-scale problem is unacceptable. The basic formulation has then to be modified to overcome local overshooting and undershooting. Both the basic formulation and this additional modification are treated in this work.

The basic finite element formulation employed here is the Streamline-Upwind/Petrov-Galerkin (SUPG) method. A thorough description of this approach 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. Although our main concern will be the numerical solution of the steady-state problem, Chapter 2 deals 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. Its main interest relies on the fact that it allows to obtain steady-state solutions via a transient relaxation, an usual procedure in CFD. Chapter 3 is concerned with the problem of removing the localized oscillations that remain about abrupt layers of the solution. A method based on the introduction of a nonlinear crosswind dissipation is proposed to remove them.

Some of the results presented in the first chapter have appeared in an article written in collaboration with Prof. E. Oñate and Dr. M. Cervera. A want to express my gratitude to both of them for initiating my research in the field of computational mechanics. I would like also to thank Prof. J. Miquel, for some fruitful discussions, and my colleagues of the Department of Structures at the Polytechnical University of Catalonia, for their support and stimulation during my work in the doctoral thesis on which part of this monograph is based.

Ramon Codina
Barcelona, October 1992

CHAPTER 1

THE STREAMLINE–UPWIND/PETROV–GALERKIN 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' (SD). Nowadays, the use of this method has become widespread and the name SD seems to prevail over SUPG in mathematical circles, but SUPG is in general preferred. For this reason, the latter 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 SUPG 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 SUPG 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 SUPG method will be treated in Chapter 3.

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 $0 = x_0 < x_1 < \ldots < x_N = \ell$ be a uniform partition of the interval $[0, \ell]$, with $x_{m+1} - x_m = h$, $m = 0, \ldots, N-1$. 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:

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

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 $\mathrm{sgn}(k^*) = \mathrm{sgn}(k)$. 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 SUPG 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

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

(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 very different approaches in the next two chapters. Therefore, there are several reasons for taking $|\alpha| \ge \alpha_c$.

1.1.3 Multidimensional case: the Streamline-Upwind/Petrov-Galerkin method

Consider first the continuous steady-state convection-diffusion problem. Let Ω be an open bounded polyhedral 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 [Hu1], [HHZ]. The main idea underlying the SUPG 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 SUPG 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 SUPG method: Find a function $\phi_h \in \Phi_h$ such that

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

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

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

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

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 SUPG 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 SUPG 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 SUPG 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 SUPG method for the convection-diffusion equation. However, it can be proved [HFB] that it allows to circumvent the Babuška-Brezzi stability condition 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 SUPG formulation. If the zero convection limit is considered, the perturbation of the test function (1.37) for the SUPG 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 SUPG 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 SUPG 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 SUPG method will read: Find $\phi \in V_h$ such that

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

where

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

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

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_{su} 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, $a_{su}(\psi_h, \psi_h) \geq C|||\psi_h|||^2$ *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

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

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 SUPG 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*

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

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_{su} will be used:

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

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

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

$$
\begin{aligned}
E(\eta) &\leq 5k\|\nabla\eta\|^2 + 4\tau^{-1}\|\eta\|^2 + \sigma\|\eta\|^2 + 2\tau\|\mathbf{u}\cdot\nabla\eta\|^2 + \sigma\tau^2\|\mathbf{u}\cdot\nabla\eta\|^2 \\
&\leq \kappa_1\left(5kh^{2m} + \frac{8|\mathbf{u}|}{\alpha h}h^{2m+2} + \sigma h^{2m+2} + \frac{\alpha h}{|\mathbf{u}|}|\mathbf{u}|^2h^{3m} + \sigma\frac{\alpha^2h^2}{4|\mathbf{u}|^2}|\mathbf{u}|^2h^{2m}\right) \\
&\leq C|\mathbf{u}|\left(\frac{1}{\gamma}h^{2m+1} + \frac{8}{\alpha}h^{2m+1} + \alpha h^{2m+1}\right) + C'\alpha^2\sigma h^{2m+2} \qquad (1.64)
\end{aligned}
$$

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]:

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

See References [Na], [HFH].

(4) The fact that $\sigma > 0$ implies that (1.62) will also hold in the L^2 norm. Since the error of the SUPG formulation is of order $h^{m+1/2}$ in this norm and the interpolation error is of order h^{m+1}, the SUPG 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 SUPG method.

A possible extension of the SUPG 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 SUPG 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 SUPG 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 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.

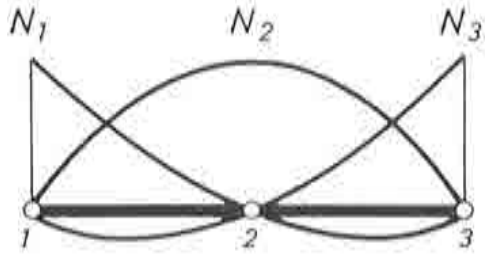

Figure 1.1 Three noded quadratic element and shape functions

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

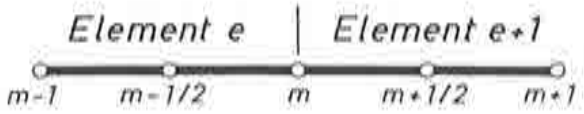

Figure 1.2 Indexes referring the nodes of two adjacent elements

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

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

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

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

1.3.2 Standard formulation of the SUPG 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 SUPG 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)

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

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 SUPG 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

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

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

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 '*1*'. Functions (1.16), (1.86) and (1.87) are called *asymptotic approximations.*

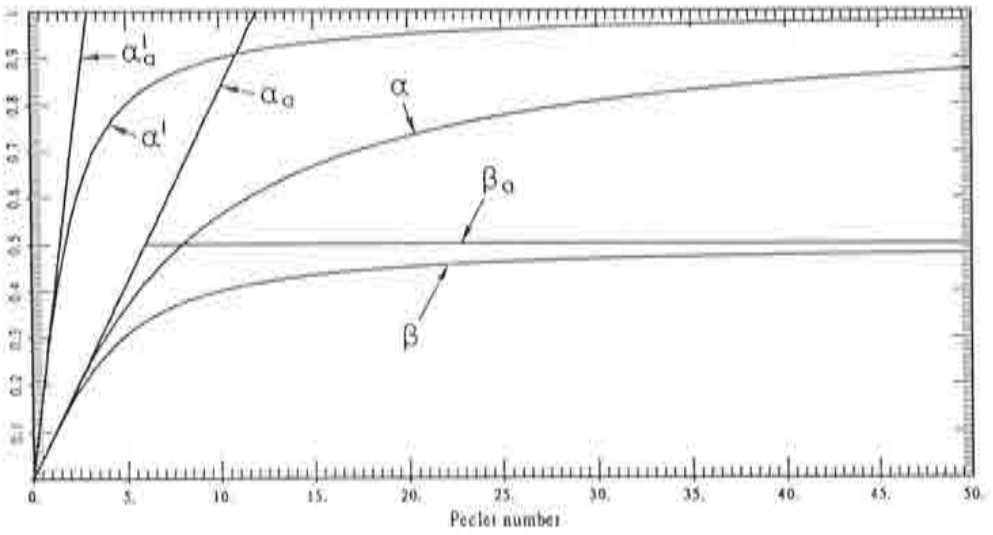

Figure 1.3 Upwind functions for linear and quadratic elements and their asymptotic approximations

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 SUPG 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.

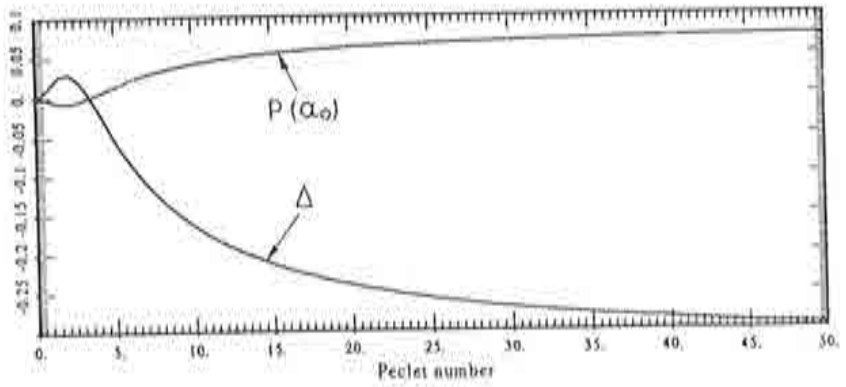

Figure 1.4 Discriminant Δ of equation (1.89) and values of $P(\alpha_0)$ for α_0 given by (1.92)

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.

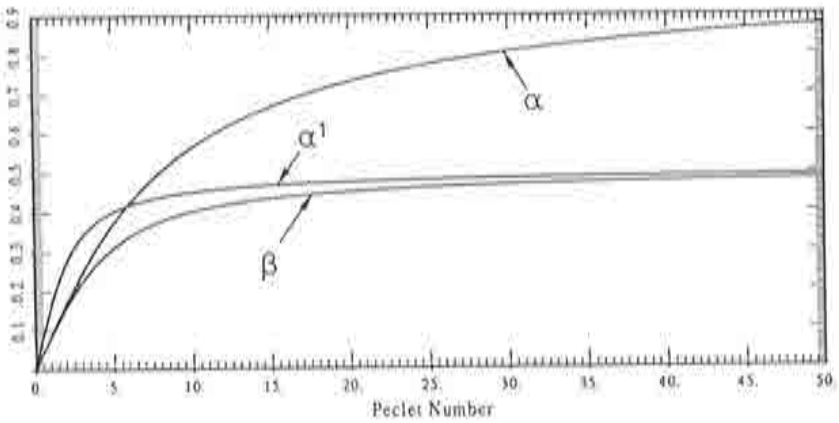

Figure 1.5 Upwind functions for quadratic elements

1.3.3 Hierarchic formulation of the SUPG 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

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

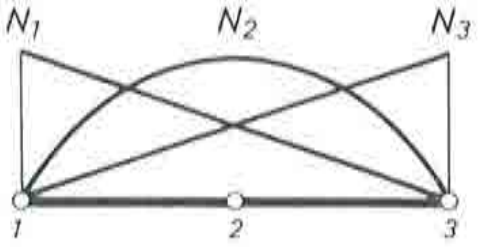

Figure 1.6 Hierarchic shape functions for three noded quadratic elements

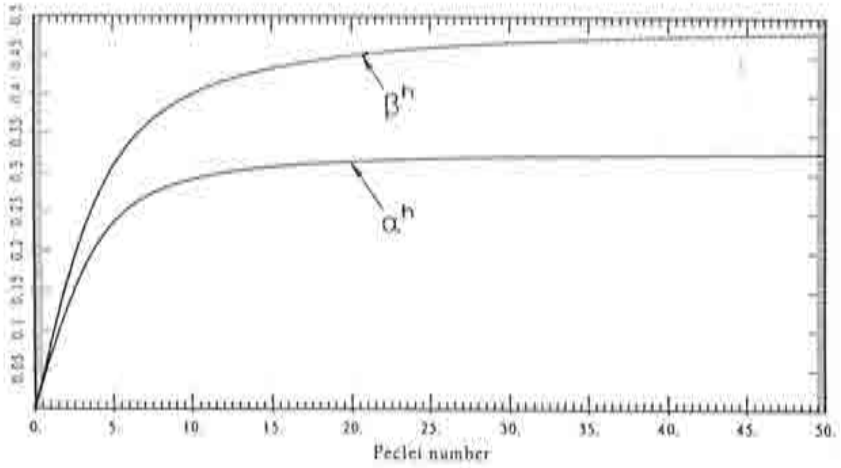

Figure 1.7 Upwind functions for hierarchic quadratic elements

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

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.

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 SUPG 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 SUPG 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):

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

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

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 $\gamma\to\infty$.

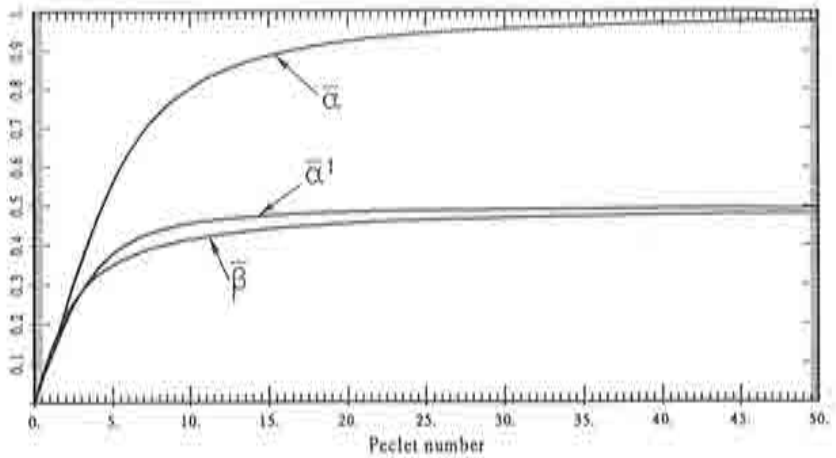

Figure 1.8 Upwind functions for quadratic elements using the GLS method

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 SUPG 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, SUPG, 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 $\pi_h : C^0(\bar{\Omega}) \longrightarrow H_h$ 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 continuous and 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 SUPG method using linear elements and the upwind function (1.12)
(ii) The SUPG method using standard quadratic elements and the upwind functions (1.83), (1.82)
(iii) The SUPG 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 $f(x)$ piecewise is constant, the numerical solution is nodally exact in all the cases. If $f(x)$ is 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

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

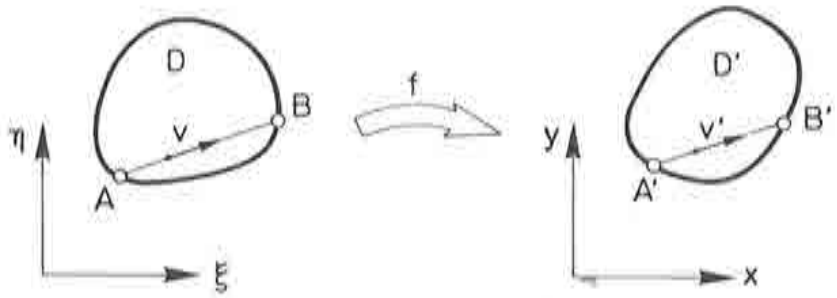

Figure 1.9 Transformation of a domain in $\mathbb{R}^2$ by an affine mapping

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 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

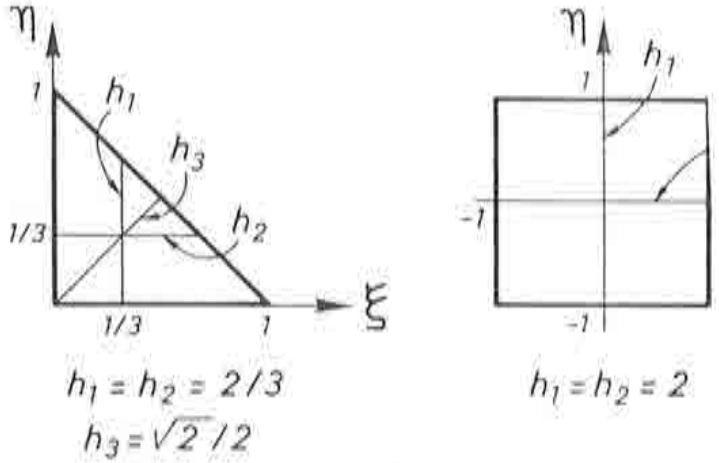

Figure 1.10 Parent domains for triangular and quadrilateral elements

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

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

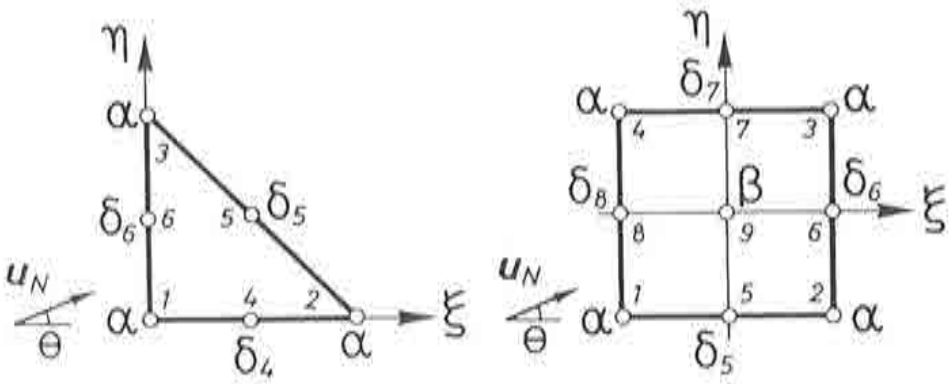

Figure 1.11 Assignment of upwind functions for the 6-noded and 9-noded elements

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 SUPG 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

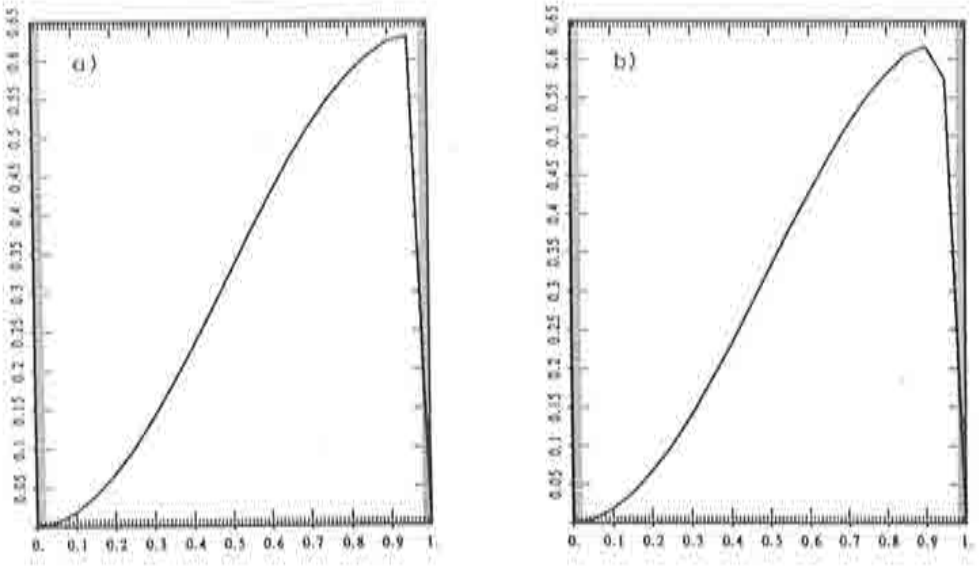

Figure 1.12 Solutions of Example 1.1. *a*) Using the upwind functions (1.82) and (1.83). *b*) Using the unique upwind function (1.93)

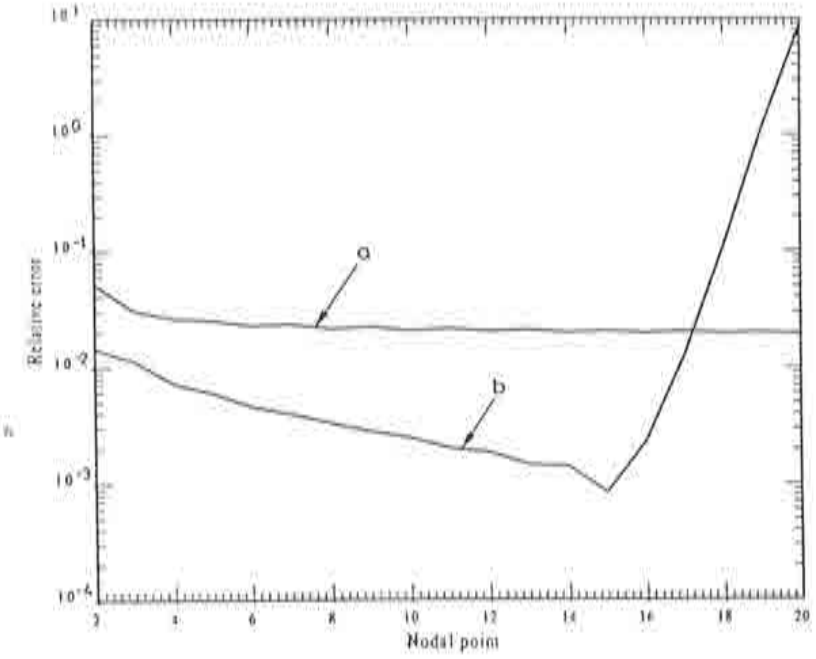

Figure 1.13 Relative errors for the solutions *a* and *b* of Figure 1.12

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.

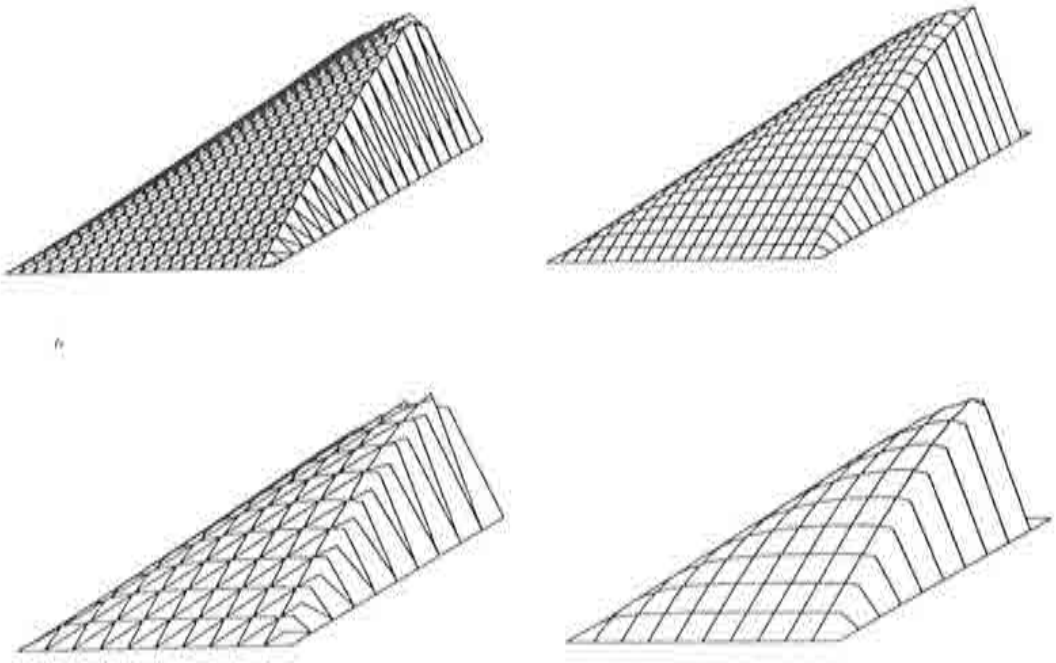

Figure 1.14 Results of Example 1.2 using triangular (3 and 6 nodes) and quadrilateral (4 and 9 nodes) elements

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

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

$$\Omega = \left]\frac{-1}{2},\frac{1}{2}\right[\times \left]\frac{-1}{2},\frac{1}{2}\right[- \{0\} \times \left]\frac{-1}{2},0\right]$$
$$\Gamma_D = \partial\Omega, \quad \Gamma_N = \emptyset$$
$$\mathbf{u}(x,y) = (-y,x)$$
$$k_{ij}(x,y) = 10^{-8}\delta_{ij}$$
$$f(x,y) = 0$$
$$g(x,y) = \begin{cases} 1 - \sin\left[\frac{\pi}{2}(1+8y)\right] & \text{if } x = 0 \text{ and } -\frac{1}{2} \le y \le 0 \\ 0 & \text{else} \end{cases}$$

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.

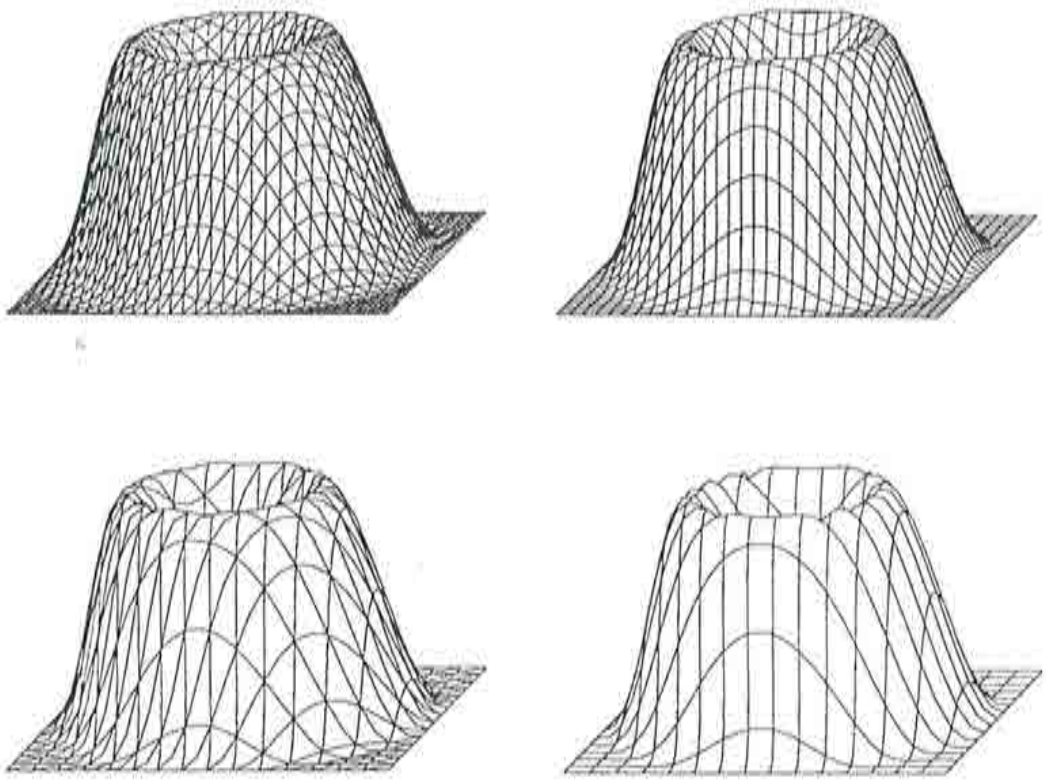

Figure 1.15 Results of Example 1.3 using triangular (3 and 6 nodes) and quadrilateral (4 and 9 nodes) elements

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 SUPG 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.

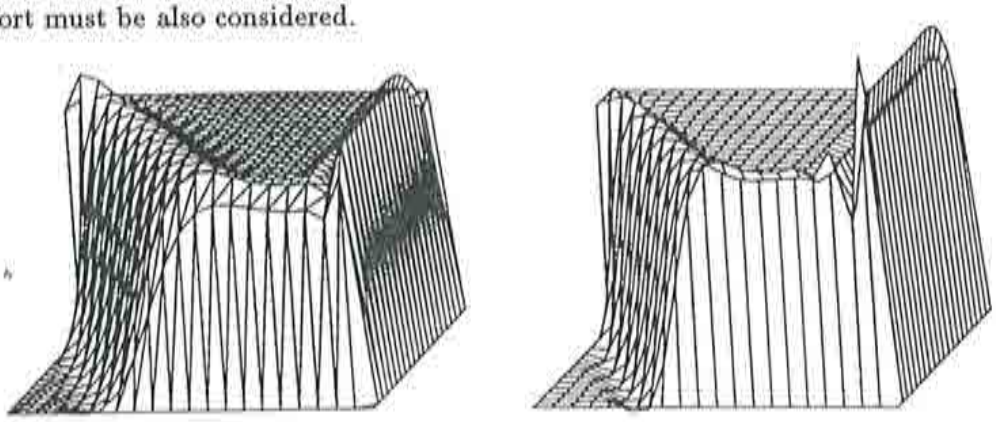

Figure 1.16 Results of Example 1.4 using 3-noded linear triangular and 4-noded bilinear quadrilateral elements

1.6 Summary and conclusions

In this chapter, a complete description of the Streamline-Upwind/Petrov-Galerkin method for solving the stationary convection-diffusion equation has been presented.

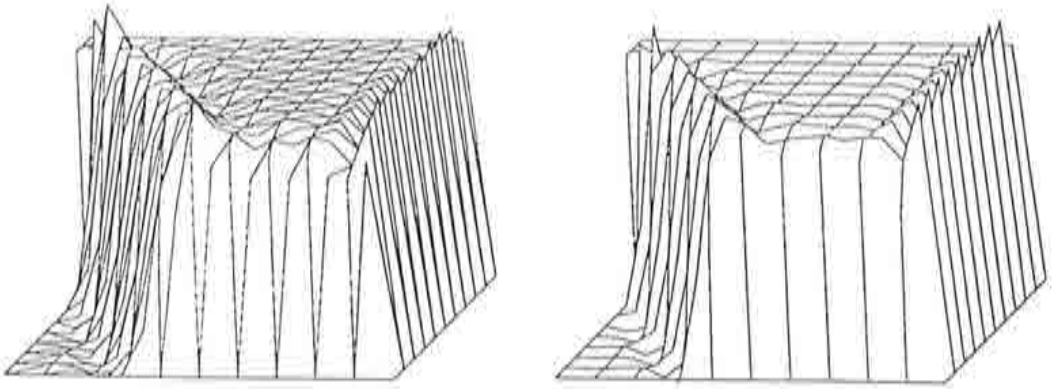

Figure 1.17 Results of Example 1.4 using quadratic triangular (6 nodes) and biquadratic quadrilateral (9 nodes) elements with the upwind functions (1.82) and (1.83)

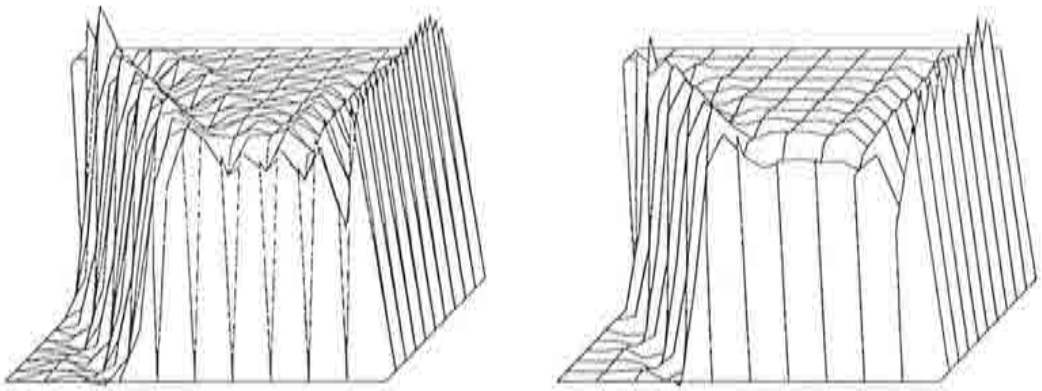

Figure 1.18 Results of Example 1.4 using quadratic triangular (6 nodes) and biquadratic quadrilateral (9 nodes) elements with the upwind function (1.93)

The motivation of the method has been shown for a very simple problem. Nevertheless, all the features of 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, nodally exact results will be obtained whenever they can be found for the homogeneous equation. The same holds true for linear elements if the source term is piecewise constant.
- *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: SUPG, 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: SUPG, 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.* 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. *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. Anal.*, 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/Upwind-Petrov-Galerkin 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 SUPG 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) (cf. [Pi]).

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-Upwind/Petrov-Galerkin method (SUPG) 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

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

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

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

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

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

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

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

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_{su} are evaluated at time $t = t^n$, we will write a^n_{su} and similarly for the linear form l_{su}. 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

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

This equation (2.20) may be rewritten as

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

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 bilinear forms $(\cdot,\cdot)_{su}$, a_{su} and the linear form l_{su} 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* $\mathbf{M}$ 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 are measured in the discrete L^2 norm [Co1]. So, in what follows

the choice of τ^e and therefore the perturbation of the test functions (2.19) that defines the SUPG 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-Upwind/Petrov-Galerkin 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 SUPG 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 $0 = x_0 < x_1 < \ldots < x_{N_{el}} = \ell$ be a uniform partition of diameter h of the interval $[0, \ell]$. Element e, $e = 1, 2, \ldots, N_{el}$, 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 $x_{e-\frac{1}{2}} := \frac{1}{2}(x_{e-1} + x_e)$. If $\boldsymbol{\phi} = \boldsymbol{\phi}(t)$ is the vector containing the nodal unknowns of $\phi(x,t)$, the application of the SUPG method to (2.29)–(2.31) will lead to an initial value problem of the form (2.22):

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

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 SUPG 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

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

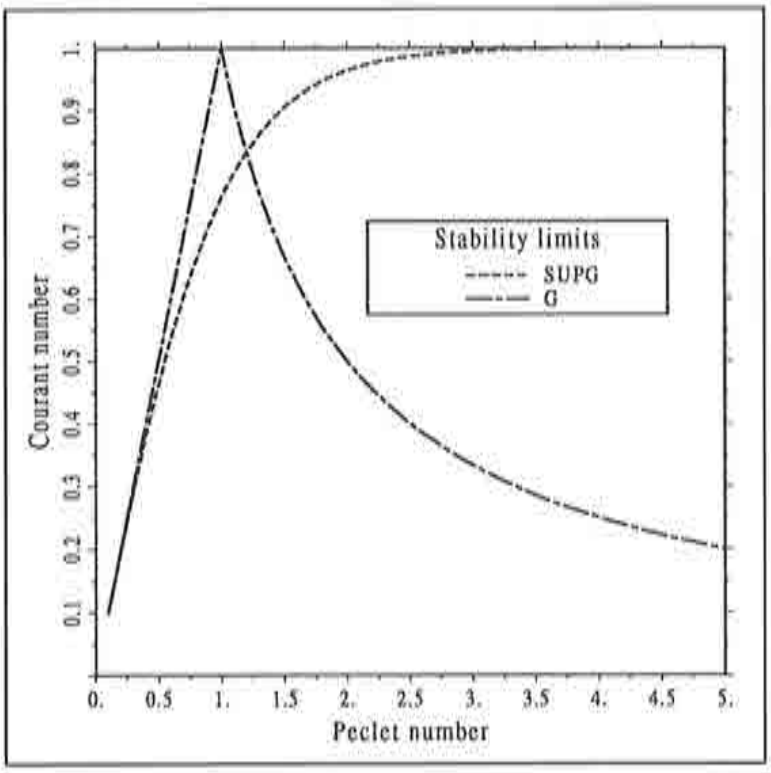

Figure 2.1 Stability limits for the convection-diffusion equation using linear elements. G: Galerkin method, SUPG: Streamline-Upwind/Petrov-Galerkin method.

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 $\alpha = \min(\frac{7}{3}, 1)$ (SUPG 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 SUPG 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 SUPG 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 $\alpha = \min(\frac{7}{3}, 1)$.

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

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

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 SUPG 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:

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

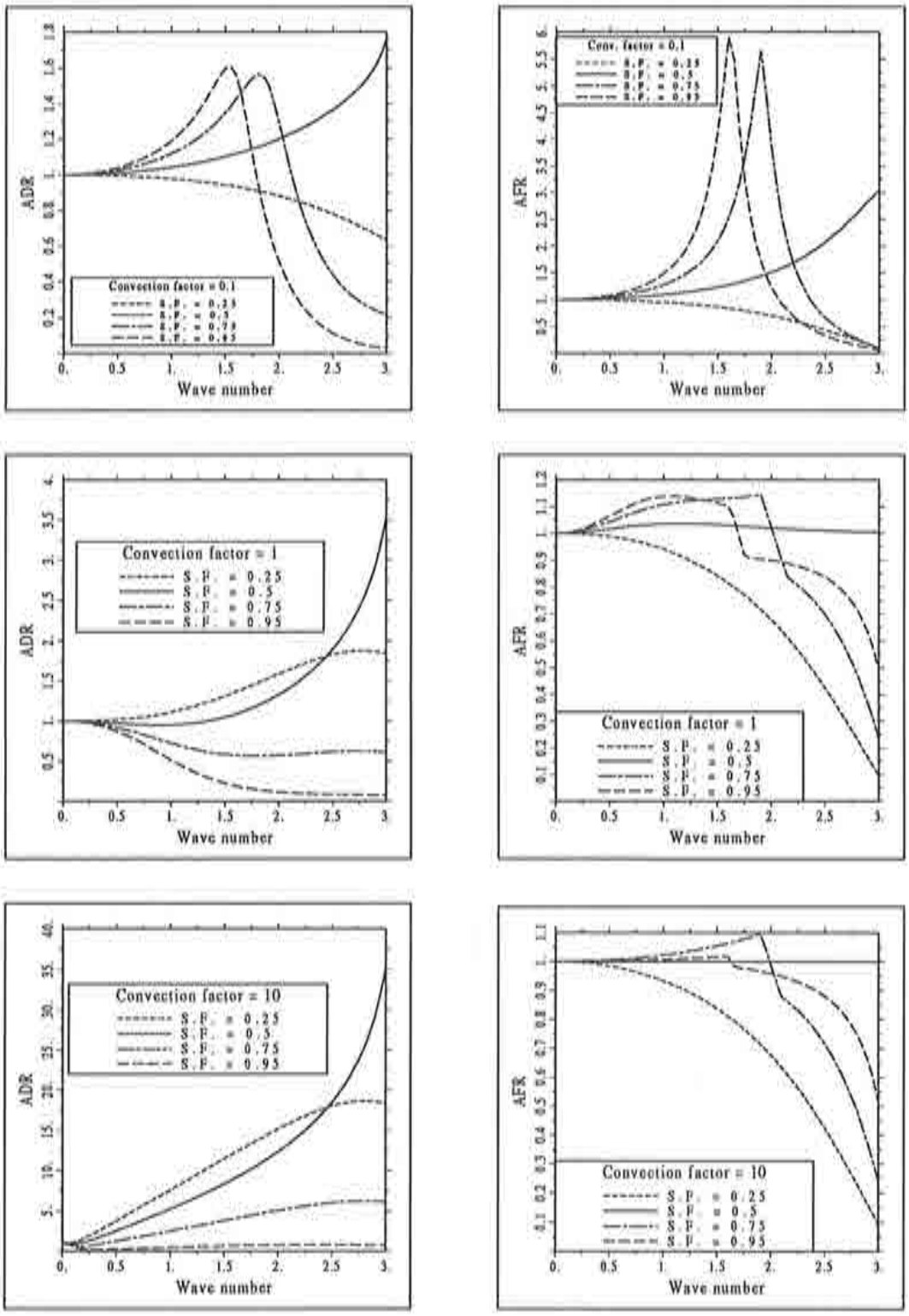

Figure 2.2 Algorithmic damping ratio (ADR) and algorithmic frequency ratio (AFR) for the SUPG method using linear elements and for different values of the security factor (S.F.) c_0 and the convection factor γ_0.

• Extreme nodes:

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

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

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

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

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

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

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

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 $\alpha = \beta = \min(\frac{7}{6}, \frac{1}{2})$ 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 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.

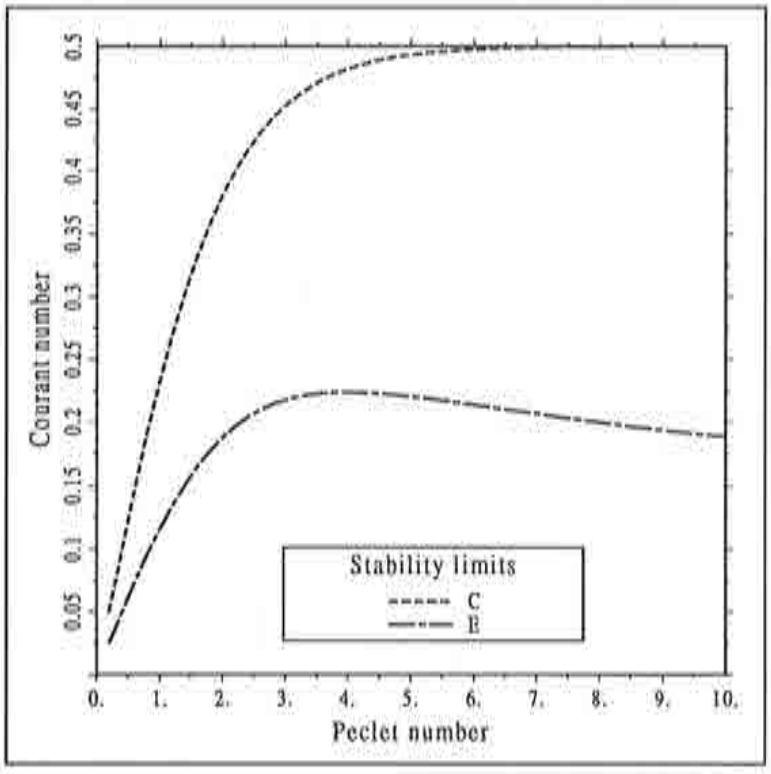

Figure 2.3 Stability limits for the convection-diffusion equation using quadratic elements. C: Central nodes, E: extreme nodes.

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

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

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

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

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,*

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

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

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

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 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 SUPG 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

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

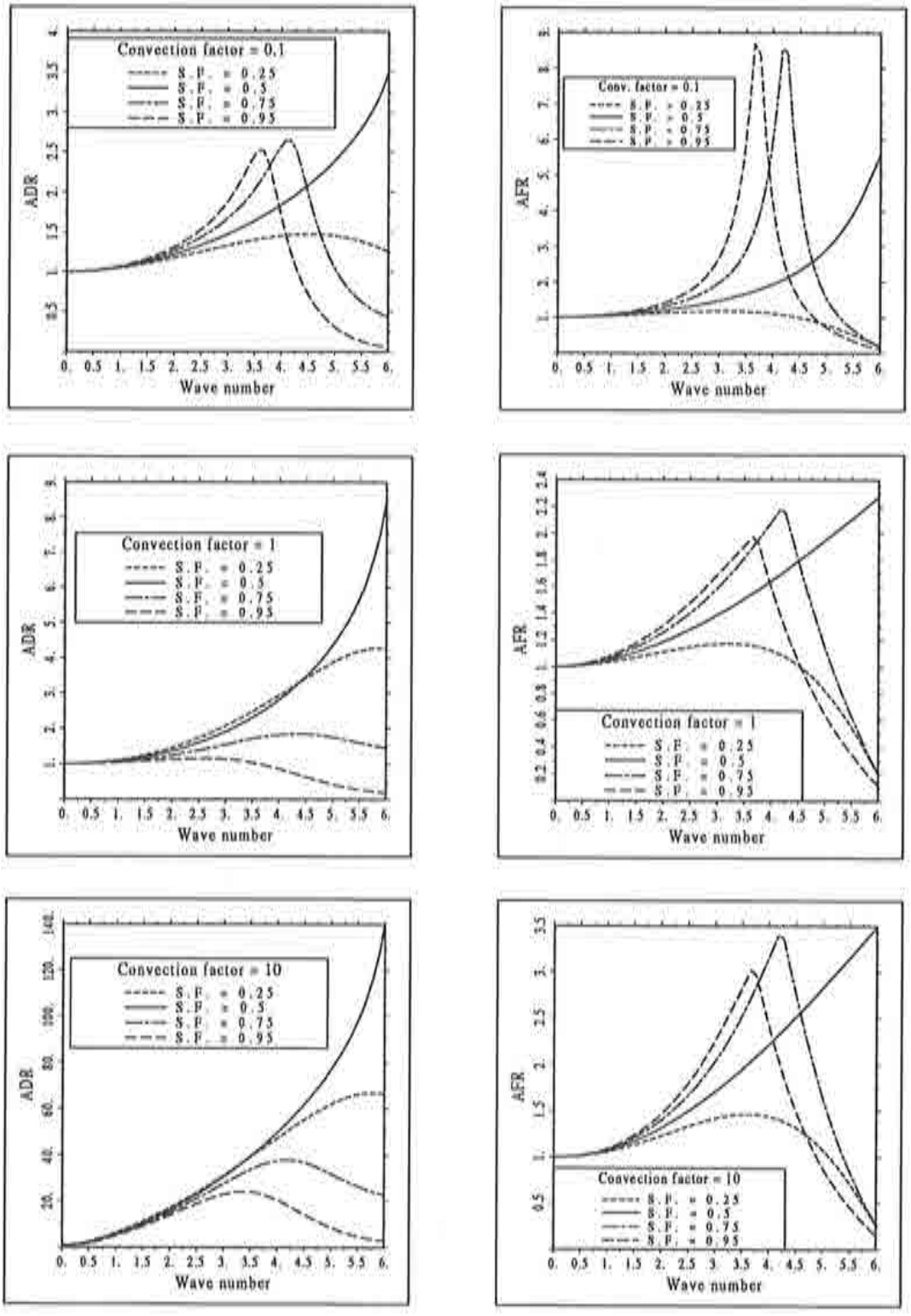

Figure 2.4 Algorithmic damping ratio (ADR) and algorithmic frequency ratio (AFR) for the SUPG method using quadratic elements and for different values of the security factor (S.F.) c_0 and the convection factor γ_0.

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

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

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 $\bar{\mathbf{M}}_d^e$ 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 $0 \le \eta \le 1$, then $\mu' \le 0$. 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 integration 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

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

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

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

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

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

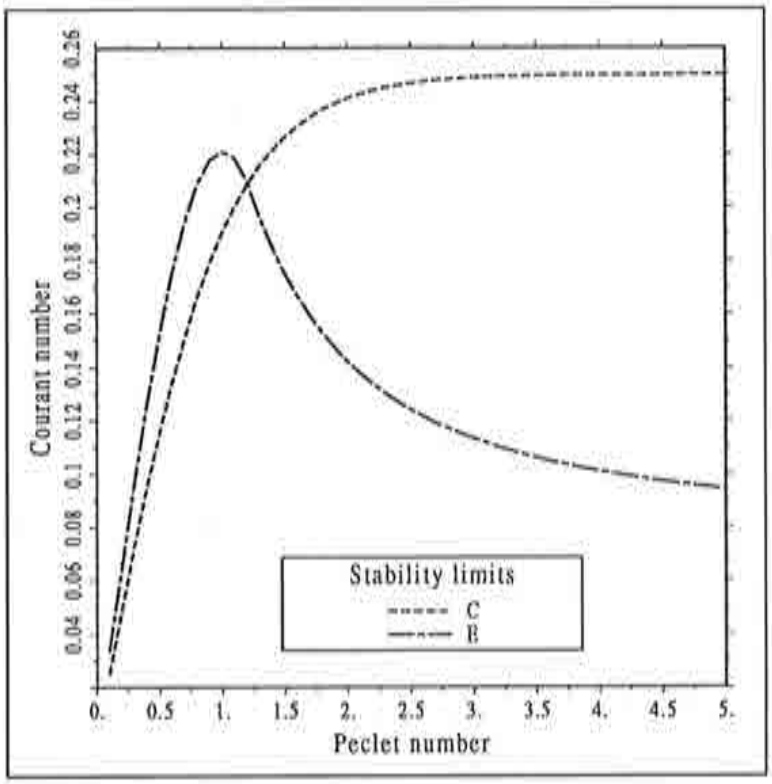

Figure 2.5 Stability limits for the convection-diffusion equation using quadratic hierarchic elements. C: Central nodes, E: extreme nodes.

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,*

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

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

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

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.

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 SUPG method. This result is indistinguishable from the analytical solution given above. Both using the Galerkin and the SUPG 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.

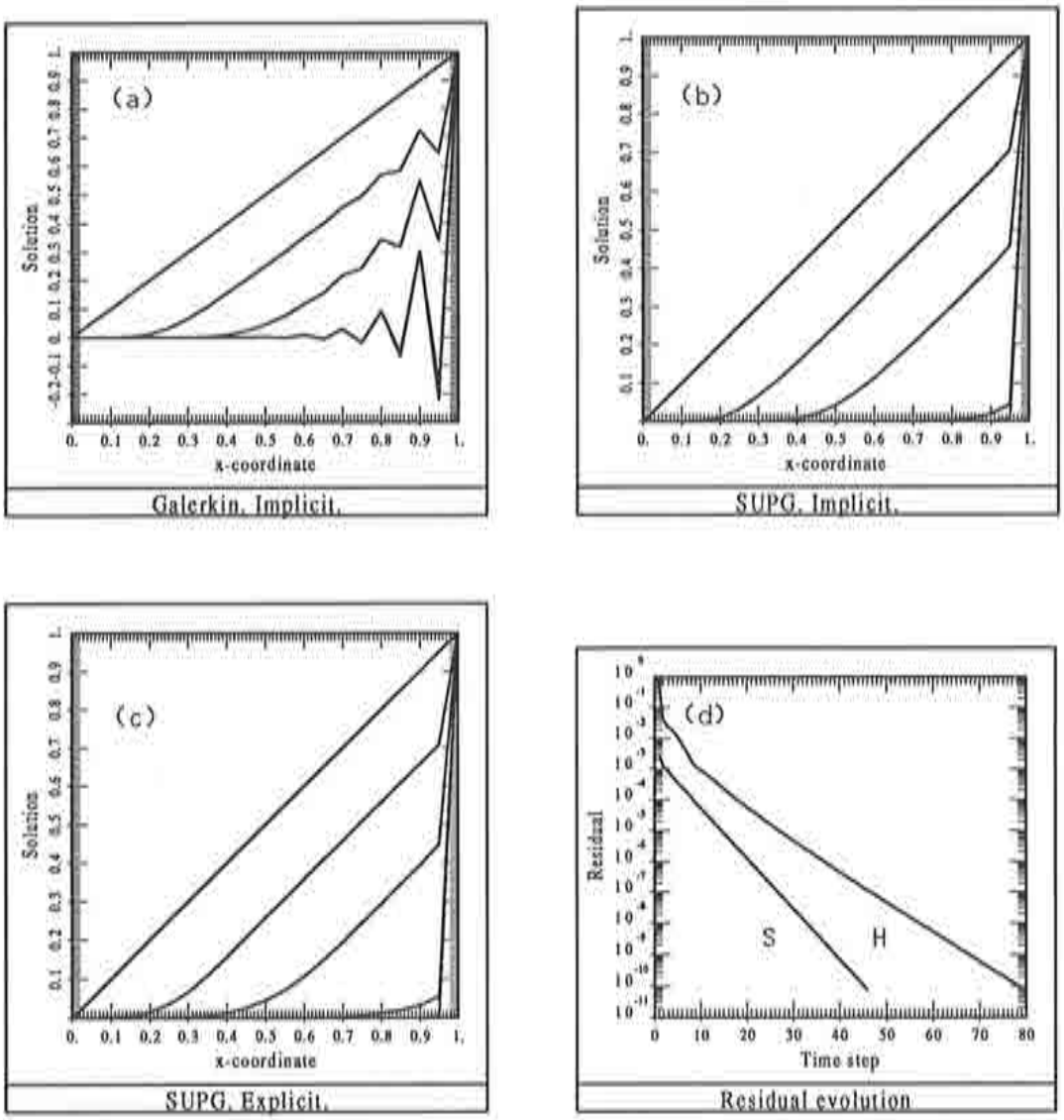

Figure 2.6 Results for example 2.1. (a): Galerkin solution using the Crank-Nicolson scheme. (b): SUPG solution using the Crank-Nicolson scheme. (c): SUPG solution using the forward Euler scheme. (d): Evolution to the steady-state using the standard (S) and the hierarchic (H) shape functions.

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

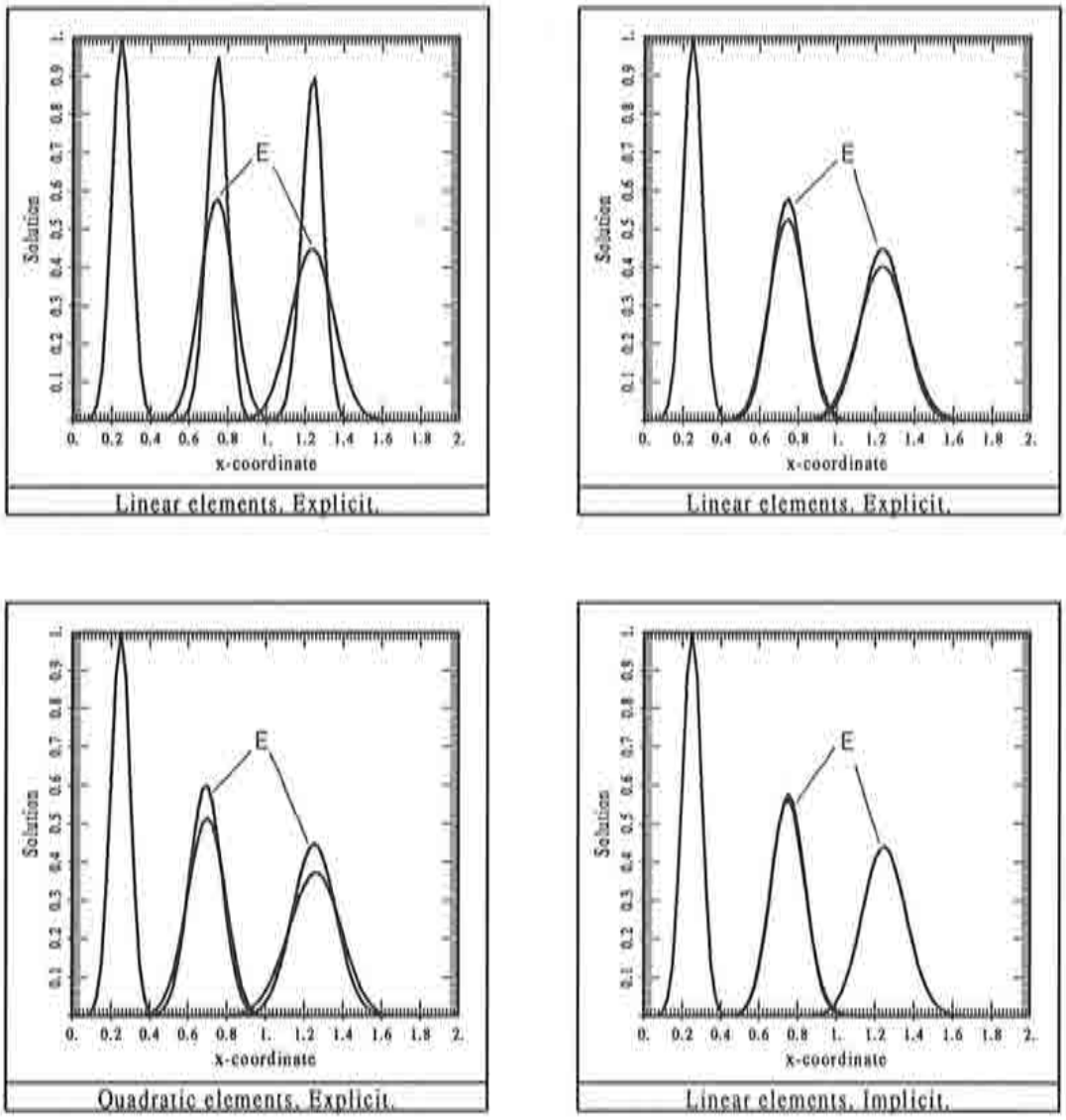

Figure 2.7 Results for example 2.2. Marker 'E' stands for exact solution. (a): Linear elements and forward Euler with $c_0 = 1$. (b): Linear elements and forward Euler with $c_0 = 0.5$. (c): Quadratic elements and forward Euler with $c_0 = 1$. (d): Linear elements and Crank-Nicolson.

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.

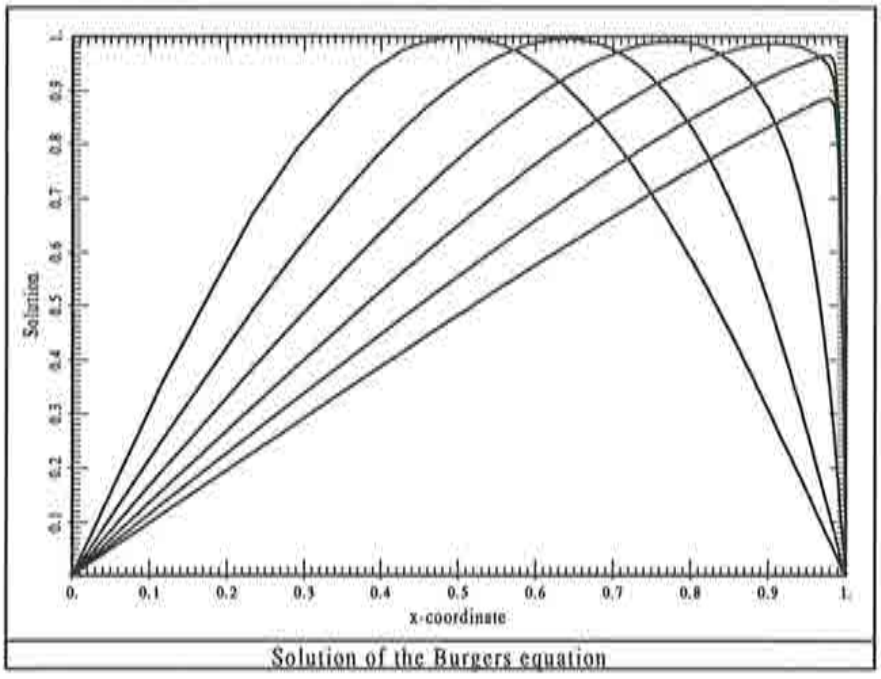

Figure 2.8 Solution of example 2.3

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

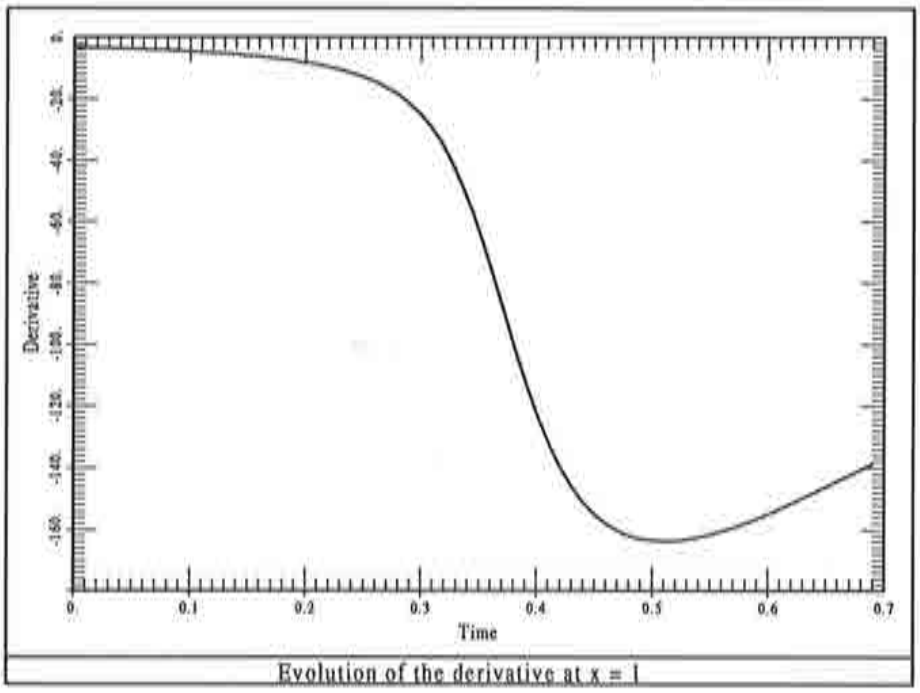

Figure 2.9 Evolution of the derivative at $x = 1$ for example 2.3

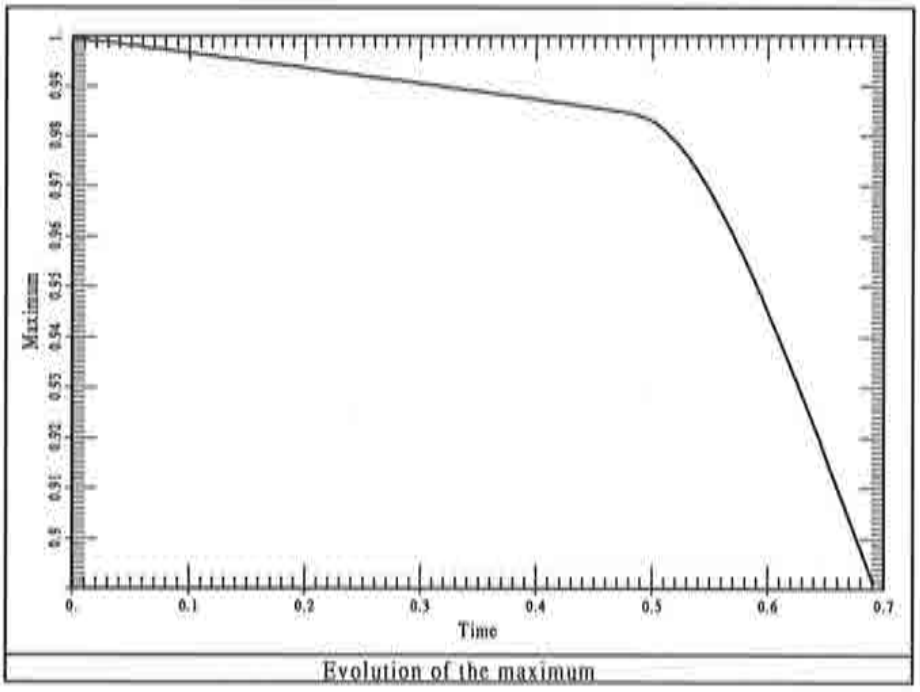

Figure 2.10 Evolution of the maximum value of the solution for example 2.3

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 SUPG 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	144.15
4	4 (2×2)	32.73
6	4	34.22
9	9 (3×3)	28.87

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

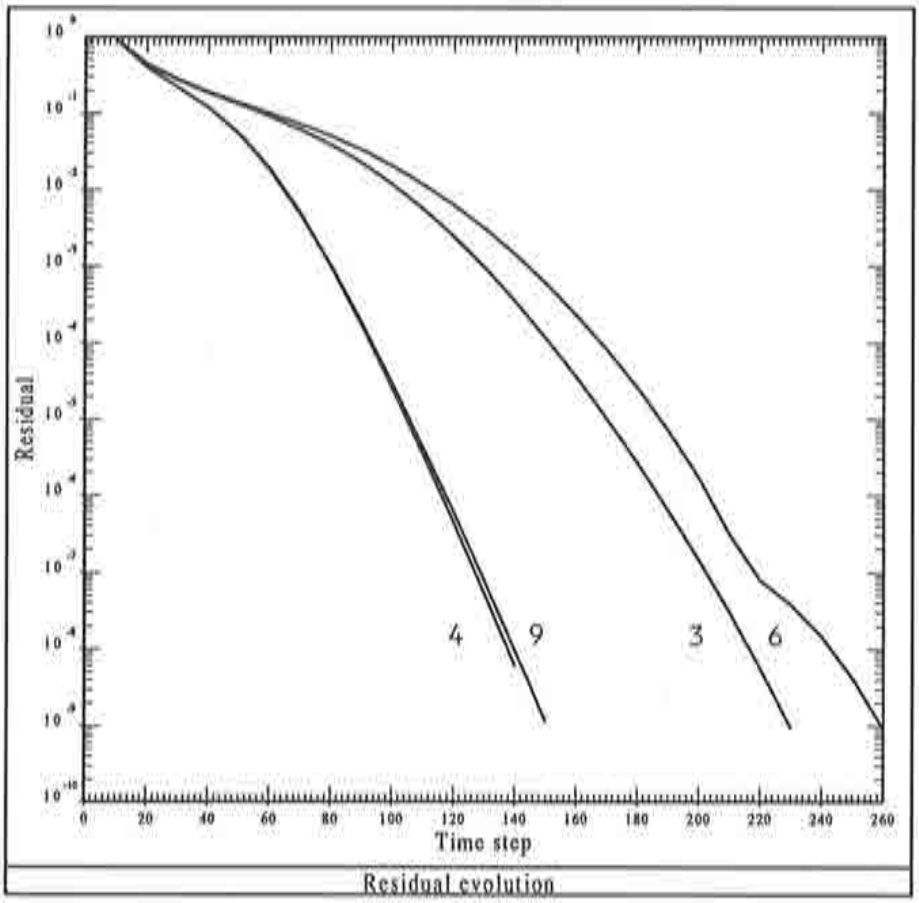

Figure 2.11 Evolution towards the steady-state in example 1.2. The number of nodes of the elements is indicated on each curve.

(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 SUPG 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 back-

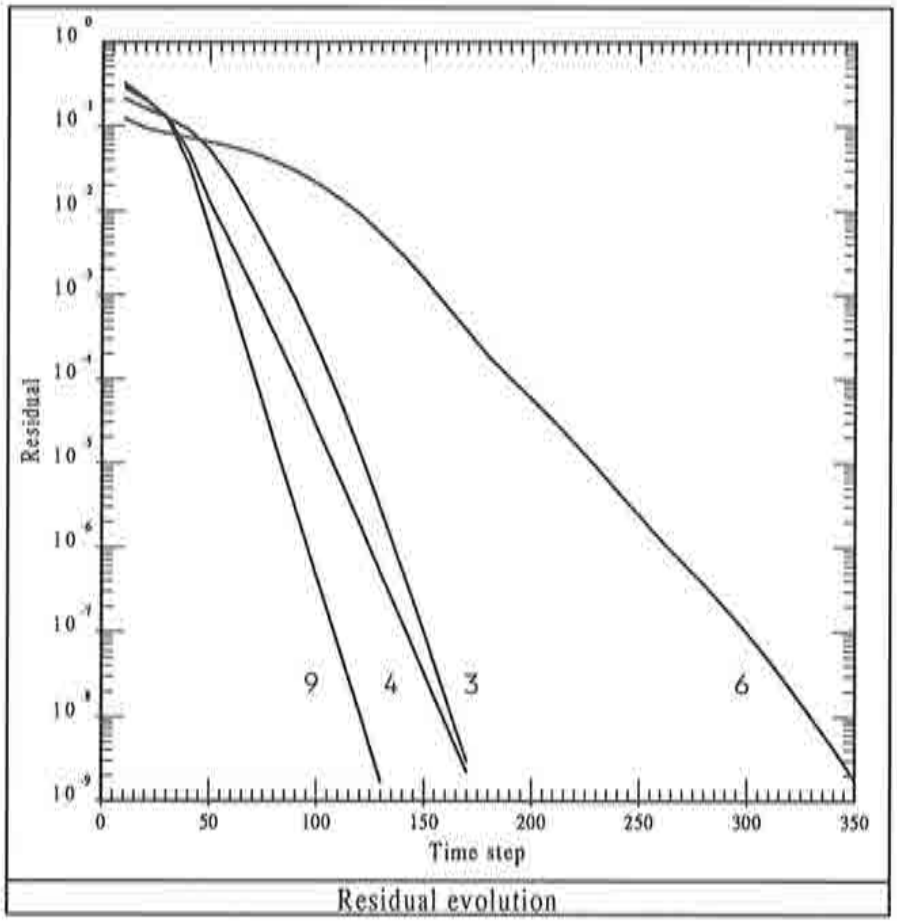

Figure 2.12 Evolution towards the steady-state in example 1.4. The number of nodes of the elements is indicated on each curve.

ward Euler method in some situations has been discussed, as well as the choice of the intrinsic time of the SUPG 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.

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
$c \le \begin{cases} \min(\mu' c_1, 2\mu c_2) & \text{if } \gamma > \gamma_c \\ \min(\mu' c_1, 2\mu c_3) & \text{if } \gamma \le \gamma_c \end{cases}$	$c \le \min\left(\dfrac{\mu'}{4}, \dfrac{\mu}{8}\right)$	$\Delta t \le \min\left(\dfrac{\mu' h^2}{8k}, \dfrac{\mu h^2}{3k}\right)$

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.* Grade thesis, Universitat Politècnica de Catalunya (1989)

[Co2] R. Codina. Generació de malles estructurades a partir d'equacions el.líptiques. *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

[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).

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

[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 discretization: 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)

CHAPTER 3

A DISCONTINUITY–CAPTURING CROSSWIND–DISSIPATION FOR THE STEADY–STATE PROBLEM

3.1 Introduction

In Chapter 1 we have presented the basic formulation of the Streamline-Upwind/Petrov-Galerkin (SUPG) method. As it has already been noted there, this method does not preclude the presence of overshoots and undershoots in the vicinity of sharp gradients of the solution. Near optimal *global* error estimates have been obtained (Section 1.2) and it is also possible to obtain near optimal *local* error estimates outside a small neighborhood containing the layers [JNP], [Na]. This ensures that in the presence of sharp layers of the solution it will not be globally deteriorated.

In certain situations, not even the small overshooting and undershooting found using the SUPG method are permissible. This happens for instance in the numerical simulation of compressible flow problems, where the solution may develop discontinuities (shocks) whose poor resolution may affect the global stability of the numerical calculations due to the nonlinear nature of the problem. Although we will not consider this application, the numerical solution of the convection-diffusion equation with a very high Péclet number provides a good model to develop numerical strategies to remove small oscillations about abrupt layers of the solution.

The reason why overshooting and undershooting appear using the SUPG method is that it is neither a *monotone* nor a *monotonicity preserving* method. A numerical method is said to be monotone if the numerical solution for all time steps retains the sign of the previous time step at all the nodes of the spatial mesh. If only the monotonicity of the initial data is maintained, the method is called monotonicity preserving [LV]. To design a high order accurate and monotone method is not easy. Godunov's theorem (cf. [LV]) establishes that a linear, monotonicity preserving method is at most first order accurate. Therefore, the only feasible way to achive the goals of high accuracy in regions where the solution is smooth and to avoid oscillations about layers is to design a *nonlinear* method, that is, a numerical scheme which depends itself on the numerical solution. The main idea of any shock-capturing (or discontinuity-capturing) technique is to increase the amount of numerical dissipation in the neighborhood of layers.

Several shock-capturing methods have been developed, both using finite difference

and finite element techniques. We will not treat here the former, for which a vast literature exists. We refer to the books [Hi], [LV], [OB], [PT] for a review of these methods, with special reference to compressible flow problems and systems of conservation laws. Concerning finite element methods, some of them are reviewed in Section 3.2.

For the particular case of the convection-diffusion equation, a possible way to treat the problem is the satisfaction of the discrete maximum principle (see [IK] for a thorough description of early finite element methods designed with this property). In Section 3.3 we adopt this point of view and consider several particular cases (1D problems with linear elements, linear and multilinear elements in multidimensional problems). Although from these examples the main conclusion is that it is difficult to take the discrete maximum principle as a starting point to design shock-capturing techniques, it provides the underlying idea for the method proposed in Section 3.4. Assuming that the streamline dissipation introduced by the SUPG method is enough to avoid oscillations in this direction, only the *crosswind* diffusion has to be increased. For consistency, the new dissipation added must be proportional to the element residual and, for accuracy, it must vanish quickly in regions where the solution is smooth and also where the convective term of the residual is small. The previous study of the discrete maximum principle is used to set the expression of the numerical crosswind dissipation. All these ideas are developed in Section 3.4, where also some results of Johnson *et al.* [JSW] are discussed. Numerical results using this new approach are presented in Section 3.5, showing a good resolution of the layers and also excellent convergence properties, in the sense to be precised later.

3.2 Some shock–capturing techniques

The problem we shall consider in this chapter is the same as in Chapter 1, namely, (1.19)–(1.21), that we recall here:

$$\mathbf{u}\cdot\nabla\phi - \nabla\cdot(\mathbf{k}\cdot\nabla\phi) = f, \qquad \text{in } \Omega \tag{3.1}$$

$$\phi = g, \qquad \text{on } \Gamma_D \tag{3.2}$$

$$\mathbf{n}\cdot\mathbf{k}\cdot\nabla\phi = r, \qquad \text{on } \Gamma_N \tag{3.3}$$

In Section 1.1.3 we have seen that the basic idea behind the SUPG method is to introduce numerical diffusion along the streamlines in a consistent manner. The streamline as the best upwind direction was questioned by Mizukami & Hughes in Reference [MH], and an upwind scheme satisfying the maximum principle was especially developed for linear triangular elements. Another monotone algorithm was presented by Rice & Schnipke in Reference [RS], now for bilinear quadrangular elements. Both methods are restricted to the elements they have been designed for, and no generalization seems easy. Since the oscillations observed using the SUPG formulation were placed in directions normal to the gradient of transported quantity, Hughes *et al.* proposed to introduce another diffusion in this direction [HMM], in a similar way to that proposed by Davis for finite differences [Da]. This new diffusion is consistently introduced as another term in the weighting functions called *discontinuity capturing*. Extensions to systems were studied in Refence [HM]. The method was initially adopted by Johnson & Szepessy in Reference [JS] using space-time finite element discretizations and, with a slight modification, used to prove convergence for the inviscid Burgers equation to

an entropy solution in Reference [JSH]. Other discontinuity capturing terms have been proposed, altough all of them keeping the SUPG (or GLS) terms (see, e.g., Tezduyar & Park [TP], Galeão & Dutra do Carmo [DG], [GD] and Shakib's thesis [Sh]).

The purpose of this section is basically to review two of these methods, namely, those proposed in References [HMM] and [GD], and to introduce a slight modification that allows to improve their behavior.

Let us consider the situation depicted in Figure 3.1, where $\mathbf{u}_{||}$ is the projection of the velocity $\mathbf{u}$ onto $\nabla\phi_h$.

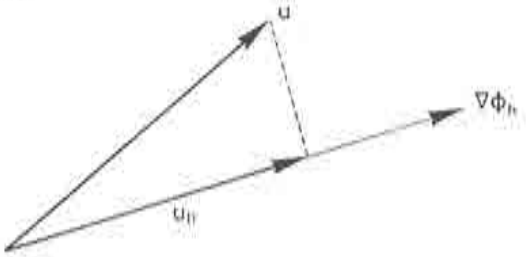

Figure 3.1 Definition of $\mathbf{u}_{||}$.

The expression of $\mathbf{u}_{||}$ will be (for $|\nabla\phi_h| \neq 0$):

$$\mathbf{u}_{||} = \frac{\mathbf{u}\cdot\nabla\phi_h}{|\nabla\phi_h|^2}\nabla\phi_h \tag{3.4}$$

Clearly, $\mathbf{u}_{||}\cdot\nabla\phi_h = \mathbf{u}\cdot\nabla\phi_h$, and for any value of the scalar s if we introduce the vector

$$\mathbf{v} = (1-s)\mathbf{u} + s\mathbf{u}_{||} \tag{3.5}$$

we will have that $\mathbf{v}\cdot\nabla\phi_h = \mathbf{u}\cdot\nabla\phi_h$, so that condition (1.32) holds and therefore we know from Chapter 1 how to introduce a numerical diffusion along the curves tangent to $\mathbf{v}$.

Since numerical experiments show that the oscillations that still remain using the SUPG method are normal to $\nabla\phi_h$, the first natural idea is to introduce another numerical dissipation in the direction of $\mathbf{u}_{||}$ [HMM]. This can be easily done by taking the weighting functions as $\psi_h + \zeta_h$, where

$$\zeta_h = \tau_1^e\mathbf{u}^e\cdot\nabla\psi_h + \tau_2^e\mathbf{u}_{||}^e\cdot\nabla\psi_h \tag{3.6}$$

instead of the expression (1.36) that defines the SUPG formulation. Omitting the superscript referring to element values, Hughes *et al.* proposed in Reference [HMM] two alternatives for the choice of τ_1 and τ_2, leading to the methods they called DC1 and DC2 (standing for discontinuity capturing of type 1 and of type 2). These two possibilities are:

$$\begin{aligned} &\text{DC1}: \quad \tau_1 = \tau, \qquad \tau_2 = \tau_{||} \\ &\text{DC2}: \quad \tau_1 = \tau, \qquad \tau_2 = \max(0, \tau_{||} - \tau) \end{aligned} \tag{3.7}$$

where τ is computed as indicated in (1.36) and $\tau_{||}$ is computed as τ but replacing the velocity $\mathbf{u}$ by its projection onto $\nabla\phi_h$, $\mathbf{u}_{||}$ (recall that this will affect the calculation of the upwind function α and the characteristic element length h appearing in Eqn.(1.36)).

From expression (3.6) it is clear that both a streamline diffusion and a diffusion in the direction of $\mathbf{u}_{\|}$ will be introduced. The choice DC2 seeks to avoid the doubling effect of DC1 along the streamlines. The final dissipation in this direction will be the maximum between the one induced by the SUPG term and the discontinuity capturing (DC) term.

However, let us look closely at the dissipation introduced by the second term in (3.6). If we only consider what happens when the convective term of Eqn.(3.1) is weighted, we will have that

$$\left(\tau_2 \mathbf{u}_{\|} \cdot \nabla \psi_h\right)\left(\mathbf{u} \cdot \nabla \phi_h\right) = \tau_2 \frac{\left(\mathbf{u} \cdot \nabla \phi_h\right)^2}{|\nabla \phi_h|^2} \nabla \psi_h \cdot \nabla \phi_h \tag{3.8}$$

that is, an *isotropic* diffusion has been introduced in the weak form of Eqn.(3.1). This will always happen if a rank-one diffusion tensor is introduced with eigenvector $\nabla \phi_h$.

Let us write the residual of the differential equation (3.1) within each element as

$$\mathcal{R}(\phi_h) := \mathbf{u} \cdot \nabla \phi_h - \nabla \cdot (\mathbf{k} \cdot \nabla \phi_h) - f \tag{3.9}$$

When the DC term in (3.6) is multiplied by this residual we will have that

$$\left(\tau_2 \mathbf{u}_{\|} \cdot \nabla \psi_h\right) \mathcal{R}(\phi_h) = \left[\tau_2 \frac{\mathbf{u} \cdot \nabla \phi_h}{|\nabla \phi_h|^2} \mathcal{R}(\phi_h)\right] \nabla \psi_h \cdot \nabla \phi_h \tag{3.10}$$

It may happen that

$$\operatorname{sgn}\left[(\mathbf{u} \cdot \nabla \phi_h) \mathcal{R}(\phi_h)\right] = -1 \tag{3.11}$$

so that a *negative* numerical diffusion may have been introduced. It must also be noticed that since the dissipation introduced by the DC term is isotropic, it may reduce the streamline diffusion introduced by the SUPG term when condition (3.11) holds true.

These problems are circumvented if the method proposed by Galeão & Dutra do Carmo [GD] is employed. The basic idea of this method is the following. Assume that the upwind direction is given by the vector

$$\mathbf{v}_h = \alpha_t \mathbf{w}_h + \beta_t (\mathbf{u} - \mathbf{w}_h) \tag{3.12}$$

where α_t and β_t are parameters to be defined later and $\mathbf{w}_h$ is a vector chosen as the solution of the following problem:

$$\begin{aligned} &\text{Minimize} \quad \|\mathbf{w}_h - \mathbf{u}\|_{0,\Omega}^2 \\ &\text{Subject to} \quad \mathbf{w}_h \cdot \nabla \phi_h - \nabla \cdot (\mathbf{k} \cdot \nabla \phi_h) - f = 0 \end{aligned} \tag{3.13}$$

The solution of this variational problem is given by

$$\mathbf{w}_h = \mathbf{u} - \lambda \nabla \phi_h, \quad \lambda = \frac{\mathcal{R}(\phi_h)}{|\nabla \phi_h|^2} \tag{3.14}$$

and hence the upwind direction will be given by

$$\mathbf{v}_h = \alpha_t \mathbf{u} + (\beta_t - \alpha_t) \lambda \nabla \phi_h \tag{3.15}$$

It is observed from (3.12) and (3.13) that the main idea of the method is to modify the streamline as the upwind direction so that the new one be as close as possible to it but satisfying the differential equation within each element.

If now we take as in References [DG], [GD] $\alpha_t = \tau_1$ and $\beta_t - \alpha_t = \tau_2$, the final result is that instead of using $\mathbf{u}_{||}$ in methods DC1 and DC2 one should take the vector

$$\mathbf{u}_r := \frac{\mathcal{R}(\phi_h)}{|\nabla\phi_h|^2}\nabla\phi_h \tag{3.16}$$

So the perturbation ζ_h will be given by

$$\zeta_h = \tau_1 \mathbf{u}\cdot\nabla\psi_h + \tau_2 \mathbf{u}_r\cdot\nabla\psi_h \tag{3.17}$$

instead of the expression (3.6).

This formulation was called *consistent approximate upwind* (CAU) method by Galeão & Dutra do Carmo.

The problem of the negative diffusion of the DC method of Hughes *et al.* is not present in this case. Instead of (3.10) we will now have

$$(\tau_2\mathbf{u}_r\cdot\nabla\psi_h)\,\mathcal{R}(\phi_h) = \tau_2\frac{\mathcal{R}(\phi_h)^2}{|\nabla\phi_h|^2}\nabla\psi_h\cdot\nabla\phi_h \tag{3.18}$$

and if τ_2 is computed as indicated before the isotropic diffusion introduced by this method will be given by

$$k_{\text{cau}} = \frac{1}{2}\alpha(\mathbf{u}_r)h(\mathbf{u}_r)\frac{|\mathcal{R}(\phi_h)|}{|\nabla\phi_h|} \tag{3.19}$$

with the upwind function α and the characteristic element length h calculated using the vector $\mathbf{u}_r$.

Summarizing, the CAU method consists in adding a diffusion proportional to the discrete residual of the differential equation within each element. This is also what Johnson *et al.* proposed in Reference [JSH]. The same approach was used by Shakib in his thesis [Sh].

To close this section, let us consider a slight modification of both the DC and the CAU methods. In order to avoid the superposition of the numerical streamline diffusion and the diffusion in the direction of $\mathbf{u}_{||}$ or $\mathbf{u}_r$, one can take the perturbation ζ_h of the test function as

$$\zeta_h = \tau\mathbf{u}^*\cdot\nabla\psi_h \tag{3.20}$$

with

$$\begin{aligned} \mathbf{u}^* &= (1-s)\mathbf{u} + s\mathbf{u}_{||} \quad \text{(modified DC)} \\ \text{or}\quad \mathbf{u}^* &= (1-s)\mathbf{u} + s\mathbf{u}_r \quad \text{(modified CAU)} \end{aligned} \tag{3.21}$$

The term $\tau(1-s)\mathbf{u}$ will introduce a streamline diffusion of value

$$k_s = \tau|\mathbf{u}|^2(1-s)$$

whereas for the modified CAU method an isotropic diffusion of value

$$k_{\text{iso}} = \tau|\mathbf{u}_r|^2 s$$

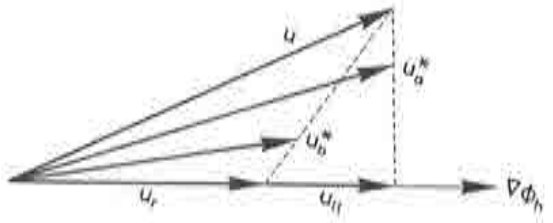

Figure 3.2 Modified upwind direction. $\mathbf{u}_a^*$ corresponds to the modification of the DC method and $\mathbf{u}_b^*$ to the modification of the CAU method.

will also be introduced. It has already been explained why the DC method (and also its modification) may result in the introduction of a negative numerical dissipation. The modified upwind direction $\mathbf{u}^*$ is shown in Figure 3.2.

From the expressions of k_s and k_{iso} it is clear that we must have $0 \leq s \leq 1$. A possible way to make the SUPG term dominant when $|\nabla\phi_h|$ is small and of little influence when this gradient is large is to take s as

$$s = \frac{|\nabla\phi_h|^2}{C + |\nabla\phi_h|^2} \tag{3.22}$$

In the numerical experiments presented in Section 3.5 we have always taken $C = 1$.

3.3 The discrete maximum principle

3.3.1 General considerations

For the continuous problem (3.1)–(3.3) it is well known that the maximum principle holds (see, e.g., [CH]), that is, the solution attains its maximum at the boundary when the source term f is non-positive. The question is whether this property is also inherited by the discrete problem or not, in a sense to be precised later.

The discrete maximum principle (DMP, for short) has important consequences in what concerns the convergence properties of the numerical scheme and, in particular, in uniform convergence. Here we will briefly state the main results obtained by Ciarlet & Raviart [CR] for the problem we consider in the case $\Gamma_N = \emptyset$, $\Gamma_D = \Gamma := \partial\Omega$.

The basic regularity assumptions for the data are that the components k_{ij} of the diffusion tensor $\mathbf{k}$ and the components u_i of the velocity field $\mathbf{u}$ belong to $L^\infty(\Omega)$, the source term f belongs to $L^p(\Omega)$, with $2 \leq N_{sd} < p$, and the extension $\tilde{g}$ of the Dirichlet data g to the whole domain belongs to the Sobolev space $W^{1,p}(\Omega)$, again with $2 \leq N_{sd} < p$. From Sobolev's imbedding Theorem, $\tilde{g}$ will be continuous in $\bar{\Omega}$, the closure of Ω ($\partial\Omega$ being Lipschitzian). Under these assumptions, it can be proved for the continuous problem that (cf. [CR])

$$\phi \in H^1(\Omega) \cap L^\infty(\Omega) \quad \text{and} \quad \|\phi\|_{L^\infty(\Omega)} \leq \|g\|_{L^\infty(\Gamma)} + C\|f\|_{L^p(\Omega)} \tag{3.23}$$

The important point is to obtain the analogue of the stability estimate for the discrete solution and, if possible, uniform convergence. These results were obtained in Reference [CR] for the particular case of linear (simplicial) finite elements and using

triangulations of strictly accute type, a concept to be defined in the following. All this, though, provided that the *the discrete maximum principle* holds true:

$$\max_{\hat{\Omega}} \phi_h \leq \max\{0, \max_{\Gamma} g_h\} \tag{3.24}$$

the function g_h being the finite element interpolant of g (g is continuous on Γ, under the assumptions stated above). More precisely, the basic results of Reference [CR] are:

$$\|\phi_h\|_{L^\infty(\Omega)} \leq \|g\|_{L^\infty(\Gamma)} + C\|f\|_{L^p(\Omega)} \tag{3.25}$$

$$\text{If } \phi \in W^{1,p}(\Omega),\ p > N_{sd}, \text{ then } \lim_{h\to 0} \|\phi - \phi_h\|_{L^\infty(\Omega)} = 0 \tag{3.26}$$

$$\text{If } \phi \in W^{2,p}(\Omega),\ p > N_{sd}/2, \text{ then } \|\phi - \phi_h\|_{L^\infty(\Omega)} = O(h) \tag{3.27}$$

Clearly, estimate (3.27) is not optimal, in the sense that one would hope to obtain an $O(h^2)$ estimate for linear elements. For the particular case of elliptic problems (Eqn.(3.1) without the convective term) but considering also the effect of numerical integration, Wahlbin proved that [Wa]

$$\|\phi - \phi_h\|_{L^\infty(\Omega)} = O(h^{3-\varepsilon}) \tag{3.28}$$

using quadratic elements, referring to the work of Nitsche [Nt] for a proof of an estimate of the form

$$\|\phi - \phi_h\|_{L^\infty(\Omega)} = O(h^{2-\varepsilon}) \tag{3.29}$$

using linear elements. In both cases, ε is a positive constant arbitrarily small. To prove (3.28) one needs to have $f \in W^{1,3}(\Omega)$ and $f \in W^{1,2}(\Omega)$ is needed for proving (3.29).

Anyway, what is important for us is to know that if the DMP holds, then pointwise stability estimates and uniform convergence can be proved. In any case, the DMP is an important property of the numerical scheme, since it ensures monotonicity (for the steady state convection-diffusion equation) and that no spurious oscillations will appear, not even in the vicinity of sharp layers.

The crucial point will be to check condition (3.24). In the next subsection a sufficient condition will be stated for an abstract discrete problem and next it will be applied to several particular cases using finite elements. Unfortunately, this condition will be too restrictive to decide if the finite element method satisfies the DMP or not.

3.3.2 A sufficient condition for the discrete problem

Let N_{tp} be the total number of nodes of the finite element mesh and N_{fp} the number of interior nodes. The finite element discretization of the problem will lead to an algebraic system of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.30}$$

where $\mathbf{x}$ stands for the vector containing the nodal unknowns x_i, $i = 1, ..., N_{tp}$. The values x_i, $i = N_{fp}+1, ..., N_{tp}$ are known from the Dirichlet boundary conditions. Matrix $\mathbf{A}$, whose components will be denoted a_{ij}, will have dimensions $N_{fp} \times N_{tp}$ and the vector $\mathbf{b}$ coming from the source term will have components b_i, $i = 1, ..., N_{fp}$

Our purpose now is to give a condition on the matrix $\mathbf{A}$ from which it will be possible to ensure that the DMP holds, viz.,

$$\max_{i=1,...,N_{tp}} \{x_i\} = x_m, \quad \text{with} \quad N_{fp}+1 \leq m \leq N_{tp} \tag{3.31}$$

First, let us introduce the following definition [CR], [Ki]: the matrix $\mathbf{A}$ is called of *nonnegative type* if the following conditions hold:

$$a_{ij} \leq 0 \quad \text{for} \quad i \neq j, \quad i = 1, ..., N_{fp}, \; j = 1, ..., N_{tp} \tag{3.32}$$

$$\sum_{j=1}^{N_{tp}} a_{ij} \geq 0, \quad i = 1, ..., N_{fp} \tag{3.33}$$

Let us call $\mathbf{A}_r = [a_{ij}]$, $i, j = 1, ..., N_{fp}$. The sufficient condition mentioned above is the following (cf. [CR], [Ki]):

Theorem 3.1 *Assume that* $\mathbf{A}$ *is of nonnegative type,* $\mathbf{A}_r$ *is nonsingular and* $b_i \leq 0$, $i = 1, ..., N_{fp}$. *Then the discrete maximum principle as expressed in (3.31) holds.*

Proof: Let us write the jth equation of the linear system (3.30) as

$$a_{jj} x_j = b_j - \sum_{\substack{k=1 \\ k \neq j}}^{N_{tp}} a_{jk} x_k \tag{3.34}$$

From conditions (3.32) and (3.33), together with the fact that $\mathbf{A}_r$ is nonsingular, it follows that

$$a_{jj} > 0, \quad 1 + \sum_{\substack{k=1 \\ k \neq j}}^{N_{tp}} \frac{a_{jk}}{a_{jj}} \geq 0 \tag{3.35}$$

Since $b_j \leq 0$, $-a_{jk}/a_{jj} \geq 0$ for all $k \neq j$ and from (3.34) and (3.35) we obtain

$$\begin{aligned} x_j &= \frac{b_j}{a_{jj}} - \sum_{\substack{k=1 \\ k \neq j}}^{N_{tp}} \frac{a_{jk}}{a_{jj}} x_k \\ &\leq \frac{b_j}{a_{jj}} - \max_{k \neq j}\{x_k\} \sum_{\substack{k=1 \\ k \neq j}}^{N_{tp}} \frac{a_{jk}}{a_{jj}} \\ &\leq \frac{b_j}{a_{jj}} + \max_{k \neq j}\{x_k\} \\ &\leq \max_{k \neq j}\{x_k\} \end{aligned} \tag{3.36}$$

Let us argue by contradiction and suppose that

$$\begin{aligned} &\max_{k \neq j}\{x_k\} = x_m, \quad \text{with} \;\; 1 \leq m \leq N_{fp}, \\ &\text{and} \;\; x_k < x_m, \quad k = N_{fp} + 1, ..., N_{tp} \end{aligned} \tag{3.37}$$

From (3.36) we will have that $x_j \leq x_m$ and using the same argument as that to arrive at (3.36) for this m,

$$x_m \leq \max_{k \neq m}\{x_k\} = x_j$$

so that $x_m = x_j$. Without loss of generality, suppose that $m = 1$ and $j = 2$. Since we know that $x_1 = x_2$ we can eliminate the first equation in (3.30) and consider the system $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$, with

$$\begin{aligned}
&\mathbf{x}' = (x_2, ..., x_{N_{tp}})^T \\
&\mathbf{b}' = (b_2, ..., b_{N_{fp}})^T \\
&a'_{i,1} = a_{i+1,1} + a_{i+1,2}, \quad i = 1, ..., N_{fp} - 1 \\
&a'_{i,j} = a_{i+1,j+1}, \quad i = 1, ..., N_{fp} - 1, \quad j = 2, ..., N_{tp} - 1
\end{aligned}$$

Since $\mathbf{A}_r$ is nonsingular, $\mathbf{A}'$ will be of nonnegative type and (3.35) will also hold for its components. Using the same argument as before, we find that $x_2 = x_j$ for some j, that we may take as $j = 3$. Repeating this process we arrive at the conclusion that $x_1 = x_2 = ... = x_{N_{fp}}$ and in the last step the analogous of (3.36) will be

$$x_{N_{fp}} \leq \max_{k=N_{fp}+1,...,N_{tp}} \{x_k\}$$

which is a contradiction with the second statement in (3.37). Therefore, this assumption (3.37) must be false, that is, condition (3.31) holds true. □

3.3.3 Some particular cases

We shall check now the hypothesis of Theorem 3.1 for three different particular discretizations using finite elements. For all these cases it will be assumed that the problem is well posed and thus the matrix $\mathbf{A}_r$ is nonsingular. The condition $b_i \leq 0$, $i = 1, ..., N_{fp}$ will be very easy to verify and therefore our main concern will be to check if the matrix $\mathbf{A}$ is of nonnegative type, that is, to verify conditions (3.32) and (3.33). This last condition is trivial to prove when standard finite elements are used. In fact, using the SUPG formulation and denoting by $\psi_{h,i}$ the shape function associated to the node i we will have that

$$\begin{aligned}
\sum_{j=1}^{N_{tp}} a_{ij} &= \sum_{j=1}^{N_{tp}} a_{su}(\psi_{h,j}, \psi_{h,i}) \\
&= \int_\Omega \left[\psi_{h,i}\mathbf{u} \cdot \nabla \left(\sum \psi_{h,j} \right) + \nabla \psi_{h,i} \cdot \mathbf{k} \cdot \left(\sum \psi_{h,j} \right) \right] d\Omega \qquad (3.38) \\
&+ \sum_{e=1}^{N_{el}} \int_{\Omega^e} (\tau^e \mathbf{u} \cdot \psi_{h,i}) \left[\mathbf{u} \cdot \nabla \left(\sum \psi_{h,j} \right) - \nabla \cdot \left(\mathbf{k} \cdot \left(\sum \psi_{h,j} \right) \right) \right] d\Omega
\end{aligned}$$

where the summations are understood to be from $j = 1$ to $j = N_{tp}$ and the bilinear form $a_{su}(\cdot, \cdot)$ is given by (1.39). Since $\sum \psi_{h,j} = 1$ if standard finite elements are used, it follows from (3.38) that condition (3.33) holds with the equal sign.

So we only have to check condition (3.32). Moreover, since the assembly operator is linear, it has to be verified only for the element matrices and this is not difficult to do for some particular cases.

One dimensional problem using linear elements

The first case we shall consider is the one-dimensional convection-diffusion equation using linear uniform elements of length h, that is, the situation considered in Section 1.1.1. For this case it is possible to obtain explicit bounds for the upwind function α in order to satisfy the discrete maximum principle. The continuous problem is

$$\begin{aligned} u\,\frac{d\phi}{dx} - k\,\frac{d^2\phi}{dx^2} &= f(x), \quad 0 < x < \ell \\ \phi(0) &= \phi_0, \quad \phi(\ell) = \phi_\ell \end{aligned} \tag{3.40}$$

with $f(x) \leq 0$.

Proposition 3.1 *Assume that problem (3.40) is discretized using the Galerkin method and an artificial diffusion $k_a = \alpha h u/2$ is introduced. Then the numerical scheme satisfies the discrete maximum principle (DMP) iff the upwind function α is such that*

$$|\alpha| \geq 1 - \frac{1}{|\gamma|} \tag{3.41}$$

If the SUPG method is employed and f is piecewise constant, then the DMP holds provided that α verifies condition (3.41) and

$$|\alpha| \leq 1 \tag{3.42}$$

Proof: The off-diagonal components of the element stiffness matrix $\mathbf{K}^e$ are

$$K^e_{12} = -\frac{k}{h}(1 + \alpha\gamma - \gamma), \quad K^e_{21} = -\frac{k}{h}(1 + \alpha\gamma + \gamma),$$

both using the Galerkin method with artificial diffusion and the SUPG formulation. Requiring $K^e_{12} \leq 0$ and taking into account that $\mathrm{sgn}(\alpha) = \mathrm{sgn}(\gamma)$ leads to

$$\alpha\gamma = |\alpha||\gamma| \geq \gamma - 1, \quad |\alpha| \geq \mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}$$

Condition $K^e_{21} \leq 0$ yields

$$\alpha\gamma = |\alpha||\gamma| \geq -\gamma - 1, \quad |\alpha| \geq -\mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}$$

Both $K^e_{12} \leq 0$ and $K^e_{21} \leq 0$ iff

$$|\alpha| \geq \max\{\mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}, -\mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}\} = 1 - \frac{1}{|\gamma|}$$

Let us check now that $b_i \leq 0$. For the Galerkin method with artificial diffusion this is obvious, since $\psi_{h,i} \geq 0$ for all i. Thus, the DMP holds for this case. If the SUPG formulation is employed and f is piecewise constant, with value f_i in the element $[(i-1)h, ih]$, we will have

$$\begin{aligned} b_i &= \int_0^\ell \left(\psi_{h,i} + \frac{1}{2}\alpha h \frac{d\psi_{h,i}}{dx}\right) f\,dx \\ &= f_i \int_{(i-1)h}^{ih} \left(\psi_{h,i} + \frac{1}{2}\alpha\right) dx + f_{i+1} \int_{ih}^{(i+1)h} \left(\psi_{h,i} - \frac{1}{2}\alpha\right) dx \end{aligned} \tag{3.43}$$

for $i = 1, ..., N_{el} - 1$. For arbitrary values of f_i and f_{i+1} both integrals must be non-negative in order to ensure that $b_i \leq 0$. Since their values are

$$\frac{1}{2}h + \frac{1}{2}\alpha h \quad \text{and} \quad \frac{1}{2}h - \frac{1}{2}\alpha h$$

this will only happen if condition (3.42) holds. □

Remarks 3.1

(1) In Chapter 1 it has been proved that the SUPG method using the optimal upwind function given by (1.12) must give nodally exact results when f is piecewise constant, but not for arbitrary source functions f. Now we see that in the general case it is not even possible to satisfy the DMP, since from (3.43) it is observed that it is possible to choose $f \leq 0$ such that $b_i > 0$ for any $\alpha > 0$. When f is piecewise linear, it is easy to show that $|\alpha| \leq 2/3$ is needed in order to have $b_i \leq 0$, but if this condition is fulfilled it is possible to violate (3.41) for certain values of the Péclet number γ.

(2) If f is *constant*, condition (3.42) is not needed except for the case in which a Neumann type boundary condition is prescribed at the outflow. At this point, only one of the two integrals in (3.43) will appear in the expression of the associated component of $\mathbf{b}$.

(3) Recall that condition (3.42) has already been found in the previous two chapters using very different arguments. In Chapter 1 it has been shown that it is the condition under which no oscillations appear in the numerical solution, and in Chapter 2 we have seen that it is the condition that ensures that the diffusive stability limit of the forward Euler scheme in time dominates the convective limit (see Section 2.3.2). □

Multidimensional problem using simplicial linear elements

Suppose now that the domain Ω is discretized using linear $N_{sd}-$simplices. Under a certain condition on the finite element partition $\{\Omega^e\}$ it will also be possible in this case to obtain bounds for the upwind function in order to satisfy the DMP. The results to be presented here are an extension of the work of Kikuchi [Ki].

For each element e, let us introduce the following notation:

ρ^e: supremum of the diameter of the spheres inscribed in Ω^e
κ^e: minimum perpendicular length of Ω^e
λ^e: maximum perpendicular length of Ω^e
h^e: diameter of Ω^e
$h = \max_e\{h^e\}$

As usual, the finite partition is assumed to be regular [Ci]. This means that there exists a positive constant C_1 such that

$$\min_e \frac{\rho^e}{h^e} \geq C_1 \quad \text{as } h \to 0 \tag{3.44}$$

and in particular this implies that there exists another positive constant C_2 such that

$$\min_e \frac{\kappa^e}{\lambda^e} \geq C_2 \quad \text{as } h \to 0 \tag{3.45}$$

Let us also introduce the constant

$$\sigma^e := \max_{i \neq j} \frac{\nabla \psi^e_{h,i} \cdot \nabla \psi^e_{h,j}}{|\nabla \psi^e_{h,i}||\nabla \psi^e_{h,j}|} = \max_{i \neq j} \cos(\nabla \psi^e_{h,i}, \nabla \psi^e_{h,j}) \tag{3.46}$$

The main restriction of what follows is that we shall assume the finite element partition $\{\Omega^e\}$ to be of *strictly accute type*. By definition (cf. [CR], [Ki]), this means that there exists a constant $\sigma_0 > 0$ such that

$$\sigma^e \leq -\sigma_0, \quad e = 1, ..., N_{el} \tag{3.47}$$

For two-dimensional problems ($N_{sd} = 2$) this happens only if all the angles of the triangles are $< \pi/2$. Observe that $\sigma^e \geq -1$ and therefore $\sigma_0 \leq 1$.

To prove the following result we shall make two simplifications. The velocity $\mathbf{u}$ will be considered constant within each element and the diffusion tensor $\mathbf{k}$ isotropic and also piecewise constant, the diffusion coeffient being k. Since $\nabla \phi_h$ will be piecewise constant, the vector $\mathbf{u}_{||}$ defined in (3.4) will also be piecewise constant. The Péclet number computed with this vector will be denoted by $\gamma_{||}$.

Proposition 3.2 *Under the assumptions stated above, assume that problem (3.1)-(3.2) with $\Gamma_D = \partial\Omega$ and $f \leq 0$ is discretized using the Galerkin method and an isotropic artificial diffusion*

$$k_a = \frac{1}{2}\alpha^e h^e |\mathbf{u}^e_{||}| \tag{3.48}$$

is introduced within each element. Then the numerical scheme satisfies the DMP if the function α^e is such that

$$\alpha^e \geq \frac{C}{N_{sd}+1} - \frac{1}{\gamma^e_{||}}, \tag{3.49}$$

for a certain constant C.

Proof: Since $\psi_{h,i} \geq 0$ for each node i, it is clear that $b_i = \int_\Omega \psi_{h,i} f d\Omega \leq 0$. As before, since the assembly operator is linear, it suffices to prove condition (3.32) for each element stiffness matrix. Let us denote one of them by $\mathbf{K}^e$ and let $\mathbf{K}^e_{||}$ be the matrix calculated with $\mathbf{u}_{||}$. We shall have that

$$\mathbf{K}^e \mathbf{\Phi}^e = \mathbf{K}^e_{||} \mathbf{\Phi}^e$$

where $\mathbf{\Phi}^e$ is the vector of nodal unknowns of ϕ_h, and therefore it is enough to check (3.32) for $\mathbf{K}^e_{||}$.

Observe first that

$$(\lambda^e)^{-1} \leq |\nabla \psi^e_{h,i}| \leq (\kappa^e)^{-1} \tag{3.50}$$

$$\int_{\Omega^e} \psi^e_{h,i} d\Omega = \frac{\text{meas}(\Omega^e)}{N_{sd}+1} \tag{3.51}$$

for all $i = 1, ..., N_{sd}$.

The ij component of the matrix $\mathbf{K}^e_{||}$, for $i \neq j$, can be bounded as follows:

$$\begin{aligned}
(k^e_{||})_{ij} &= (k + k_a) \int_{\Omega^e} \nabla\psi^e_{h,i} \cdot \nabla\psi^e_{h,j} d\Omega + \int_{\Omega^e} \psi^e_{h,i} \mathbf{u}^e_{||} \cdot \nabla\psi^e_{h,j} d\Omega \\
&\leq (k + k_a)\text{meas}(\Omega^e)\sigma^e |\nabla\psi^e_{h,i}||\nabla\psi^e_{h,j}| + |\mathbf{u}^e_{||}||\nabla\psi^e_{h,j}| \frac{\text{meas}(\Omega^e)}{N_{sd}+1} \\
&\leq (k + k_a)\text{meas}(\Omega^e)\sigma^e \frac{1}{(\lambda^e)^2} + |\mathbf{u}^e_{||}| \frac{\text{meas}(\Omega^e)}{\kappa^e (N_{sd}+1)} \\
&= (\lambda^e)^{-2}\text{meas}(\Omega^e) \left[\frac{(\lambda^e)^2}{\kappa^e} \frac{|\mathbf{u}^e_{||}|}{N_{sd}+1} + \left(k + \frac{1}{2}\alpha^e h^e |\mathbf{u}^e_{||}| \right) \sigma^e \right]
\end{aligned}$$

where we have used (3.50) and (3.51). Taking into account that

$$\frac{(\lambda^e)^2}{\kappa^e} \leq \frac{\lambda^e}{\kappa^e} h^e \leq \frac{h^e}{C_2}$$

it follows that

$$(k^e_{||})_{ij} \leq \frac{1}{2}(\lambda^e)^{-2}\text{meas}(\Omega^e)|\mathbf{u}^e_{||}|h^e \left[\frac{2}{C_2(N_{sd}+1)} + \left(\frac{2k}{h^e|\mathbf{u}^e_{||}|} + \alpha^e \right) \sigma^e \right]$$

Therefore, the condition $(k^e_{||})_{ij} \leq 0$ will hold if

$$\frac{2}{C_2(N_{sd}+1)} + \left(\frac{1}{\gamma^e_{||}} + \alpha^e \right) \sigma^e \leq 0$$

Since $\sigma^e < 0$, this is equivalent to

$$\alpha^e \geq \frac{(-2/C_2\sigma^e)}{N_{sd}+1} - \frac{1}{\gamma^e_{||}}$$

The Proposition follows for $C = -2/C_2\sigma^e$. □

Remarks 3.2

(1) As for the one-dimensional problem, an upper bound for the upwind function α^e is needed when the SUPG formulation is used in order to have $b_i \leq 0$ for the case of piecewise constant source f, that is, $\alpha^e \leq C'$ for a certain constant C'. Let us prove this for each elemental contribution to the vector $\mathbf{b}$. If f^e is the value of f in element e,

$$\begin{aligned}
b^e_i &= f^e \int_{\Omega^e} \left(\psi^e_{h,i} + \frac{\alpha^e h^e}{2|\mathbf{u}^e|} \mathbf{u}^e \cdot \nabla\psi^e_{h,i} \right) d\Omega \\
&= f^e \left[\frac{\text{meas}(\Omega^e)}{N_{sd}+1} + \frac{\alpha^e h^e}{2|\mathbf{u}^e|} \mathbf{u}^e \cdot \nabla\psi^e_{h,i} \text{meas}(\Omega^e) \right] \qquad (3.52)
\end{aligned}$$

On the other hand, using the fact that $\lambda^e \geq \rho^e$ we get

$$\begin{aligned}
\frac{1}{N_{sd}+1} + \frac{\alpha^e h^e}{2|\mathbf{u}^e|} \mathbf{u}^e \cdot \nabla\psi^e_{h,i} &\geq \frac{1}{N_{sd}+1} - \frac{\alpha^e h^e}{2} |\nabla\psi^e_{h,i}| \\
&\geq \frac{1}{N_{sd}+1} - \frac{\alpha^e h^e}{2\kappa^e} \\
&\geq \frac{1}{N_{sd}+1} - \frac{\alpha^e h^e}{2\kappa^e} \frac{\lambda^e}{\rho^e} \\
&\geq \frac{1}{N_{sd}+1} - \frac{\alpha^e}{2C_1C_2}
\end{aligned}$$

and from (3.52) it follows that $b_i^e \leq 0$ provided that

$$\alpha^e \leq \frac{2C_1C_2}{N_{sd}+1}$$

But even though $b_i^e \leq 0$ the DMP may not hold since *the SUPG formulation only introduces streamline diffusion* and thus Proposition 3.2 does not guarantee the satisfaction of the DMP for this case. This argument is the basic idea of the method introduced in the next section.

(2) Proposition 3.2 might be useful if one wishes to use an artificial diffusion method. In fact, it explains why these methods may succeed in removing the oscillations of the Galerkin approach. □

A two-dimensional problem using bilinear elements

The situations considered so far are general. Now we shall analyse a very simple two-dimensional example using bilinear finite elements. Let the domain Ω be discretized using equal elements of length h_x in the $x-$direction and length h_y in the $y-$direction, and let $\mathbf{u} = (u, 0)$, with $u \geq 0$ (see Figure 3.3).

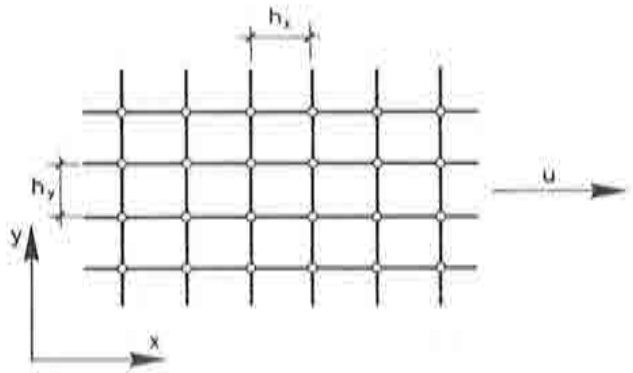

Figure 3.3 Description of the problem

Our purpose in analizing this case is again to get insight in the role played by the upwind function in the satisfaction of the DMP. For that, we shall assume that the physical diffusion is isotropic, of value k, and that numerical diffusion is added both in the $x-$ and the $y-$directions, so that the final diffusion to be considered in each direction is

$$k_x = k + \frac{1}{2}\alpha_x u h_x, \quad k_y = k + \frac{1}{2}\alpha_y u h_y \tag{3.53}$$

We want to study under which conditions on the functions α_x and α_y and also on the ratio $\eta := h_y/h_x$ is possible to satisfy the DMP. As in the previous cases, this reduces to check condition (3.32) for the element stiffness matrix $\mathbf{K}^e$, that shall be split into the diffusive contribution $\mathbf{K}_d^e$ and the convective contribution $\mathbf{K}_c^e$, with components

$$(k_d^e)_{ij} = \int_{\Omega^e} \left(k_x \partial_x \psi_{h,i}^e \partial_x \psi_{h,j}^e + k_y \partial_y \psi_{h,i}^e \partial_y \psi_{h,j}^e\right) dx dy$$

$$(k_c^e)_{ij} = \int_{\Omega^e} \psi_{h,i} u \partial_x \psi_{h,j}^e dx dy$$

Working out the explicit expression of $(k_d^e)_{ij}$ and $(k_c^e)_{ij}$ it is found that

$$\mathbf{K}_d^e = \begin{pmatrix} \frac{1}{3}(\eta k_x + \frac{1}{\eta}k_y) & \frac{1}{6}(-2\eta k_x + \frac{1}{\eta}k_y) & \frac{1}{6}(-\eta k_x - \frac{1}{\eta}k_y) & \frac{1}{6}(\eta k_x - \frac{2}{\eta}k_y) \\ & \frac{1}{3}(\eta k_x + \frac{1}{\eta}k_y) & \frac{1}{6}(\eta k_x - \frac{2}{\eta}k_y) & \frac{1}{6}(-\eta k_x - \frac{1}{\eta}k_y) \\ & & \frac{1}{3}(\eta k_x + \frac{1}{\eta}k_y) & \frac{1}{6}(-2\eta k_x + \frac{1}{\eta}k_y) \\ \text{Symm} & & & \frac{1}{3}(\eta k_x + \frac{1}{\eta}k_y) \end{pmatrix}$$

$$\mathbf{K}_c^e = \frac{uh_y}{12}\begin{pmatrix} -2 & 2 & 1 & -1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ -1 & 1 & 2 & -2 \end{pmatrix}$$

Let us first consider the case $u = 0$ and thus $\alpha_x = \alpha_y = 0$, $k_x = k_y = k$. Requiring condition (3.32) to hold yields

$$(k^e)_{12} = (k_d^e)_{12} = \frac{k}{6}\left(-2\eta + \frac{1}{\eta}\right) \leq 0 \implies \eta \geq \frac{\sqrt{2}}{2}$$

$$(k^e)_{14} = (k_d^e)_{14} = \frac{k}{6}\left(\eta - \frac{2}{\eta}\right) \leq 0 \implies \eta \leq \sqrt{2}$$

Hence, *even without convection* we need to have

$$\frac{\sqrt{2}}{2} \leq \eta \leq \sqrt{2} \tag{3.54}$$

to ensure that the DMP will hold.

Suppose now that $\eta = 1$. We obtain now the following conditions on the different off-diagonal terms of the matrix $\mathbf{K}^e$:

$$(k^e)_{12} = \frac{1}{6}\left(-k - \alpha_x uh + \frac{1}{2}\alpha_y uh\right) + \frac{1}{6}uh \leq 0 \implies \alpha_x - \frac{1}{2}\alpha_y \geq 1 - \frac{1}{2\gamma}$$

$$(k^e)_{13} = \frac{1}{6}\left(-2k - \frac{1}{2}\alpha_x uh - \frac{1}{2}\alpha_y uh\right) + \frac{1}{12}uh \leq 0 \implies \alpha_x + \alpha_y \geq 1 - \frac{2}{\gamma}$$

$$(k^e)_{14} = \frac{1}{6}\left(-k + \frac{1}{2}\alpha_x uh - \alpha_y uh\right) - \frac{1}{12}uh \leq 0 \implies \alpha_x - 2\alpha_y \leq 1 + \frac{1}{\gamma}$$

$$(k^e)_{21} = \frac{1}{6}\left(-k - \alpha_x uh + \frac{1}{2}\alpha_y uh\right) - \frac{1}{6}uh \leq 0 \implies \alpha_x - \frac{1}{2}\alpha_y \geq -1 - \frac{1}{2\gamma}$$

$$(k^e)_{23} = \frac{1}{6}\left(-k + \frac{1}{2}\alpha_x uh - \alpha_y uh\right) + \frac{1}{12}uh \leq 0 \implies \alpha_x - 2\alpha_y \leq -1 + \frac{1}{\gamma}$$

$$(k^e)_{24} = \frac{1}{6}\left(-2k - \frac{1}{2}\alpha_x uh - \frac{1}{2}\alpha_y uh\right) - \frac{1}{12}uh \leq 0 \implies \alpha_x + \alpha_y \geq -1 - \frac{2}{\gamma}$$

The only independent conditions are

$$\alpha_x - \frac{1}{2}\alpha_y \geq 1 - \frac{1}{2\gamma}, \quad \alpha_x + \alpha_y \geq 1 - \frac{2}{\gamma}, \quad \alpha_x - 2\alpha_y \leq -1 + \frac{1}{\gamma} \tag{3.55}$$

Consider the case $\alpha_y = 0$. The last condition in (3.55) leads to $\alpha_x \leq -1 + 1/\gamma$, which is incompatible with the first, $\alpha_x \geq 1 - 1/2\gamma$, for high values of the Péclet number. Therefore, *it is impossible to satisfy condition (3.32) adding only streamline diffusion.*

Suppose now that $\alpha_x = \alpha_y = \alpha$. From (3.55) it is found that

$$\alpha \geq 2 - \frac{1}{\gamma} \tag{3.56}$$

is needed if condition (3.32) is to be verified. Again we observe, as in the previous cases, that the form of the upwind funtion must be $C - 1/\gamma$ for a certain constant C.

3.3.4 Discussion

Several conclusions may be drawn from the application of Theorem 3.1 to the cases considered above. The first is that this result is limited, in the sense that it does not apply to many cases of interest. Maybe the most obvious one is the one-dimensional problem using quadratic elements, for which we know from the results of Chapter 1 that it is possible to obtain nodally exact solutions for piecewise linear source functions. The element stiffness matrix for the problem without convection and using elements of equal length h is

$$\mathbf{K}^e = \frac{2k}{h} \begin{pmatrix} \frac{7}{6} & -\frac{4}{3} & \frac{1}{6} \\ -\frac{4}{3} & \frac{8}{3} & -\frac{4}{3} \\ \frac{1}{6} & -\frac{4}{3} & \frac{7}{6} \end{pmatrix}$$

Neither this matrix nor the assembled one are of nonnegative type. Clearly, if the numerical solution is nodally exact condition (3.31) will hold, although (3.24) may not (ϕ_h is piecewise quadratic and may have local extrema between two nodes). Nevertheless, for linear elements we have obtained bounds for the upwind function (conditions (3.41) and(3.42)) that ratify which must be its behavior in terms of γ predicted in the previous chapters.

The next two cases studied (linear simplicial elements and a simple problem using bilinear elements) provide several conclusions. The first is that the stiffness matrix will not be of nonnegative type if the mesh is distorted, in particular, if the triangulation is not of strictly accute type when simplices are used and if the aspect ratio h_x/h_y or h_y/h_x is large using bilinear elements. Also, we have observed that the streamline diffusion is not enough to end up with a matrix of nonnegative type, but also an addition of crosswind diffusion is needed. This is perhaps the most salient result. When an isotropic artificial diffusion $k_a = \alpha h|\mathbf{u}|/2$ is introduced, α must be greater than $C - 1/\gamma_{||}$ for a certain constant C. Upper bounds for α have been obtained when the SUPG method is employed. All this agrees with the behavior of the upwind function dictated by the convergence analysis of Section 1.2.

All these facts will serve us to design the method proposed in the next section.

3.4 A discontinuity–capturing crosswind–dissipation

From the discussion of the previous section and the comments of Section 3.2, it is clear that the streamline diffusion introduced by the SUPG formulation is not enough to avoid overshooting and undershooting in the vicinity of sharp layers and an additional crosswind diffusion is needed. The methods discussed in Section 3.2 introduce this new dissipation but also modify the streamline diffusion. From the expression of the upwind function we use, this streamline diffusion satisfies all the general requirements we have found in Section 3.3 for some particular cases. The main idea of the method to be introduced here is to keep unaltered the diffusion in the direction of the streamlines and only to modify the crosswind diffusion.

The new crosswind dissipation (CD, from now onwards) must satisfy two conditions. To avoid excessive overdamping, it must be small in regions where convective effects are not very important, that is, where $|\mathbf{u}\cdot\nabla\phi_h|$ is small. For consistency, it must be proportional to the element residual defined in (3.9). Guided by the results of the previous section, the magnitude of the CD could by taken within each element as

$$k_c^e = \frac{1}{2}\alpha_c^e h^e \frac{|\mathbf{u}^e\cdot\nabla\phi_h|}{|\nabla\phi_h|} \tag{3.57}$$

when $|\nabla\phi_h| \neq 0$ and zero otherwise. Since when linear elements are used $\Delta\phi_h = 0$ within each element, instead of (3.57) we could take

$$k_c^e = \frac{1}{2}\alpha_c^e h^e \frac{|\mathcal{R}(\phi_h)|}{|\nabla\phi_h|} \tag{3.58}$$

Observe that the presence of f in $\mathcal{R}(\phi_h)$ does not modify the final stiffness matrix, but only the right-hand-side. Therefore, for simplicial elements this matrix will be still of nonnegative type, as proved in Proposition 3.2. Clearly, the use of (3.58) will yield a consistent numerical scheme (the exact solution ϕ will satisfy it) but the use of (3.57) will not.

The function α_c^e will be taken as

$$\alpha_c^e = \max\{0, C - 1/\gamma_{\|}^e\} \tag{3.59}$$

This ensures that $k_c^e = 0$ when $|\mathbf{u}^e\cdot\nabla\phi_h|$ is small. The question that remains is how to choose the constant C. From 2D numerical experiments we have found that $C \approx 0.7$ for linear and bilinear elements, and $C \approx 0.35$ for quadratic and biquadratic elements are effective. For the first case, it is observed that this corresponds to (3.49) with the constant in this expression equal to 2.

Having defined the magnitude of the CD by (3.57) or (3.58) and the function α_c^e by (3.59), the description of the method is now complete. The final problem will be written next.

In order to introduce the CD, a tensor

$$\mathbf{k}_c^e = k_c^e\mathbf{k}_0 := k_c^e\frac{1}{|\mathbf{u}|^2}(\mathbf{u}_{n,1}\otimes\mathbf{u}_{n,1} + \mathbf{u}_{n,2}\otimes\mathbf{u}_{n,2}) = k_c^e\left(\mathbf{I} - \frac{1}{|\mathbf{u}|^2}\mathbf{u}\otimes\mathbf{u}\right) \tag{3.60}$$

must be added within each element to the real diffusion tensor $\mathbf{k}$. In (3.60), $\mathbf{u}_{n,1}$ and $\mathbf{u}_{n,2}$ denote the vectors normal to $\mathbf{u}$ and with the same norm for $N_{sd} = 3$. For $N_{sd} = 2$ there is only one vector normal to $\mathbf{u} = (u_1, u_2)$, that may be taken as $\mathbf{u}_{n,1} = (-u_2, u_1)$.

Using the notation of Chapter 1, the problem to be solved is the following: Find a function $\phi_h \in \Phi_h$ such that

$$\begin{aligned}
&\int_\Omega (\psi_h \mathbf{u}\cdot\nabla\phi_h + \nabla\psi_h\cdot\mathbf{k}\cdot\nabla\phi_h)\,d\Omega \\
&+\sum_{e=1}^{N_{el}}\int_{\Omega^e}(\tau^e\mathbf{u}^e\cdot\nabla\psi_h)\left[\mathbf{u}\cdot\nabla\phi_h - \nabla\cdot(\mathbf{k}\cdot\nabla\phi_h)\right]d\Omega \\
&+\sum_{e=1}^{N_{el}}\int_{\Omega^e}\frac{1}{2}\alpha_c^e h^e\frac{|\mathcal{R}(\phi_h)|}{|\nabla\phi_h|}(\nabla\psi_h\cdot\mathbf{k}_0\cdot\nabla\phi_h)\,d\Omega \\
&=\int_\Omega \psi_h f\,d\Omega + \int_{\Gamma_N}\psi_h r\,d\Gamma + \sum_{e=1}^{N_{el}}\int_{\Omega^e}(\tau^e\mathbf{u}^e\cdot\nabla\psi_h)f\,d\Omega
\end{aligned} \tag{3.61}$$

for all $\psi_h \in \Psi_h$. Observe that the only difference with the original SUPG method is the third term in the left-hand-side.

From the expressions of the intrinsic time τ^e in terms of the upwind function given in Chapter 1 and from (3.59) it follows that the CD will always be smaller than the streamline diffusion introduced by the SUPG formulation. The total diffusion ellipsoid in 2D is schematically represented in Figure 3.4.

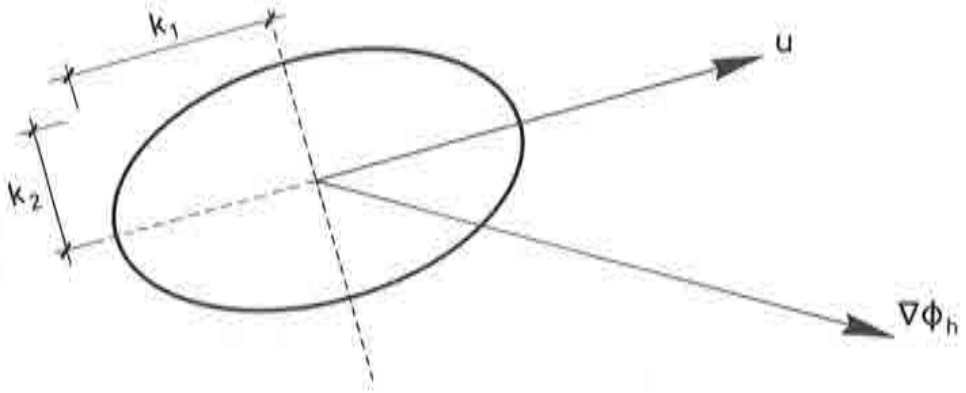

Figure 3.4 Total diffusion ellipsoid. The values of k_1 and k_2 are $k_1 = k + \alpha h|\mathbf{u}|/2$, $k_2 = k + \alpha_c h|\mathcal{R}(\phi_h)|/2|\nabla\phi_h|$ when the physical diffusion is isotropic and of magnitude k.

Let us close this section mentioning some results obtained by Johnson *et al.* in Reference [JSW] and slightly improved by Niijima [Nj]. They analyzed a 2D model problem using linear elements and the SUPG formulation and introducing a *constant* crosswind diffusion $k_c = \max(k, h^{3/2})$, k being the physical diffusion. For this choice of k_c, the global L^2-estimates of the SUPG method are not deteriorated, since they are $O(h^{3/2})$. Suppose that $k \leq h^{3/2}$. Under certain regularity assumptions on the data, the main results of the quoted references are

$$\begin{array}{lll}
 & \text{Ref. [JSW]} & \text{Ref. [Nj]} \\
|(\phi-\phi_h)(x_0,y_0)| = & O\left(h^{5/4}\log^{3/2}(1/h)\right) & O\left(h^{11/8}\log(1/h)\right)
\end{array} \tag{3.62}$$

$$\|\phi_h\|_{L^\infty(\Omega)} = \quad O\left(h^{-1/4}\log^{3/2}(1/h)\right) \qquad O\left(h^{-1/8}\log(1/h)\right) \quad (3.63)$$

$$\|\phi - \phi_h\|_{L^1(\Omega)} = \quad O\left(h^{1/2}\log^{5/2}(1/h)\right) \qquad O\left(h^{5/8}\log^2(1/h)\right) \quad (3.64)$$

where (x_0, y_0) is a point in Ω. The important result is the pointwise error estimate (3.61), the other two are consequences of it. Observe that (3.62) is a weak version of (3.25) and (3.61) a weak version of (3.27).

All these theoretical results seem to confirm that the introduction of a crosswind diffusion might improve the numerical solution. This idea will be further strengthened by the numerical experiments to be presented next.

3.5 Numerical examples

In all the numerical examples to be presented below, we have solved the nonlinear system of equations arising using the methods discussed in this chapter via a fictious transient. The forward Euler scheme in time, with a lumped mass matrix, has been used to step in time until the steady state has been reached. The stability limits derived in Chapter 2 have been employed with success. Unless otherwise specified, the safety factor in (2.97) has always been taken equal to unity. When talking about the convergence rate of a certain method, we shall refer to the convergence towards the steady state solution. The *residual* for the nth time step will be

$$\text{Residual} = \frac{|\Phi^{n+1} - \Phi^n|}{|\Phi^{n+1}|}$$

Eight different numerical methods will be considered, with the following acronyms:

1.- SUPG: Original SUPG formulation
2.- DC type 1: Discontinuity capturing DC1 of Hughes *et al.* [HMM], defined by (3.6) and the first equation in (3.7).
3.- DC type 2: Discontinuity capturing DC2 of Hughes *et al.* [HMM], defined by (3.6) and the second equation in (3.7).
4.- CAU: Method of Galeão & Dutra do Carmo [GD], defined by (3.16) and (3.17).
5.- Modified DC: Method defined by (3.20) and the first equation in (3.21).
6.- Modified CAU: Method defined by (3.20) and the second equation in (3.21).
7.- CD type 1: Introduction of a crosswind dissipation given by (3.57) and (3.60).
8.- CD type 2: Introduction of a crosswind dissipation given by (3.58) and (3.60). The final problem in this case is (3.61).

The method we currently favor is number 8. Although not consistent, method number 7 has also interesting features. It is slightly more overdiffusive than method number 8, but sometimes the convergence towards the steady state is faster.

In all the cases we have used a value 1 for the constant in Eqn.(3.22) and for the constant in Eqn.(3.59) values 0.7 for linear elements and 0.35 for quadratic elements.

All the calculations have been carried out on a CONVEX-320 computer using double arithmetic precision.

Example 3.1 In this first example we solve again the same problem as in Example 1.4 of Chapter 1, where we first noticed the inability of the SUPG method to preclude overshooting and undershooting in the vicinity of sharp layers. The computational domain is the unit square $[0,1] \times [0,1]$.

Figure 3.5 shows the results obtained using a mesh of 20×20 bilinear elements. Since there are no source terms for this problem and the Laplacian of the shape functions within each element is zero, the CAU method coincides with DC of type 2, and so do the modified DC and the modified CAU methods, and the CD of types 1 and 2. Results using the DC method of type 1 are not shown (see Reference [HMM]). They are the most overdiffusive of all. The best results are those obtained using the modified DC and the CD methods, the latter being slightly less diffusive.

The convergence history of the four methods for which results are shown in Figure 3.5 is plotted in Figure 3.6. It is observed that the convergence rate of the CD method is very similar to that of the original SUPG algorithm, whereas for the DC type 2 and the modified DC it is substancially deteriorated. For these two methods the residual does not decrease from a certain time step onwards. If fact, for the modified DC method a safety factor $f_t = 0.5$ in Eqn.(2.97) is needed in order to obtain a stable time stepping algorithm.

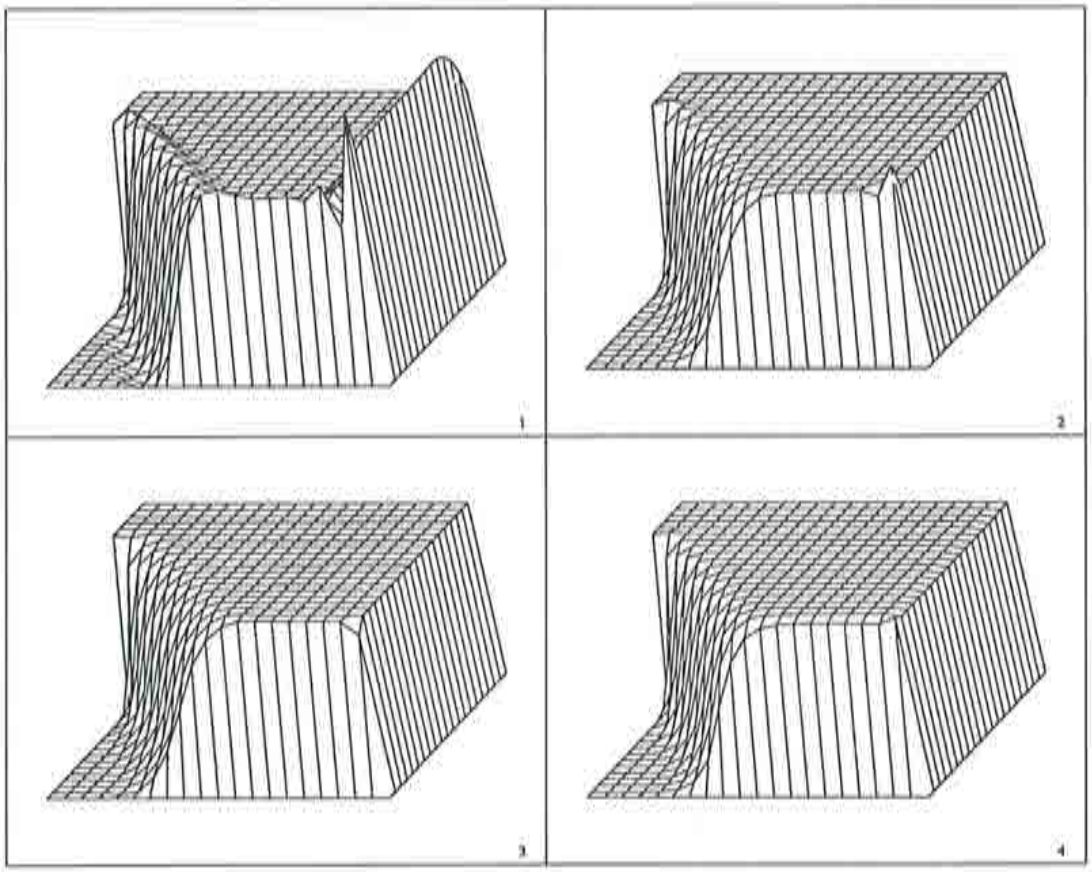

Figure 3.5 Numerical solution of Example 3.1 using bilinear elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

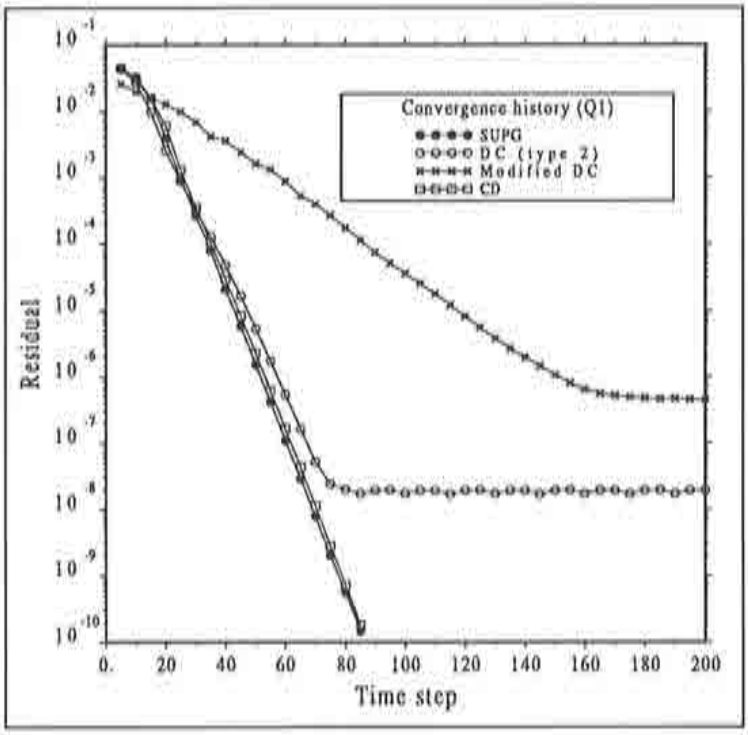

Figure 3.6 Convergence history for Example 3.1 using bilinear elements.

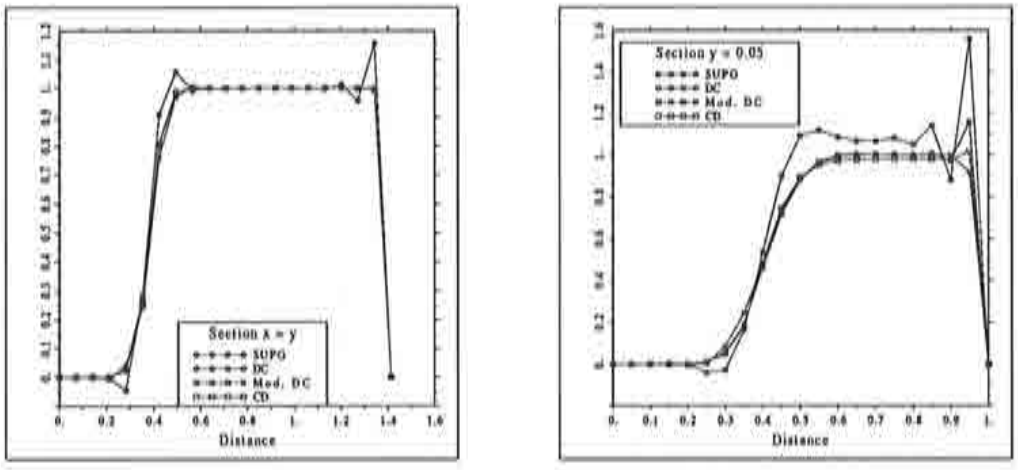

Figure 3.7 Sections $y = 0.05$ and $y = x$ for Example 3.1 using bilinear elements.

In Figure 3.7 we have plotted the sections $y = 0.05$ and $y = x$ (diagonal section). The oscillations found using the SUPG formulation as well as the good resolution of the layers obtained using the other three methods is apparent.

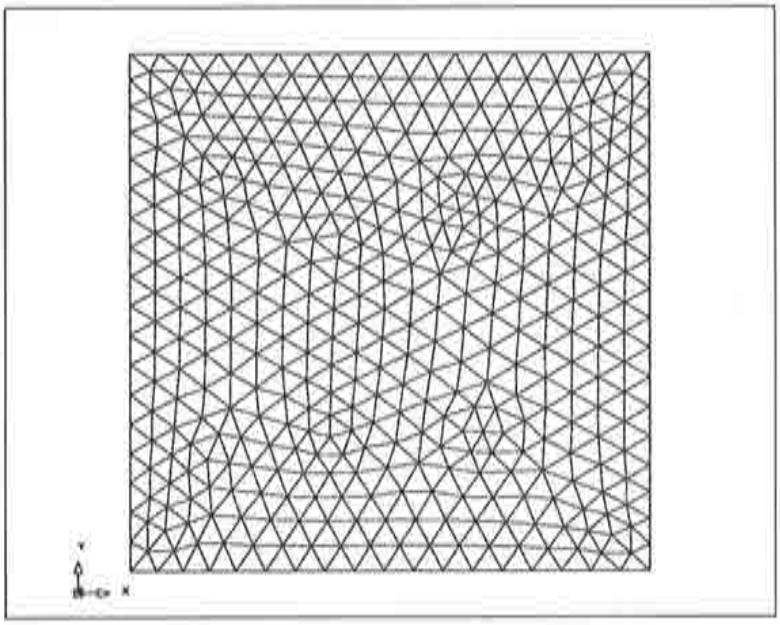

Figure 3.8 Unstructured mesh using triangles for Example 3.1. Six-noded triangles have been split into four subtriangles in this figure. Therefore, all the nodes are represented here.

Next we consider the same problem but using triangular finite elements, both linear and quadratic. For the former case, the mesh is the same as for the latter, but splitting each quadratic triangle into four linear subtriangles. The resulting unstructured finite element mesh is shown in Figure 3.8. It is composed of 437 nodal points and 800 linear triangles (or 200 quadratic triangles).

Let us consider first the case of linear triangles. Figure 3.9 shows the results obtained using the same methods as for bilinear finite elements. It is also observed here that all the shock-capturing techniques succeed in removing the localized oscillations found using the SUPG method. It is also important to note the good accuracy obtained considering the unstructured mesh we use. Only near the corner $(x, y) = (1, 1)$ results seem to be slightly overdamped.

From the proof Proposition 3.2 it follows that for this case the DMP should be satisfied provided that $\alpha_c^e \geq (-2/C_2\sigma^e)/(N_{sd}+1) - 1/\gamma_{||}$, where all the terms appearing in this expression have been defined earlier. From Figure 3.9.(4) it is observed that very small overshoots are still present introducing the crosswind dissipation given by (3.57) (or (3.58), in this case it is the same) and with α_c^e calculated as indicated in (3.59), with $C = 0.7$. If all the triangles were perfectly equilateral, we would have that $\sigma^e = \cos(-\pi/6) = -0.5$ and $C_2 = 1$ (see Eqns.(3.45) and (3.46)). Therefore, we should have $\alpha_c^e \geq 4/3 - 1/\gamma_{||}$ in order to satisfy the DMP, that is, $C = 4/3$ should be taken in (3.59). Although from Figure 3.8 it is observed that not all the elements are equilateral, we have checked that if $C = 4/3$ is taken the DMP in fact holds true. However, results are a little overdiffusive using this value of the constant (not shown).

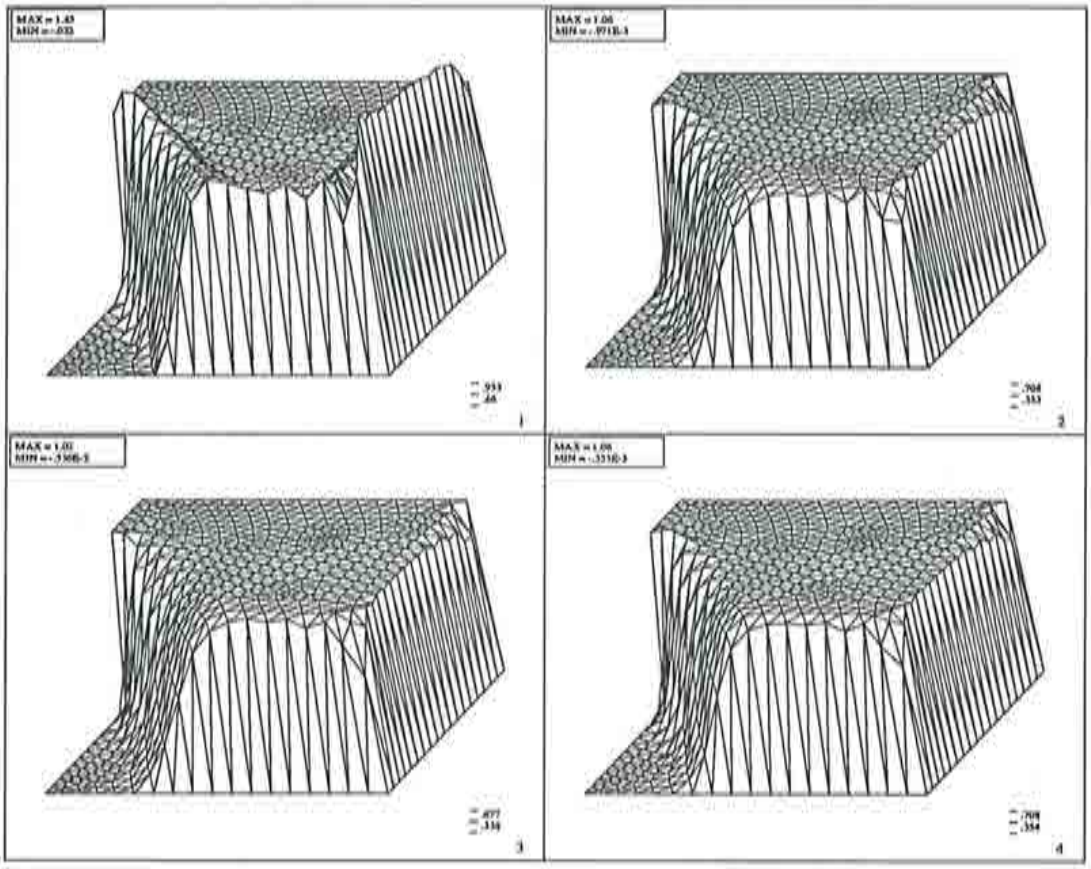

Figure 3.9 Numerical solution of Example 3.1 using linear triangular elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

The main difference between the three shock-capturing techniques considered for this problem is not the accuracy of the numerical results but their convergence properties. The convergence history is plotted in Figure 3.10. The conclusion to be drawn from this figure is the same as for bilinear elements: whereas the evolution towards the steady state introducing CD is very similar to that of the SUPG method, the convergence rate of the DC type 2 and the modified DC methods is much smaller. Once again, a safety factor $f_t = 0.5$ has been needed for the modified DC method.

Let us consider now the case of quadratic triangles. For the small value of the diffusion coefficient used for this problem ($k = 10^{-6}$), the diffusive term within each element is negligible (although the Laplacian of the shape functions is not zero) and therefore the CAU method yields the same results as the DC method, and so do the CD type 1 and the CD type 2 methods. So we shall consider the same methods as for the previous two cases using bilinear and linear elements. Numerical results are shown in Figure 3.11.

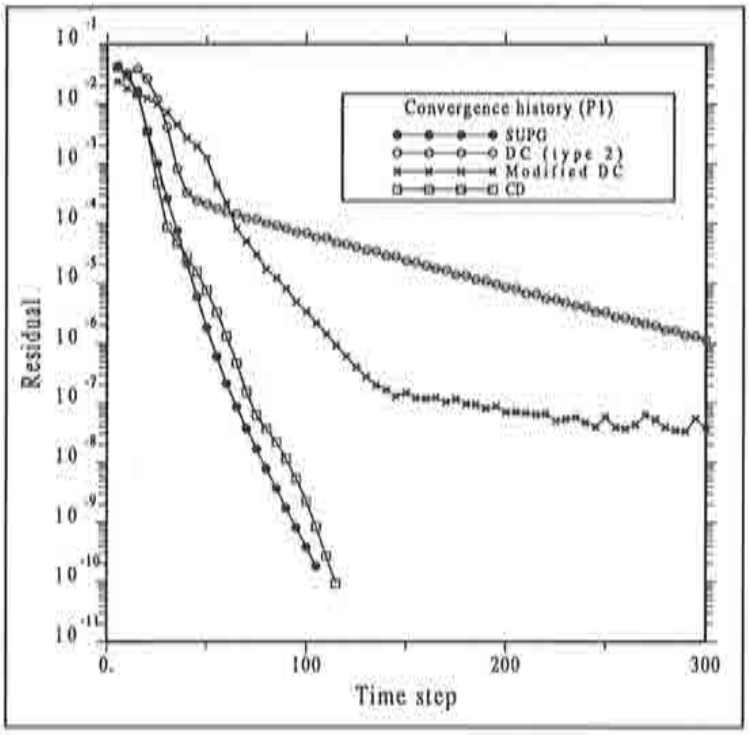

Figure 3.10 Convergence history for Example 3.1 using linear triangular elements.

A single upwind function $\alpha^e = 0.5 \min\{\gamma^e/3, 1\}$ has been used to calculate the intrinsic time of the SUPG contributions. As shown in Figures 1.17 and 1.18, better acuracy is obtained using different upwind functions for the corner nodes of the triangles and for the midside nodes. The method to assign the upwind functions has been discussed in Chapter 1. However, it is much cheaper and simpler to use a unique upwind function.

The performance of the different shock-capturing techniques is again very similar. Results are good in general, although a little smeared about the boundary layers. The convergence history plotted in Figure 3.12 shows the same trends as for bilinear and linear elements. It is interesting to note that is this case the method that has a higher convergence rate is the introduction of a CD, which reaches the steady state even faster than the SUPG formulation.

Let us discuss the numerical cost of the calculations referring to the different shock-capturing techniques and the different elements employed. Consider first the bilinear element, for which a 2×2 Gauss-Legendre integration rule has been employed. The total CPU of the computation using the SUPG, the DC and the CD methods, as well as the CPU time per iteration are the following:

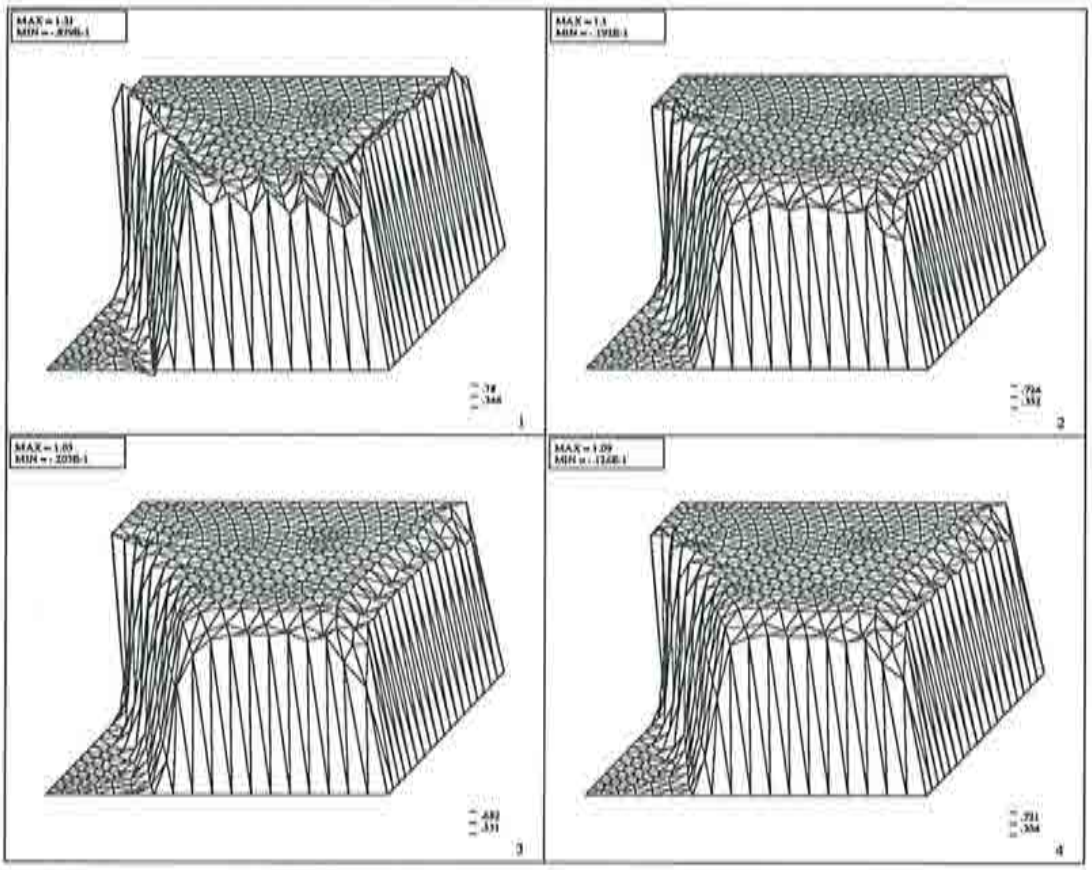

Figure 3.11 Numerical solution of Example 3.1 using quadratic triangular elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

Method	CPU (seconds)	CPU per iter. ($\times 10^{-3}$)
SUPG	6.74	79.29
DC (type 2)	18.99	94.95
CD	7.60	89.41

It is observed that the CD method needs an amount of computer time per iteration similar to the DC method (even smaller in this case). Athough this depends on the programming, it is a clear indication that the introduction of an anisotropic diffusion is not expensive from the computational standpoint. The increase of computer time with respect to the original SUPG formulation is not very important.

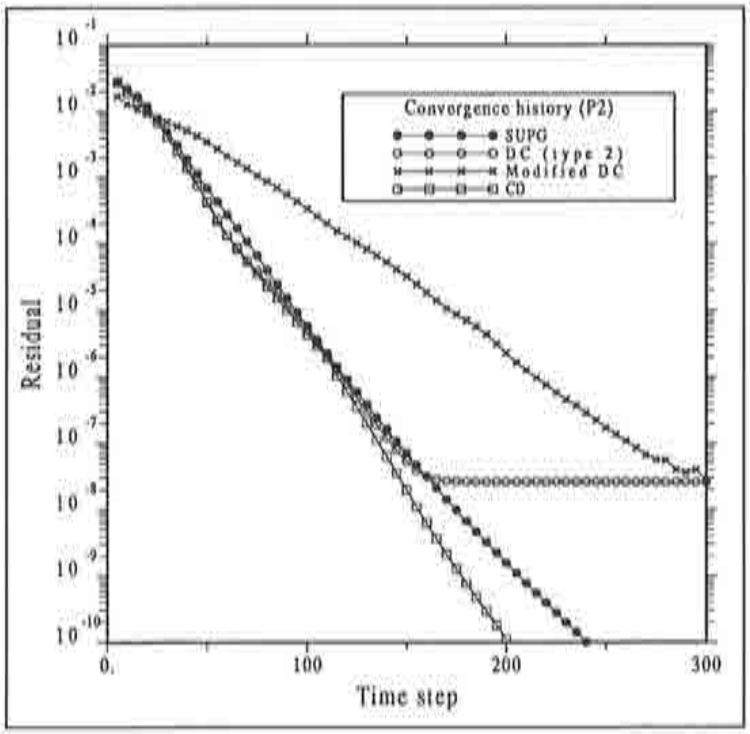

Figure 3.12 Convergence history for Example 3.1 using quadratic triangular elements.

For the linear triangle using a three-point integration rule the results are the following:

Method	CPU (seconds)	CPU per iter. ($\times 10^{-3}$)
SUPG	13.15	125.40
DC (type 2)	44.29	147.63
CD	17.18	149.39

and for the quadratic triangle using a four-point integration rule:

Method	CPU (seconds)	CPU per iter. ($\times 10^{-3}$)
SUPG	9.04	37.67
DC (type 2)	12.70	42.33
CD	8.27	41.35

The conclusions for these two cases are the same as before in what concerns the behavior of the shock-capturing techniques. Concerning the element employed in the calculation, the smallest CPU time per iteration is needed using the quadratic triangle, whereas the largest is needed using the linear triangle. As explained in Chapter 2, this is due to the *total* number of integration points of the finite element mesh, since using an explicit scheme to advance in time the cost of updating the unknowns depends on

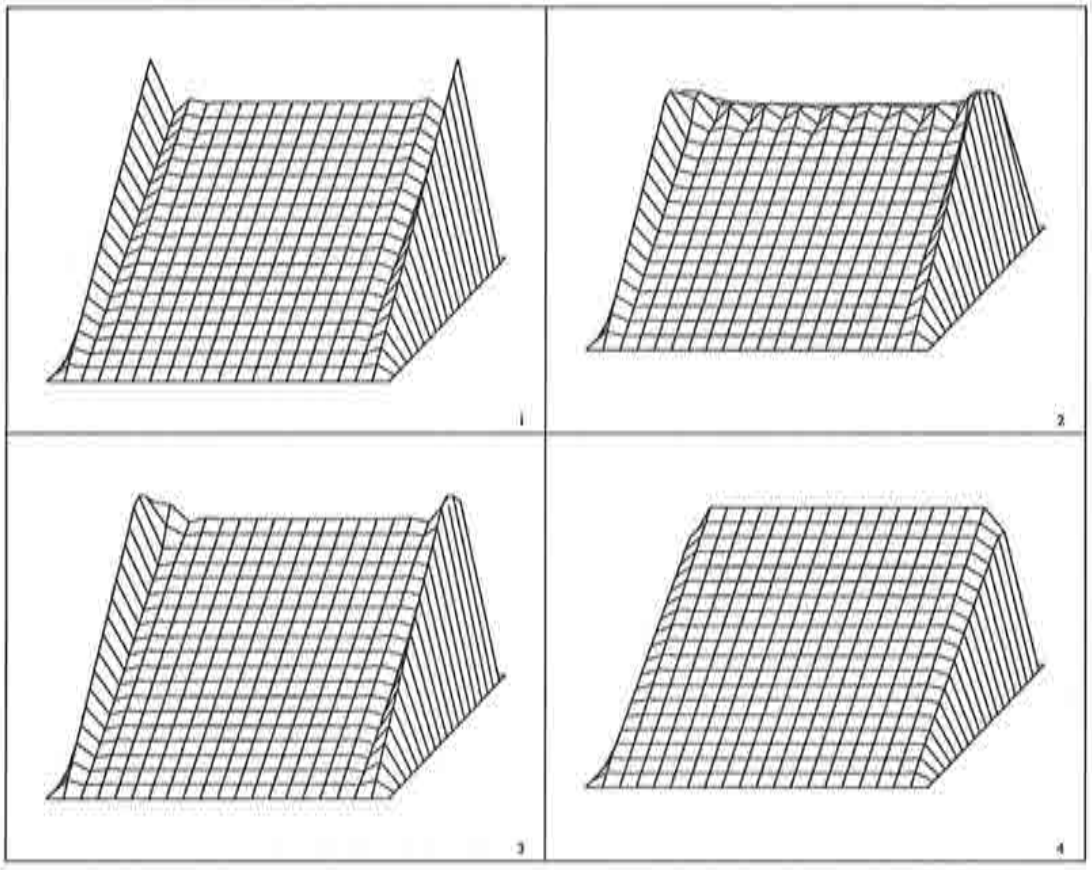

Figure 3.13 Numerical solution of Example 3.2. (1): SUPG formulation; (2): DC of type 1; (3): DC of type 2; (4): CAU.

the number of nodal points, not on the type of element. It is also observed that the total CPU is similar (although depending on the method employed) using the bilinear element and the quadratic one, even though many more time steps are needed to reach the steady state for the latter. All this was already observed in Chapter 2.

In summary, we may conclude that the CD method has a specific computational cost similar to the DC, although converges much better. The behavior of the elements using these methods is the same as for the SUPG formulation: quadratic elements may be cheaper than linear elements due to the lesser number of total integration points for a given number of nodal points of the finite element mesh.

Example 3.2 The computational domain for this example will be again the unit square, discretized now with a uniform mesh of 20×20 bilinear elements. Homogeneous boundary conditions of Dirichlet type are prescribed on the whole boundary. The data of the problem are:

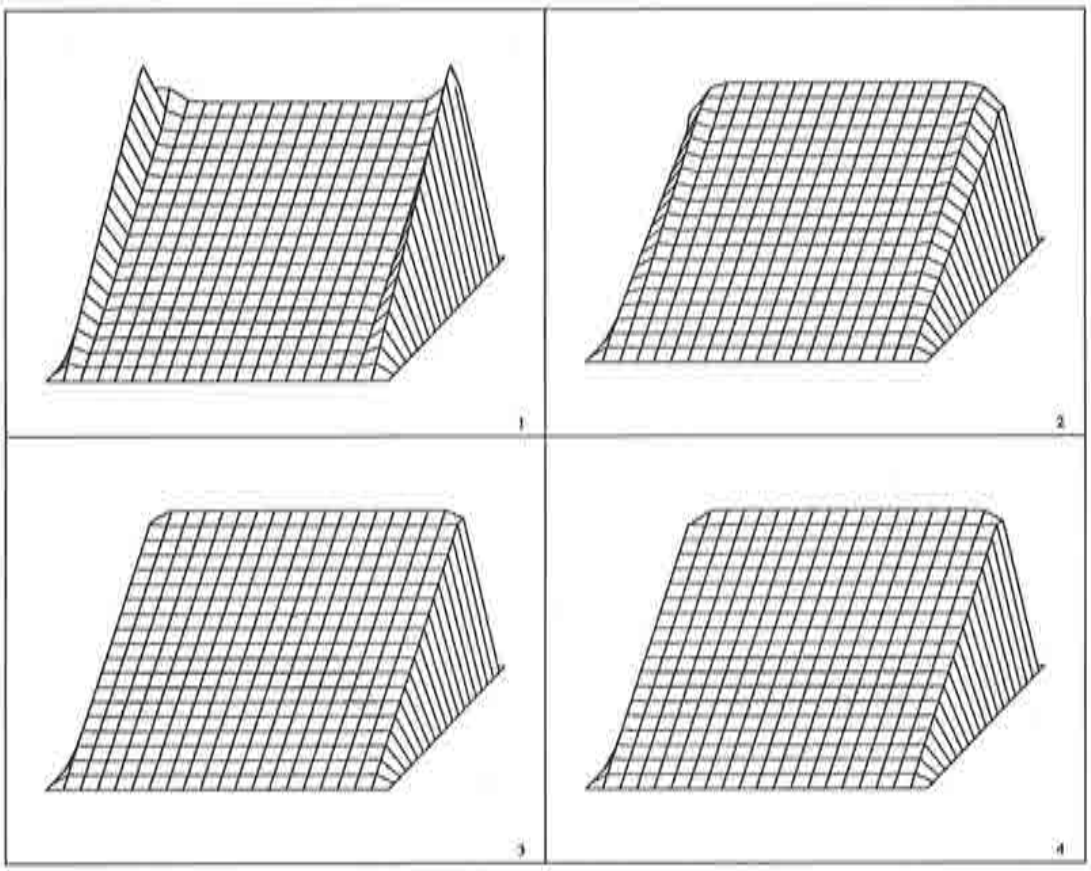

Figure 3.14 Numerical solution of Example 3.2. (1): Modified DC; (2): Modified CAU; (3): CD of type 1; (4): CD of type 2.

$$\Omega =]0,1[\times]0,1[$$
$$\Gamma_D = \partial\Omega, \quad \Gamma_N = \emptyset$$
$$\mathbf{u}(x,y) = (0,1)$$
$$k_{ij}(x,y) = 10^{-8}\delta_{ij}$$
$$f(x,y) = 1$$
$$g(x,y) = 0$$

Results obtained using the eight shock-capturing techniques indicated at the begining of this section are shown in Figures 3.13 and 3.14. It is observed that the SUPG formulation yields oscillations in the direction normal to the lateral layers, but a good resolution of the layer normal to the velocity field is obtained. This confirms our argument that in the streamline direction it is not necessary to increase the numerical diffusion, no matter how small the physical diffusion is.

The first shock-capturing technique we have considered is the DC type 1 method. We have failed to obtain a converged solution in this case. It could be argued that this is due to the instability of the forward Euler scheme in time. But this is not the case: after a large number of time steps the residual does not increase, which would

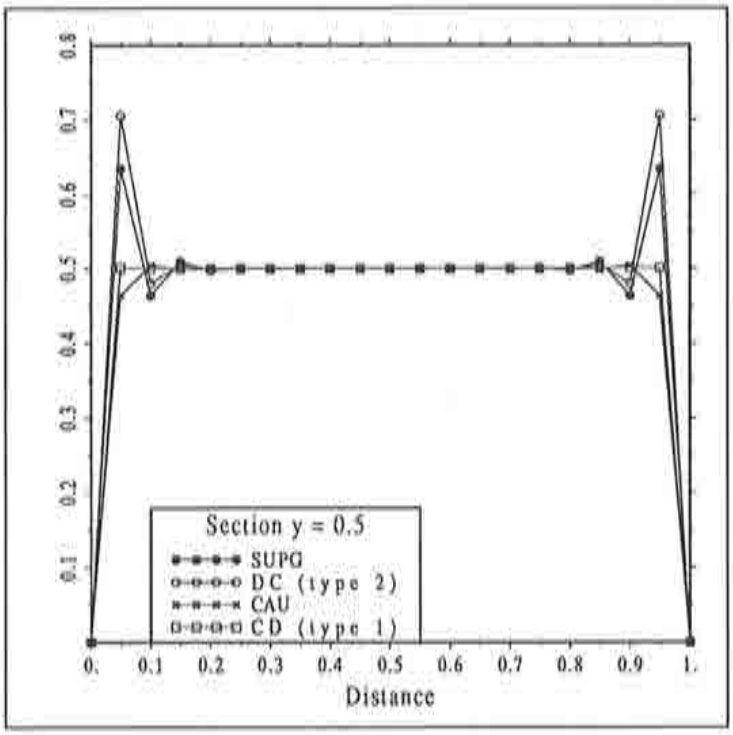

Figure 3.15 Section $y = 0.5$ for Example 3.2.

mean an unstable behavior, but rather keeps constant at an unacceptable value of order 10^{-2}. The solution plotted in Figure 3.13.(2) shows the oscillations encountered along the layer normal to the velocity. The problem relies on the strong nonlinearity of the numerical algorithm, which is not possible to trackle using the time stepping method. It is important to note that we are faced to a numerical problem because of the modification of the streamline diffusion, which we know is enough from the results of the SUPG formulation.

Let us look now at the results obtained using the DC type 2 method. Now convergence problems are not encountered, although the oscillations along the lateral layers found using the SUPG approach have not been avoided, but rather increased. The explanation to this is that using this method we may introduce a negative diffusion, as explained in Section 3.2. That the CAU approach circumvents this problem is clearly observed from Figure 3.13.(4). Good results are obtained now, although the solution is too smeared along the lateral layers.

The modified DC and the modified CAU methods show the same problems as the original DC type 2 and CAU methods, as seen from Figures 3.14.(1) and 3.14.(2). A dramatic improvement is observed when the CD type 1 and type 2 approaches are employed (see Figures 3.14.(3) and 3.14.(4)). We recall the the first of these methods is not consistent, in the sense that the exact solution does not satisfy the numerical scheme. Nevertheless, numerical results are very good, as they are using the consistent CD type 2 method. Only the solution at the nodes next to the corners $(0,1)$ and $(1,1)$ is slightly smeared.

In Figure 3.15 we have plotted the section $y = 0.5$, where the quality of the

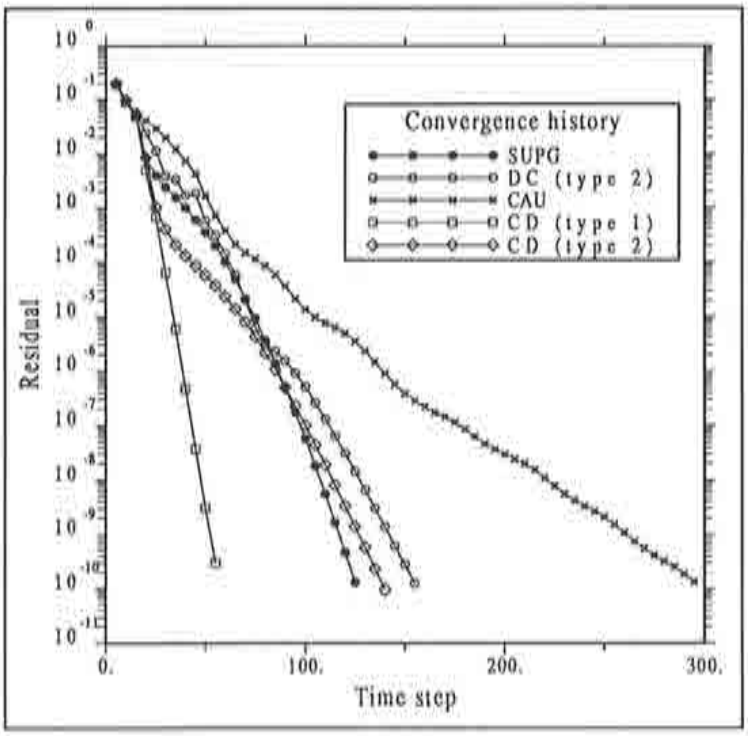

Figure 3.16 Convergence history for Example 3.2.

solution obtained using different techniques is better appreciated.

The convergence history using different methods is plotted in Figure 3.16. Concerning the SUPG, DC type 2, CAU and CD type 2 methods the conclusions are the same as in the previous example: the CD type 2 method has a convergence rate very similar to the SUPG technique and the other two methods have worse convergence properties. However, now we also observe that the CD type 1 method is the one that converges best, much better than the SUPG method. This fact though must not be taken as a general assessment, since for some other numerical experiments we have observed that the convergence of the CD type 2 method is better than that of the CD type 1 approach, and always very similar to the original SUPG technique.

In conclusion, from this example we see that the introduction of the crosswind diffusion as proposed in this chapter not only improves the numerical behavior of the iterative scheme compared to other shock-capturing techniques, but also that higher accuracy may be obtained.

Example 3.3 The problem we solve now is very similar to the previous one. The only difference will be the expression of the source term, that now we take as

$$f(x,y) = \begin{cases} 1 & \text{if } 0 \leq y \leq 0.5 \\ -1 & \text{if } 0.5 < y \leq 1 \end{cases}$$

Again, only results obtained using a mesh of 20×20 bilinear finite elements will be discussed. These results for the eight shock-capturing techniques are shown in Figures

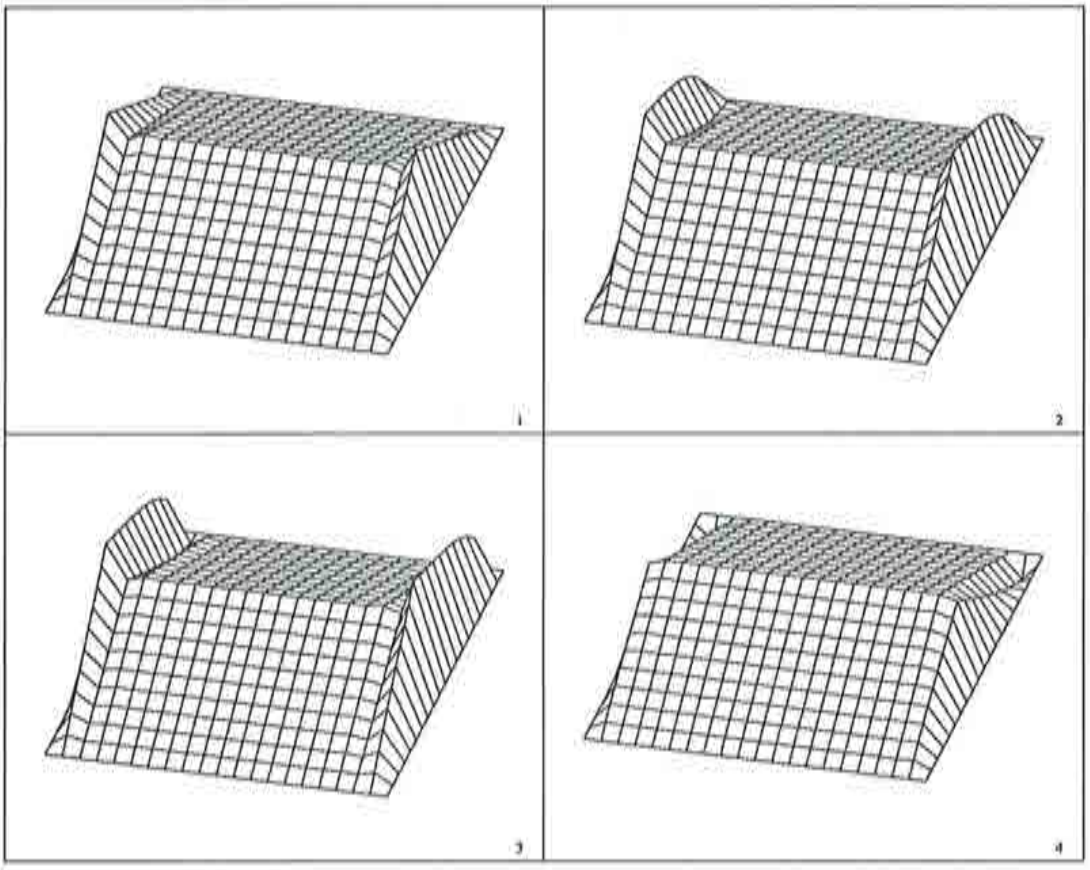

Figure 3.17 Numerical solution of Example 3.3. (1): SUPG formulation; (2): DC of type 1; (3): DC of type 2; (4): CAU.

3.17 and 3.18.

Concerning the SUPG and the DC methods, the same observations as for the previous example can be made. Neither of these formulations precludes the appearence of oscillations along the lateral layers (see Figures 3.17.(1), 3.17.(2), 3.17.(3) and 3.18.(1)). It is noted that for the DC methods (type 1, type 2 and modified DC) the magnitude of the overshoots is increased with respect to the original SUPG formulation in the zone $0.5 < y \leq 1$. This clearly indicates that we are introducing a negative diffusion in the numerical scheme. The convergence history is again very poor using the DC methods (not shown).

Results using the CAU, modified CAU, CD type 1 and CD type 2 methods are much better (see Figures 3.17.(4), 3.18.(2), 3.18.(3) and 3.18.(4), respectively). The overshoots along the lateral layers are in fact removed. However, it is observed that in the zone $0.5 < y \leq 1$ the numerical solution is a little overdamped in all the cases, and it is even negative near the corners $(0, 1)$ and $(1, 1)$. We have found this problem a severe test for the shock-capturing techniques, since the numerical solution is extremely sensitive to the numerical dissipation introduced by all these methods, especially for $0.5 \leq y \leq 1$. Nevertheless, the best answers are again obtained using the CD type 1 and type 2 methods. For the former, the solution is a little bit more smeared than for

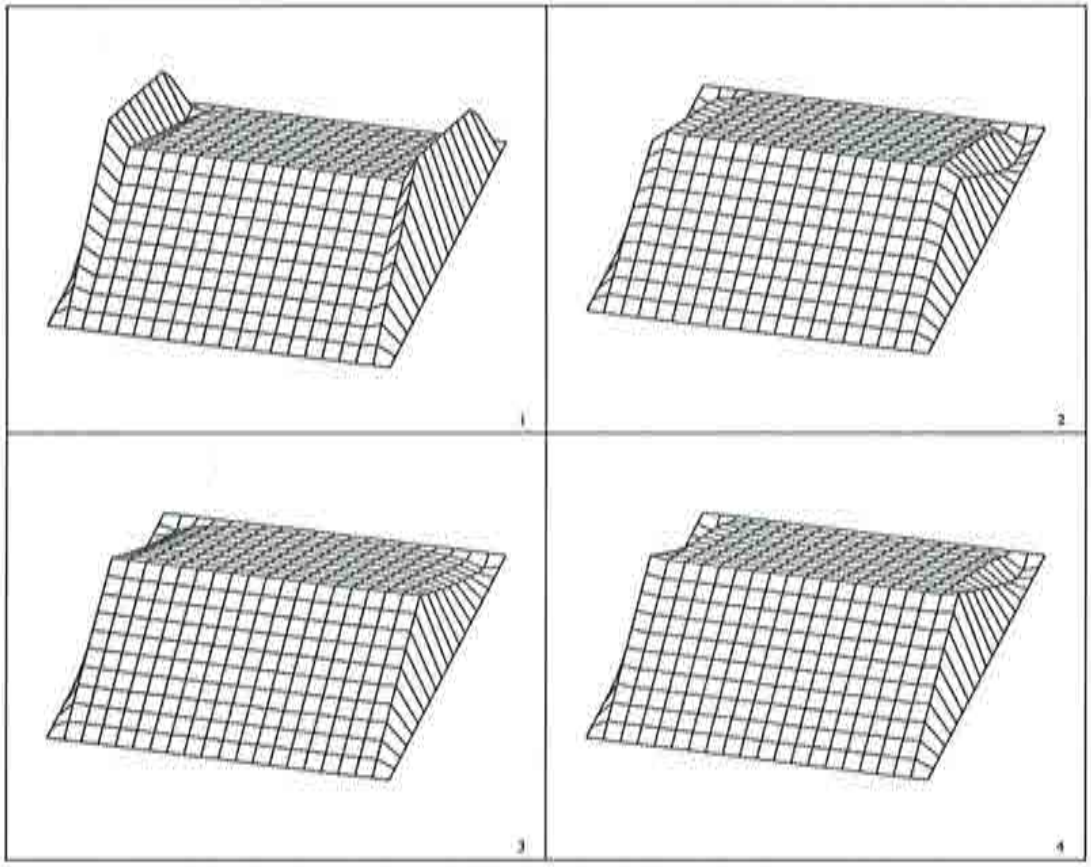

Figure 3.18 Numerical solution of Example 3.3. (1): Modified DC; (2): Modified CAU; (3): CD of type 1; (4): CD of type 2.

the latter, showing the effect of adding a dissipation not proportional to the residual of the equation within each element.

In Figure 3.19 we have plotted the sections $y = 0.75$ and $y = 0.95$, where the quality of the numerical solutions is better observed. In the first case the exact solution has a constant value $\phi = 0.25$ and in the second a constant value $\phi = 0.05$. The solution using the CD type 1 method is the most smooth one, although too smeared. It is also observed that the CD type 2 method is less overdiffusive than the CAU method. This, however, depends on the values of the algorithmic constants of all the numerical methods. Although the numerical results are not very sensitive to these values, a certain influence does certainly exist.

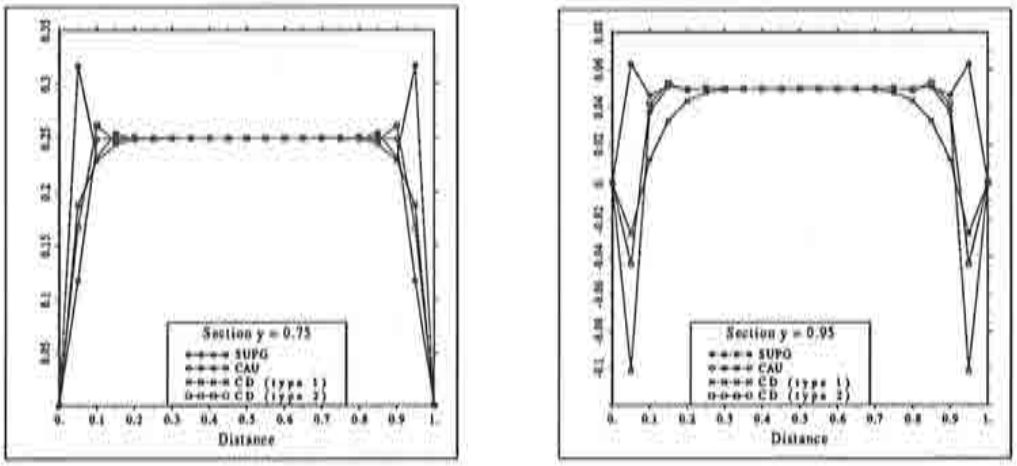

Figure 3.19 Sections $y = 0.75$ and $y = 0.95$ for Example 3.3.

Example 3.4 To study further how overdiffusive are the numerical formulations we use, now we study again the problem presented in Example 1.3, with slightly different data:

$$\Omega =]0,1[\times]0,1[- \left]\frac{1}{2},1\right] \times \{\frac{1}{2}\}$$
$$\Gamma_D = \partial\Omega, \quad \Gamma_N = \emptyset$$
$$\mathbf{u}(x,y) = \left(-y - \frac{1}{2}, x + \frac{1}{2}\right)$$
$$k_{ij}(x,y) = 10^{-8}\delta_{ij}$$
$$f(x,y) = 0$$
$$g(x,y) = \begin{cases} \sin\left[\frac{\pi}{2}(8x-5)\right] + 1 & \text{if } \frac{1}{2} \le x \le 1 \text{ and } y = \frac{1}{2} \\ 0 & \text{else} \end{cases}$$

We shall solve this problem using a uniform mesh of 31×31 nodal points. First we consider the case in which 30×30 bilinear elements are used to discretize the domain Ω. The numerical solution using the SUPG, the DC type 2, the modified DC and the CD (type 1 and type 2 coincide) methods is shown in Figure 3.20. For the small diffusion considered, the amplitude of the sine profile should be constant in the whole domain. As it could be expected, it is observed that a certain loss of this amplitude results from the use of any of the shock-capturing techniques. The less overdiffusive method is the introduction of a crosswind dissipation (CD). As can be seen from the section $y = 0.5$ shown in Figure 3.21, the highest amplitude is obtained using the CD method, although the improvement with respect to the other techniques is not very pronounced for this problem.

The same results obtained using 15×15 biquadratic elements are shown in Figures 3.22 and 3.23 (elevation plots and section $y = 0.5$, respectively). In the figures, each biquadratic element has been split into four bilinear elements to compare with the previous solution. It is interesting to note that all the shock-capturing techniques yield much better results than using the bilinear element. So, we may conclude that in

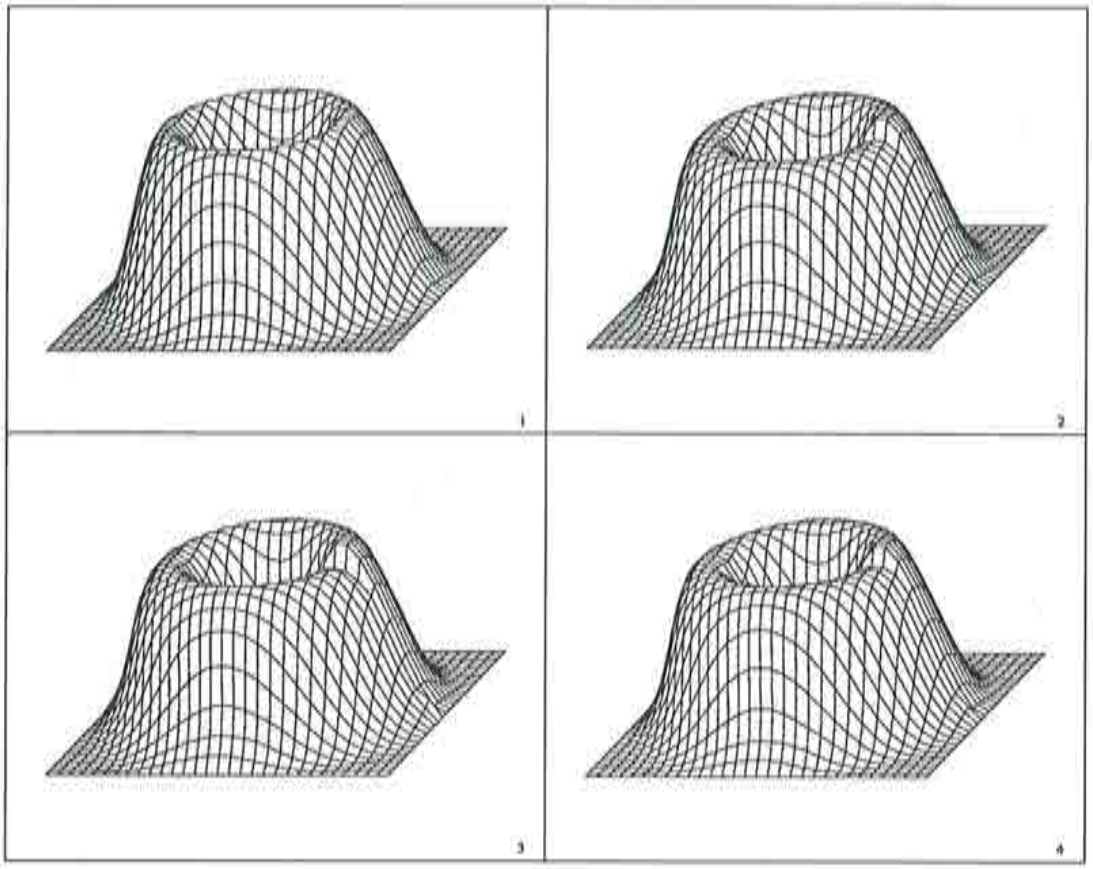

Figure 3.20 Numerical solution of Example 3.4 using bilinear elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

regions where the solution is smooth, quadratic elements yield much less overdiffusive numerical answers than linear elements. This behavior is observed for all the shock-capturing techniques we are considering. The reason is that using quadratic elements we have a much better approximation to the residual of the continuous equation within each element. It should be noted that the Laplacian of the shape functions is not zero, but since the diffusion is very small, the diffusion term of this residual within each element is negligible. Therefore, both the CD type 1 and the CD type 2 methods yield almost the same numerical results, as well as the DC type 2 and the CAU methods.

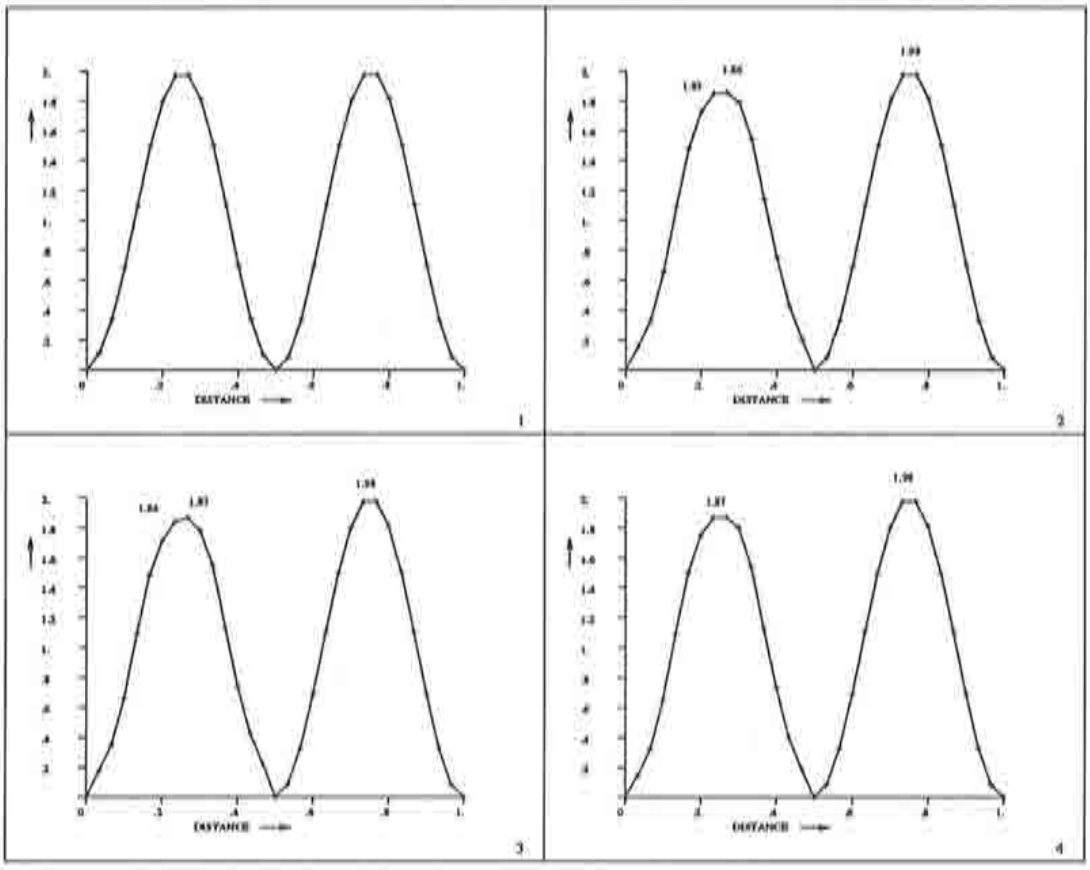

Figure 3.21 Section $y = 0.5$ for Example 3.4 using bilinear elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

3.6 Summary and conclusions

In this chapter we have been concerned with the problem of removing the localized oscillations that still remain using the SUPG formulation, that is, to devise a shock-capturing technique. The attention has been focussed first on the study of two of these methods, reinterpreted as the introduction of a nonlinear dissipation. It has been shown why the discontinuity capturing technique of Hughes *et al.* may fail in the presence of source terms. This is due to the fact that we actually may be introducing negative numerical dissipation. The introduction of a true positive dissipation proportional to the residual circumvents this problem. The crucial question is why should this new dissipation be isotropic if the streamline diffusion introduced by the original SUPG formulation seems to be enough along the streamlines. This last point is confirmed not only by numerical experiments, but also by the study of the discrete maximum principle in some simple cases.

Having in mind the idea that only a modification of the crosswind diffusion is necessary, the discrete maximum principle provides the theoretical grounds for the dessign of the new dissipation. Assuming first that an isotropic diffusion is added to

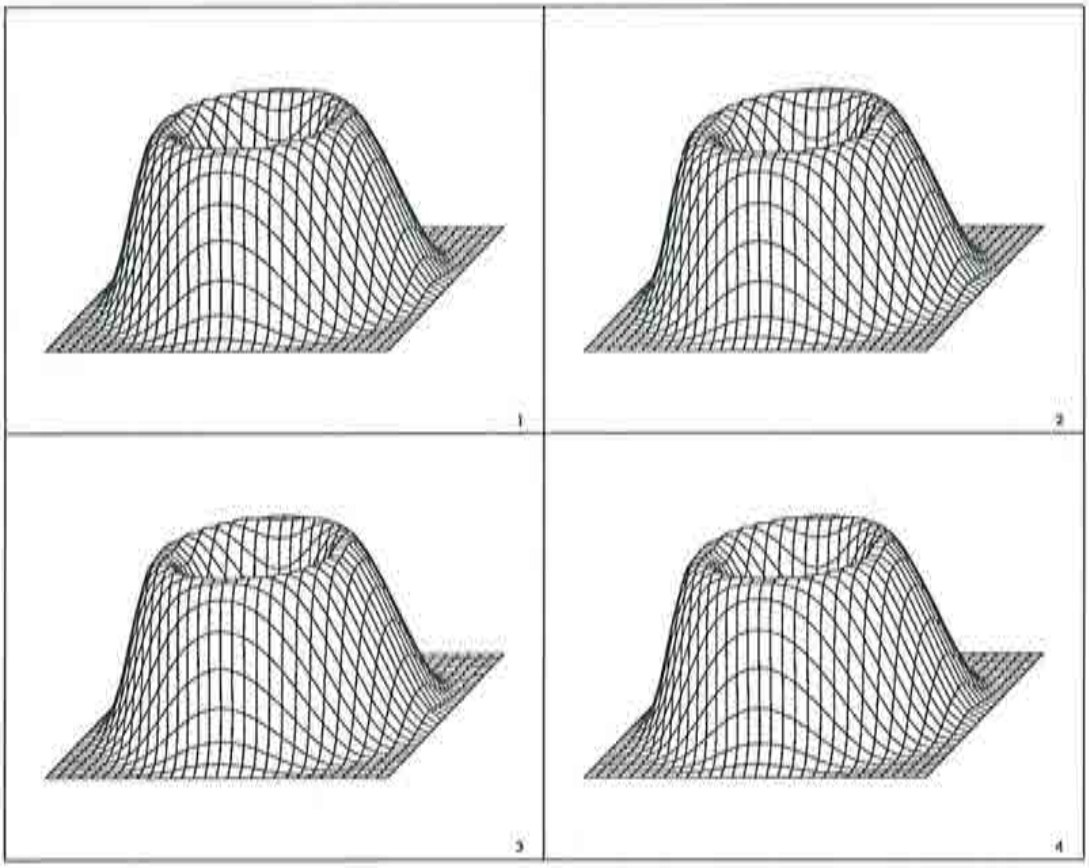

Figure 3.22 Numerical solution of Example 3.4 using biquadratic elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

the Galerkin formulation, for two particular cases it has been shown that this dissipation can be taken as indicated by (3.57), with the function α_c given by (3.59). The bound $\alpha_c \geq C - 1/\gamma_{||}$ is the sharpest we have been able to obtain. Observe that $\gamma_{||}$ is the smallest of the possible pseudo-Péclet numbers that can be computed with vectors $\mathbf{v}$ such that $\mathbf{v} \cdot \nabla \phi_h = \mathbf{u} \cdot \nabla \phi_h$. The question that remains open is, as for most numerical methods, the election of the algorithmic constants. Our choice has been based on numerical experimentation.

In order to obtain a consistent method when quadratic elements are used and/or source terms are present, the straightforward extension of the numerical dissipation (3.57) is the use of (3.58). If this is done, the exact solution shall satisfy the numerical scheme, provided it is smooth enough so that the continuous equation (3.1) makes sense.

The important point is that the streamline dissipation associated to the SUPG formulation is greater that any of the lower bounds obtained for the isotropic dissipation that must be introduced using an artificial diffusion method (up to the choice of the algorithmic constants). Therefore, the natural idea is to introduce (3.57) or (3.58) only as a crosswind diffusion, not isotropic. This is in summary what we propose.

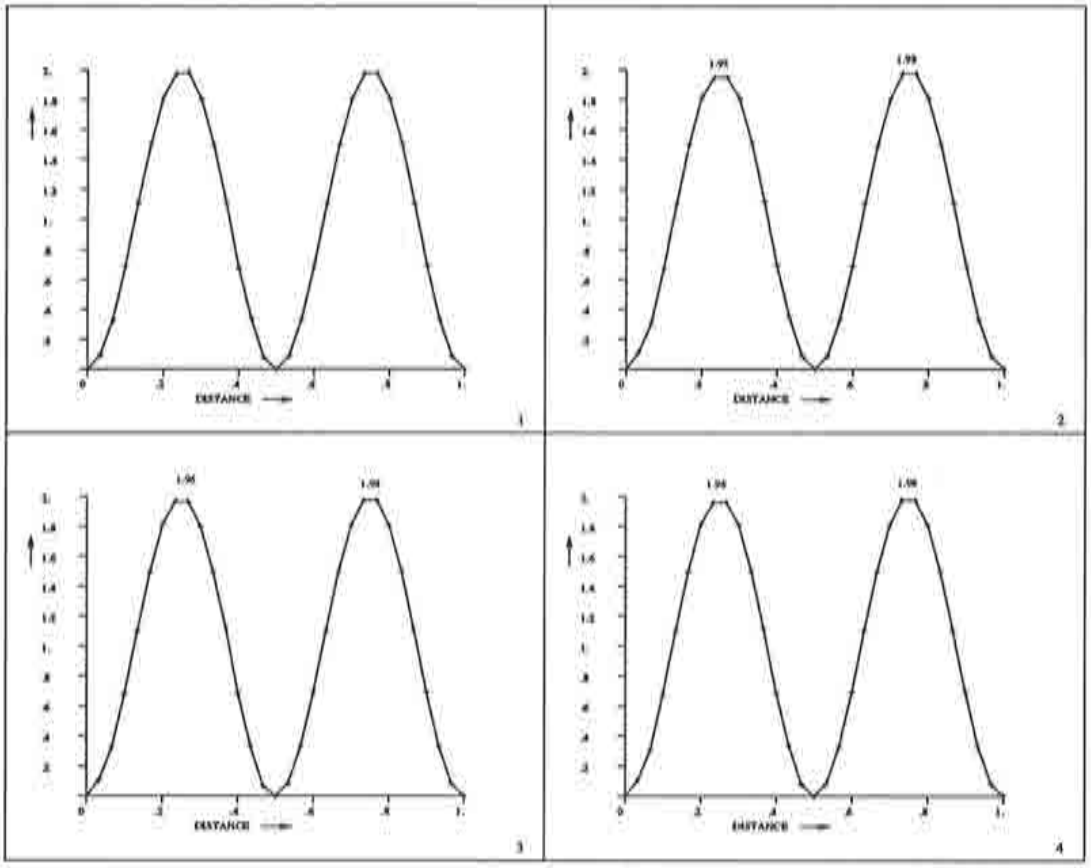

Figure 3.23 Section $y = 0.5$ for Example 3.4 using biquadratic elements. (1): SUPG formulation; (2): DC of type 2; (3): Modified DC; (4): CD.

Also, it has been noticed that the upwind function of the SUPG method may not be arbitrarily large, in contrast with what happens using purely an artificial dissipation.

Concerning the practical behavior of this new approach, from the numerical experiments presented in the last section several conclusions may be drawn. We have seen that this new method, compared to some other shock-capturing techniques,

- has a much better convergence rate towards the steady-state when a transient relaxation is used to solve the nonlinear discrete problem,
- is less overdiffusive,
- has a similar computational cost.

We may therefore conclude that its numerical performance makes it an attractive shock-capturing technique. Also, a theoretical basis exists for some particular cases, although generalizations are always necessarily *ad hoc*.

References

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

[CR] P.G. Ciarlet and P.A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Meths. Appl. Mech. Engrg.*, vol. 2, (1973) 17–31

[CH] R. Courant and D. Hilbert. *Methods of mathematical physics*, vol. 2. (Wiley-Interscience, 1962).

[Da] S.F. Davis. A rotationally biased upwind difference scheme for the Euler equations *J. Comput. Physics*, vol. 39, (1981) 164–178

[DG] E.G. Dutra do Carmo and A.C. Galeão. Feedback Petrov-Galerkin methods for convection dominated problems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 88, (1991) 1–16

[GD] A.C. Galeão and E.G. Dutra do Carmo. A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 68, (1988) 83–95

[Hi] C. Hirsch. *Numerical computation of internal and external flows*, vols. 1 and 2. (John Wiley & Sons, 1990)

[HM] T.J.R. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems. *Comput. Meths. Appl. Mech. Engrg.*, vol. 58, (1986) 329–336

[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] T. Ikeda. *Maximum principle in finite element models for convection-diffusion phenomena.* (North-Holland/Kinokuniya, 1983)

[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

[JSW] C. Johnson, A.H. Schatz and L.B. Wahlbin. Crosswind smear and pointwise errors in streamline diffusion finite element methods. *Math. Comput.*, vol. 49, (1987) 25–38

[JS] C. Johnson and A. Szepessy. On the convergence of a finite element method for a nonlinear hyperbolic conservation law. *Math. Comput.*, vol. 49, (1987) 427–444

[JSH] C. Johnson, A. Szepessy and P. Hansbo. On the convergence of shock-capturing streamline finite element methods for hyperbolic conservation laws. *Technical report 1987-21*, Mathematics Department, Chalmers University of Technology, Göteborg, 1987.

[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)

[LV] R.J. LeVeque. *Numerical methods for conservation laws.* (Birkhäuser, 1990)

[MH] A. Mizukami and T.J.R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle. *Comput. Meths. Appl. Mech. Engrg.*, vol. 50, (1985) 181–193

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

[Nj] K. Niijima. Pointwise error estimates for a streamline-diffusion finite element scheme. *Numer. Math.*, vol. 56, (1990) 707–719

[Nt] J.A. Nitsche. L^{∞}-convergence for finite element approximations, 2. *Conference on finite elements*, Rennes, France, May 12–14, 1975.

[OB] E.S. Oran and J.P. Boris. *Numerical simulation of reactive flow.* (Elsevier, 1987)

[PT] R. Peyret and D. Taylor. *Computational methods for fluid flow.* (Springer-Verlag, 1983)

[RS] J.G. Rice and R.J. Schnipke. A monotone streamline upwind finite element method for convection dominated flows. *Comput. Meths. Appl. Mech. Engrg.*, vol. 48, (1985) 313–327

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

[TP] T.E. Tezduyar and Y.J. Park. Discontinuity-capturing finite element formulations for nonlinear convection-diffusion-reaction equations. *Comput. Meths. Appl. Mech. Engrg.*, vol. 59, (1986) 307–325

[Wa] L.B. Wahlbin. Maximum norm error estimates in the finite element method with isoparametric quadratic elements and numerical analysis. *RAIRO Anal. Numer.*, vol. 12, (1978) 173–262