UNIVERSITÀ DEGLI STUDI DELL' INSUBRIA

FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI DI COMO
DIPARTIMENTO DI SCIENZA E ALTA TECNOLOGIA

DOTTORATO DI RICERCA IN MATEMATICA DEL CALCOLO - XXIV CICLO

SOLUTION OF THE 2D NAVIER-STOKES EQUATIONS BY A NEW FE FRACTIONAL STEP METHOD

Supervisor : Prof. Vincenzo Pennati

Coordinator : Prof. Stefano Serra Capizzano

Ph.D. thesis of :

ANTOINE CELESTIN KENGNI JOTSA

Matr. N. 710695

26 MARCH 2012

Contents

Abstract **V**

Riassunto **VII**

Introduction **1**

1 The mathematical model **5**
1.1 Incompressible Navier Stokes equations 5
1.1.1 Boundary conditions 6
1.2 Well-posedness 7
1.3 Problem to be solved 8
1.4 Time advancing by finite difference 9
1.5 The Stokes problem 10
1.5.1 Some definitions and results about Sobolev spaces 10
1.6 Operator splittings for a general problem 14
1.7 Fractional step methods for Navier-Stokes problem 17
1.8 Advancing time to be used 19
1.9 Spatial discretization 20
1.9.1 Classical finite difference (FD) method 20
1.9.2 Finite volume (FV) method 22

2 Solution of elliptic problems **25**
2.1 The finite element (FE) method for the Poisson equation 25
2.1.1 The discrete problem 27
2.1.2 Interpolation operator and error 29
2.1.3 Generation of the shape functions in the 1D case 29
2.1.4 Generation of the shape functions in 2D case 32
2.2 Discontinuous Galerkin (DG) finite elements methods 34
2.2.1 Enforcing Dirichlet boundary condition through penalties 34
2.2.2 The flux and the primal formulation 35
2.2.3 Choices of numerical fluxes 38
2.2.4 Other choices of the numerical fluxes 41
2.2.5 Convergence of the DG 43
2.3 The finite volume-element (FVE) method 46
2.4 New conservative finite element method 47
2.4.1 Algebraic description 49
2.4.2 Treatment of boundary conditions 50
2.4.3 Numerical results 51
2.5 Conclusions of the chapter 52

3 Domain decomposition methods **53**
3.1 Introduction . . . 53
3.2 Multi-domain techniques . . . 53
3.2.1 Fictitious domain (or domain embedding) methods . . . 53
3.2.2 Disjoint partitions . . . 54
3.2.3 Overlapping sub-domains . . . 54
3.3 Usefulness of multi-domain techniques . . . 54
3.4 Solution of algebraic systems . . . 55
3.4.1 Direct and iterative methods . . . 55
3.5 Convergence estimates . . . 57
3.5.1 Preconditioner . . . 57
3.5.2 Restriction and prolongation operators . . . 58
3.6 Solution of a Poisson equation . . . 60
3.6.1 The Schwarz approach . . . 60
3.6.2 One level Schwarz . . . 66
3.6.3 The colouring technique . . . 69
3.6.4 Two level Schwarz methods . . . 69
3.6.5 Convergence estimates . . . 70
3.6.6 Why use a Schwarz preconditioner? . . . 72
3.7 Application of multi-domain methods to other problems . . . 72
3.7.1 The generalized Stokes problem . . . 72
3.7.2 The inviscid generalized Stokes problem . . . 76
3.7.3 Convergence estimates . . . 78
3.8 Numerical experiences . . . 78
3.9 Conclusions of the chapter . . . 82

4 Parabolic and convection diffusion problems **85**
4.1 Initial boundary value problems and weak formulation . . . 85
4.1.1 Mathematical analysis . . . 86
4.1.2 Semi-discrete approximation by FE . . . 90
4.1.3 Time advancing by FD . . . 93
4.2 Approximation of convection-dominated problems . . . 95
4.2.1 The streamline diffusion method . . . 96
4.2.2 Interpretation of the additional term in the case of linear elements . . . 97
4.2.3 Analysis of the streamline diffusion method . . . 98
4.2.4 The characteristic Galerkin method . . . 102
4.3 New fractional step for convection diffusion equations . . . 105
4.3.1 Weak and algebraic formulation . . . 105
4.3.2 Computation of the step 3 of the iterative procedure . . . 106
4.3.3 Conservation property . . . 107
4.3.4 Numerical tests . . . 108
4.4 Conclusions of the chapter . . . 112

5 Solution of Navier-Stokes equations **113**
5.1 Equal-order finite element and characteristic-fractional step approach . . . 113
5.1.1 Choice of functional spaces . . . 113
5.1.2 Characteristic - fractional step and weak formulation . . . 114
5.1.3 Algebraic formulation of the equal-order approximation . . . 116
5.2 The inf-sup condition and some stabilization methods . . . 117
5.3 New algebraic stabilized method . . . 118
5.4 Numerical tests: two theoretical problems . . . 119
5.4.1 Two dimensional time dependent problem . . . 119

5.4.2 Two dimensional stationary problem 124
5.5 Numerical tests: a real application 139
5.5.1 Natural Convection in a square cavity 139
5.6 Conclusions of the chapter 146

6 Solution of Shallow-Water equations 147
6.1 An environmental problem 147
6.2 The mathematical model 148
6.3 The numerical model 148
6.4 Algebraic formulation 149
6.5 A problem with an analytical solution 151
6.6 Conclusions of the chapter 152

7 Software 157
7.1 Preprocessor 157
7.1.1 TRIANGLE 157
7.1.2 ANVI 158
7.2 ELFICS 159
7.3 ROTINS 163
7.3.1 MODULE : computation of matrices 163
7.3.2 MODULE : solution of algebraic system without Schwarz preconditioner 163
7.3.3 MODULE : solution of algebraic system with Schwarz overlapping preconditioner 163
7.4 Post-processor 164

Conclusions 165

Acknowledgements 167

Bibliography 169

Abstract

In this work, a mathematical and numerical approach for the solution of the 2D Navier-Stokes equations for incompressible fluid flow problems is investigated. A new flux conservative technique for the solution of the elliptic part of the equations is formulated.
In the new model, the non linear convective terms of the momentum equations are approximated by means of characteristics and the spatial approximations, of equal order, are obtained by polynomials of degree two. The advancing in time is afforded by a fractional step method combined with a suitable stabilization technique so that the Inf-Sup condition is respected.
In order to keep down the computational cost, the algebraic systems are solved by an iterative solver (Bi-CGSTAB) preconditioned by means of Schwarz additive scalable preconditioners.
The properties of the new method are verified carrying out several numerical tests. At first, some elliptic, parabolic and convective-diffusive problems are solved and discussed, then the results of some time dependent and stationary 2D Navier-Stokes problems (in particular the well known benchmark problem of the natural convection in a square cavity) are discussed and compared to those found in the literature.
Another, potentially very important application of the numerical tools developed, regards the solution of 1D Shallow-Water equations. In fact the use of the fractional steps scheme for advancing in time and the finite elements (of different polynomial degrees) for the spatial approximation, makes the above mentioned approach computationally profitable and convenient for real applications. The efficiency and accuracy of the numerical model have been checked by solving a theoretical test.
Finally, a brief description of the software suitably developed and used in the tests conclude the thesis.

Keywords: 2D Navier-Stokes, fractional step, finite element, equal order, stabilization, flux conservation, Schwarz preconditioner, 1D Shallow-Water

Riassunto

In questo lavoro è investigato un approccio matematico numerico per la risoluzione delle equazioni 2D di Navier-Stokes per problemi di fluidi incomprimibili ed è formulata una nuova tecnica per la conservazione dei flussi matematici nella risoluzione della parte ellittica delle equazioni.
Nel nuovo modello i termini convettivi non lineari delle equazioni della quantità di moto sono approssimati con le caratteristiche e l'approssimazione spaziale è eseguita con polinomi di grado due sia per la velocità che per la pressione. L'avanzamento temporale è compiuto secondo un metodo ai passi frazionari includendo opportune tecniche di stabilizzazione, così da rispettare la condizione Inf-Sup.
Il costo computazionale è mantenuto limitato risolvendo i sistemi algebrici con un solutore iterativo (Bi-CGSTAB) e precondizionandoli per mezzo di precondizionatori scalabili del tipo di Schwarz additivo.
Le proprietà del nuovo metodo sono verificate risolvendo numerosi test numerici. All'inizio sono risolti e discussi alcuni problemi ellittici, parabolici e convettivo-diffusivi. Successivamente i risultati di alcuni problemi 2D di Navier-Stokes, sia transitori che stazionari (tra cui in particolare il ben noto problema della convezione naturale in una cavità quadrata), sono discussi e confrontati con quelli reperibili nella letteratura scientifica.
Un'altra applicazione, potenzialmente molto importante, degli schemi numerici sviluppati riguarda la risoluzione delle equazioni 1D del moto delle acque basse (Shallow-Water equations). Infatti l'uso dello schema ai passi frazionari per l'avanzamento temporale e l'uso di elementi finiti (di grado polinomiale differente) per l'approximazione spaziale, fa sì che il nuovo approcio sia computazionalmente vantaggioso ed adatto anche allo studio del trasporto di inquinanti passivi nei corsi d'acqua. L'efficienza ed accurattezza del modello numerico sono state verificate risolvendo un test teorico con soluzione analitica nota.
Nell' ultima parte della tesi è riportata una breve descrizione del software in cui sono implementate le tecniche sopradescritte e utilizzate nei casi numerici.

Parole chiave : 2D Navier-Stokes, passi frazionari, elementi finiti, egual ordine pressione-velocità, stabilizzazione, conservazione dei flussi, precondizionatore di Schwarz, 1D Shallow-Water

Introduction

Motivation

Numerical modeling has become nowadays an important subject in applied mathematics and in several aspects of technical and industrial activity. Particularly in fluid dynamics, the incompressible Navier-Stokes equations are by now widely accepted as a mathematical model for incompressible viscous fluid flows.
Since a large amount of numerical methods has been already developed for solving fluid dynamic problems, we considered useful to reserve an important part of the thesis to an extensive review of the existing literature. In fact, the aim of the work was to develop new stable FE approaches for the solution of 2D incompressible Navier-Stokes equations and of 1D Shallow-Water equations, taking particularly into account the accuracy of the numerical solutions and the computational efficiency of the mathematical models.
In detail, at first the most popular FD schemes for advancing in time, included the fractional step ones, and the high order FE spatial approximations for solving 2D elliptic, parabolic and convective-diffusive partial differential equations was studied; then we used the best techniques (in our opinion) for solving the 2D Navier-Stokes equations and the 1D Shallow-Water equations. In spite of appearances, is possible for elliptic problems to find a connection between saddle-point and discontinuous finite element methods so that we decided to develop and to analyze the continuous version of one of the discontinuous Galerkin methods analyzed that is, due to its simple formulation, the Baumann-Oden method.
We decided to use for the spatial approximation piecewise polynomials of degree two so that in our new scheme we are able to impose the continuity of the numerical fluxes by construction. Unfortunately, the new scheme is not conservative according to the classical definition, however it could be generalized so that a scheme generally conservative, like those named finite volume-elements (FVE) could be obtained.
The solution of elliptic problems requires the solution of an algebraic system and this is, generally, computationally expensive. In order to reduce as most as possible this effort (in fact in transient fluid dynamic problems the solution of elliptic problems can be repeated several times, also hundred or thousand), it is widespread to use preconditioned iterative solvers. As the stiffness matrices of the conservative FE are not symmetric, we decided to use the well known Bi-CGSTAB solver preconditioned by a Schwarz preconditioner.
The additive two level Schwarz preconditioners that we build and used were derived from the theory of multi-domains methods with overlapping and have the scalability property.
In order to heuristically verify the properties of the new method, some numerical tests, with or without preconditioners, by conservative or traditional FE, were carried out.
Some tables reporting the error norms and the estimate order of accuracy have been build and by means of them we were able to deduce that (of course) the traditional FE solutions respect the theoretical order of accuracy while the conservative FE solutions seem not have an optimal order of accuracy. A theoretical analysis of this was not carried out, but remembering that in elliptic problems the Galerkin weak formulation guaranties the orthogonality of the error with respect to the spaces to which belong the numerical solutions, it is not surprising that the continuity constraint imposed on the numerical fluxes breaks the above property.
In the solution of transient convection-diffusion problems a particular attention has to be de-

voted to the convective terms. In the past years, in order to guarantee stability and strong consistency, many approaches have been formulated; one of them is the so called characteristic method. This method gives an easy and accurate approximation of the convective terms (particularly in 2D problems) reducing the convection-diffusion equation to a parabolic equation. Its most appealing property is the unconditioned stability in linear problems.
Some numerical tests, both for parabolic and convective-diffusive equations, were carried out and the results suitably evaluated.
The second part of this thesis is dedicated to the solution of the two very important and popular fluid dynamics problems: the 2D Navier-Stokes and 1D Shallow-Water equations.
Apparently the two physical phenomena described by the two systems of equations are very different; in fact the first one represents an incompressible fluid flow problem while the second one a long wave problem, but both are transient problems and have unknown variables the velocity and pressure (even if, actually, in the Shallow-Water equations the hypothesis of hydrostatic pressure allows the "substitution "of the pressure by the elevation of the free surface). Keeping in mind this and considering the obvious differences between the problems (in particular that the first one is 2D while the second one is 1D), we decided to use the same approach for advancing in time and for approximating the convective term, i.e. the fractional step and the characteristic techniques. In this way, at each time instant of the marching technique, the most expensive computational effort was reduced to the solution of algebraic systems stemming from the elliptic problems that we have extensively discussed previously. Both the numerical models (2D Navier-Stokes and 1D Shallow-Water equations) have approximation of first order in time, while the spatial approximation for velocity and pressure for the 2D Navier-Stokes equations is second order and second order for unit width discharge and first order for elevation for the 1D Shallow-Water equations. Another very important aspect of our approach regards the choice of equal order approximation for Navier-Stokes equations; in fact, the non respect of Inf-sup condition requires to use a suitable stabilization technique. Recently, has been presented in literature a method, simple enough from an algorithm point of view, for stabilizing the solutions in presence of equal order choice. We adopted, after the suitable modifications, this method in our models and, at our knowledge, numerical models with the features described thus far have not been published yet.
Not existing for Shallow-Water equations a condition similar to the Inf-Sup, but knowing by numerical experiences that solutions of equal order are unstable also for Shallow-Water problems, we decided to use polynomials basis functions of degree two for unit width discharge and degree one for elevation. In order to check the efficiency of our models, we solved some numerical tests with known analytical solutions both for Navier-Stokes and Shallow-Water problems; moreover, for the Navier-Stokes equations, we solved also a well known benchmark problem and compared our results with those found in literature.
The efficiency and the low computational cost of the numerical model for Shallow-Water make it a promising tool for the prediction of distribution of pollutants in rivers and basins.

Thesis outline

Chapter 1 starts with the derivation of the mathematical model for fluid flow problems. We discuss possible types of boundary conditions which are useful in fluid flow. We shortly recall known well posedness results for the Navier-Stokes equations and we announce the effective problem to be solved. Moreover, definitions of some functional spaces to be used throughout the thesis is given. We present standard approach for the approximation of the Navier-Stokes equations and finally some classical spatial discretization method are addressed.

In Chapter 2, we present the Galerkin FE method for the solution of Poisson equation. In addition we make an overview of some discontinuous Galerkin FE methods for elliptic problems as well as a short presentation of the finite volume-element method. Finally, a new

conservative method for elliptic equations resembling the discontinuous Galerkin method is introduced, followed by numerical results generated by this new method with a comparison with results obtained from the traditional Galerkin finite element method.

Chapter 3 mainly treats the Schwarz domain decomposition approach. We present the Schwarz approach as a mathematical tool for solving elliptic type problems, like Poisson equation. On the other hand, we present the Schwarz overlapping approach as an efficient preconditioner for solving algebraic linear systems and known theoretical results are given. The reasons for choosing Schwarz overlapping additive preconditioner are given followed by a short application of the multi-domain approach to other problems. Finally an efficient algorithm for the construction of the Schwarz preconditioner and its application to the test problems of chapter 2 are provided.

In Chapter 4, we review some numerical approaches for the solution of the time dependent parabolic problem with known theoretical results as well as for time dependent convection diffusion problem. Moreover, the characteristic Galerkin method is presented and finally a new fractional approach for solving linear convection diffusion equation is given; some results by some numerical tests are presented. This scheme in the next chapter will be used for the solution to the 2D incompressible Navier-Stokes equations and for the solution of the temperature equation.

Chapter 5 is devoted to our new stabilized approach. At first the fractional step for the advancing in time of 2D incompressible Navier Stokes equations is developed. Then we shortly review the Inf-Sup condition and some stabilized methods, followed by our equal order stabilized approach at the algebraic level. Numerical tests are thus presented to verify the effectiveness and the efficiency of this new approach. We consider two different problems with non trivial known solutions: they are the two dimensional unsteady flow of decaying vortices and the lid driven cavity flow. The aim is to test the convergence behavior of the new method by the maximum and square error norms that can be computed because the exact solution is known. Then we consider a real problem that is the benchmark flow problem of natural convection in a square cavity where no analytic solution is known. The aim here is to compare the results of the new method to literature values in order to compare our method to well established schemes.

Chapter 6 is devoted to the solution of 1D Shallow-Water equations by a numerical approach similar to that used for 2D Navier-Stokes equations. Also for this problem a numerical test with analytical solution has been solved and the results discussed.

Chapter 7 contains a brief description of the software TRIANGLE-ANVI-ELFICS-ROTINS-GNUPLOT in which the method and algorithms from chapters 2-5 are implemented and by which the results of the tests were provided.

Chapter 1

The mathematical model

1.1 Incompressible Navier Stokes equations

We denote the Cartesian spatial coordinates by $X = (x, y, z) = \{x_i\}_{i=1}^d$, $d \in \{2, 3\}$ and the time by t. The vectorial operator of spatial derivative is denoted by $\nabla = \{\partial x_i\}_{i=1}^d$ where ∂_ξ is the spatial derivative with respect to ξ. The scalar operator of partial derivative usually called divergence operator is devoted by $div = \nabla\cdot = \sum_{i=1}^d \partial x_i$. We consider a viscous incompressible fluid and we assume its density ρ and dynamic viscosity μ are dependent on space and time. Within a fixed spatial domain $\Omega \subset \mathbb{R}^d$, and during the time interval $(0, T)$, i.e for $(X, t) \in \Omega \times (0, T)$, the distribution of the velocity field $\underline{u} = \underline{u}(X, t) = (u_i)_{i=1}^d$ and the pressure field $p = p(X, t)$ of the fluid is modeled by the incompressible Navier Stokes equations given in their non conservative form :

$$\begin{aligned} \rho\partial_t(\underline{u}) + \rho(\underline{u}\cdot\nabla)\underline{u} - \nabla\cdot(2\mu\underline{\mathtt{D}}(\underline{u})) + \nabla p &= \underline{f} \quad \text{in} \quad \Omega\times(0,T) && (1.1)\\ \nabla\cdot\underline{u} &= 0 \quad \text{in} \quad \Omega\times(0,T) && (1.2) \end{aligned}$$

where $(\underline{\mathtt{D}}(\underline{u}))_{i,j} = \frac{1}{2}(\nabla\underline{u} + \nabla\underline{u}^T) = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})_{1\le i,j\le d}$ is the deformation rate or strain tensor also called the symmetric gradient of $\underline{u}$ and $\underline{f}$ denotes a volumetric force as for instance gravitation. Equation (1.1) enforces the conservation of momentum while equation (1.2) guarantees the incompressibility constraint equivalent to the conservation of volume. We say that a fluid is incompressible if the volume of any sub-domain $\mathtt{V} \subset \Omega$ remains constant for all times $t > 0$, i.e

$$\int_{\mathtt{V}(t)} dx = \int_{\mathtt{V}} dx \quad \text{for all} \quad t > 0$$

We note that $\nabla\cdot\underline{u}$ is the trace of the deformation tensor $(\underline{\mathtt{D}}(\underline{u}))$. The last two terms of the left hand side of equation (1.1) can also be written as $-\nabla\cdot\underline{\mathtt{T}}(\underline{u}, p)$ where

$$\underline{\mathtt{T}}(\underline{u}, p) = 2\mu(\underline{\mathtt{D}}(\underline{u})) - p\underline{\mathtt{I}}$$

is the stress tensor and $\underline{\mathtt{I}} = (\delta_{ij})_{1\le i,j\le d}$ denotes the unit tensor. We can also write the divergence of a tensor $\underline{\mathtt{S}}$:

$$(\nabla\cdot\underline{\mathtt{S}})_i = \sum_{k=1}^{d} \partial_{x_k}\mathtt{s}_{\mathtt{ik}}$$

For a detailed derivation of tensor and justification of this model see [76].
Generally when solving real problems, to make complete the formulation, we have:

- To add a (or several) convection diffusion equation (equations) for a (or several) scalar variable (variables) , e.g. for temperature $\mathtt{T}$ given by

$$\rho\frac{\partial\mathtt{T}}{\partial t} + \underline{u}\cdot\nabla\mathtt{T} - \nabla\cdot(k\nabla\mathtt{T}) = \mathtt{S} \quad \text{in} \quad \Omega\times(0,T) \qquad (1.3)$$

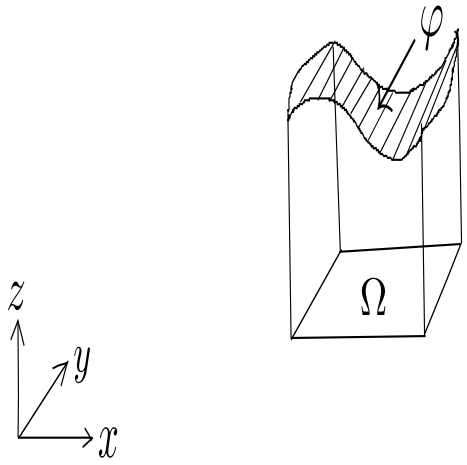

Figure 1.1: Perspective view of a physical flow .

where k is the thermal conductivity

- To take into account, for $d = 3$, the free surface evolution, because when the fluid is flowing, a part of its boundary is changing in time due to the presence of some externals forces such as the wind. Denoting by $\varphi(x, y, t)$ the free surface, its evolution should satisfy the cinematic equation
$$w = \frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y} \tag{1.4}$$
see Figure 1.1

- To include a turbulence model, e.g. $K-\varepsilon$ which are exactly two more convection diffusion equations to be solved. We refer to [80, 81] for a detailed description of the $K-\varepsilon$ model.

Finally the flow model has also to be completed with suitable initial and boundary conditions. For the initial conditions, it is clear that an initial velocity field $\underline{u}_0$ is necessary:

$$\underline{u}|_{t=0} = \underline{u}_0 \quad \text{in} \quad \Omega$$

For the boundary conditions, different possibilities are presented in Subsection 1.1.1 .

1.1.1 Boundary conditions

We partition the boundary $\partial\Omega$ of Ω into a finite number n of subsets $\gamma_i, i = 1, \ldots n$. In order for the Navier-Stokes equations to be well-posed, suitable boundary conditions need to be specified on each subset γ_i. Many different types of boundary conditions are possible in principle; also the kind of boundary condition can vary from point to point on the boundary, but at any given point only one boundary condition can be specified. We refer to [113] for an extensive enumeration of these possibilities and consider just the following ones.

Dirichlet boundary conditions

Generally, Dirichlet boundary conditions prescribe a velocity field $\underline{g}_D$

$$\underline{u} = \underline{g}_D \quad \text{on} \quad \gamma_i \tag{1.5}$$

They are usually applied to impose an inflow velocity profile $\underline{g}_D$, or to model a wall moving with velocity $\underline{g}_D$. In the latter case, they are also called no-slip boundary conditions as they impose the fluid not to slip but to stick at the wall.

Neumann boundary conditions

Neumann boundary conditions prescribes a force $\underline{g}_N$ per unit area as the normal components of the stress tensor

$$\underline{\mathtt{T}}(\underline{u},p)\cdot n = 2\mu(\underline{\mathtt{D}}(\underline{u}))\cdot n - p\cdot n = \underline{g}_N \quad \text{on} \quad \gamma_i$$

where n is the outer unit normal on γ_i. Neumann boundary conditions are used to model a given force per unit area on the boundary γ_i, often with $\underline{g}_N = 0$ for what is called free outflow. For vanishing velocity gradients, the force $\underline{g}_N$ corresponds to the pressure on the boundary. Usually, Neumann boundary conditions specify a flux condition on the boundary. See also [64] for more details about the interpretation and implication of this type of boundary conditions.

Mixed boundary conditions

Mixed boundary conditions combines Dirichlet boundary conditions in the normal direction n for velocity with null components of the stress tensor in the tangential direction(s) τ :

$$\begin{aligned} \underline{u}\cdot n &= \underline{g}_D\cdot n \quad \text{on} \quad \gamma_i \\ (\underline{\mathtt{T}}(\underline{u},p)\cdot n)\cdot\tau = (2\mu\underline{\mathtt{D}}(\underline{u})\cdot n)\cdot\tau &= 0 \quad \text{on} \quad \gamma_i \quad \forall\tau : \tau\cdot n = 0 \end{aligned}$$

In particular, when $g_D = 0$ we talk about free slip boundary conditions.

Mixed Robin boundary conditions

Sometimes, a smooth transition from slip to no-slip boundary conditions is desired. This can be realized by imposing Dirichlet boundary conditions in the normal direction as for the slip boundary conditions and to replace the boundary conditions in the tangential direction by Robin boundary conditions, as a linear combination of Dirichlet and Neumann boundary conditions :

$$\begin{aligned} \underline{u}\cdot n &= \underline{g}_D \quad \text{on} \quad \gamma_i && (1.6) \\ (\omega C_\tau\underline{u} + (1-\omega)(\underline{\mathtt{T}}(\underline{u},p)\cdot n)\cdot\tau = \omega C_\tau\underline{u} + (1-\omega)(2\mu\underline{\mathtt{D}}(\underline{u})\cdot n)\cdot\tau &= \omega C_\tau\underline{g}_D\cdot\tau \,\text{on}\,\gamma_i && (1.7) \\ \forall\tau : \tau\cdot n = 0 && && (1.8) \end{aligned}$$

Here, $\omega \in [0,1]$ depends on the actual flow problem. For $\omega = 0$, we have free slip conditions whereas for $\omega = 1$, we have no-slip boundary conditions. In practice, ω can be a smooth function of space and time with values in $[0,1]$ allowing thus a smooth transition between the two cases. This holds for $\underline{g}_D = 0$ but transition boundary conditions cover also the general Dirichlet case for $\underline{g}_D \neq 0$ and $\omega = 1$. The weight C_τ depends on the velocity field and the force per unit area. This particular type of boundary conditions has been studied in more details in [67].

1.2 Well-posedness

In this part, we consider some important issues regarding the question of well-posedness of the problems described by the mathematical model equations, and of existence and uniqueness of their solutions. A global existence result for the coupled problem $(1.1) - (1.2)$ with $\underline{f} = \rho\underline{g}$ has been proven by Lions [85]. This proof requires Ω to be a smooth, bounded, connected open subset of $\mathbb{R}^d$, and that homogeneous Dirichlet boundary conditions (i.e with $\underline{g}_D = 0$) are

imposed on the whole boundary. If ρ_0, the initial data u_0 and the source data $\underline{f}$ satisfy

$$\begin{aligned} \rho_0 &\geq 0 \quad \text{a.e. in} \quad \Omega \\ \rho_0 &\in \mathrm{L}^\infty(\Omega) \\ \rho_0 \underline{u}_0 &\in (\mathrm{L}^2(\Omega))^d \\ \rho_0 |\underline{u}_0|^2 &\in \mathrm{L}^1(\Omega) \\ \underline{f} &\in (\mathrm{L}^2(\Omega \times (0,T)))^d \end{aligned}$$

then, there exist global solutions which satisfy

$$\begin{aligned} \rho &\in \mathrm{L}^\infty(\Omega \times (0,T)) \\ \underline{u} &\in (\mathrm{L}^2(0,T;\mathrm{H}^1_0(\Omega)))^d \\ \rho |\underline{u}|^2 &\in \mathrm{L}^\infty(0,T;\mathrm{L}^1(\Omega)) \\ \nabla \underline{u} &\in \mathrm{L}^2(\Omega \times (0,T)) \\ \rho &\in \mathcal{C}([0,\infty], \mathrm{L}^s(\Omega)) \quad \forall\ s \in [1,\infty) \end{aligned}$$

These solutions are called weak solutions and nothing more than the above weak statements are known. Specifically nothing is known about the regularity of the pressure field p. We note that the pressure is not uniquely defined; in fact if $(\underline{u}, p)$ is a solution of (1.1) then $(\underline{u}, p + c)$, $c \in \mathbb{R}$ is also a solution of the same equation. When Dirichlet boundary conditions are specified on the entire boundary $\partial\Omega$, g_D has to satisfy the compatibility condition

$$\int_{\partial\Omega} \underline{g}_D \cdot n ds = 0$$

otherwise the solution of the problem couldn't respect the incompressibility constraint. These solutions are also called solution "à la Leray" by analogy with the classical global existence results for the homogeneous incompressible Navier-Stokes problems obtained by Leray in [77, 78, 79]. However, if $\rho_0 > 0$ then it is proven [85] that there is a short time smooth solution to which all weak solutions are equal.
This situation is not very satisfactory, however it does not prevent us from developing well posed numerical models approximating the mathematical models and providing solutions which are in good agreement with physical observations.
We assume in what follows that the density of the fluid ρ is a positive constant and so the kinematic viscosity $\nu = \mu(\rho)/\rho$ is also constant.

1.3 Problem to be solved

Motivations

Generally when solving the 3D incompressible Navier-Stokes equations, we have to deal at each time instant and all together with :

- 3 momentum equations for the velocity field $\underline{u}$
- 1 equation for the incompressibility for p
- 1 equation for the energy eventually written in term of the temperature
- 1 or more equations for turbulence models

- and 1 cinematic equation for the free surface.

Of course, the numerical solution of such system is difficult and heavy, in particular if we are interested in the development and in the test of new numerical tools. For this reason, we are going to consider only the 2 dimensional counter part of the incompressible Navier-Stokes equations. In this case, there is no more the presence of the free surface evolution and of the turbulence models. Therefore we are looking for the distribution of the two components of the velocity field $\underline{u}(u(x,y,t),v(x,y,t))$ and the pressure field $p(x,y,t)$ so that the equations in non conservative form are :

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}-\mu(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2})+\frac{\partial p}{\partial x} = f_u \quad \text{in} \quad \Omega\times(0,\mathtt{T}) \tag{1.9}$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}-\mu(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2})+\frac{\partial p}{\partial y} = f_v \quad \text{in} \quad \Omega\times(0,\mathtt{T}) \tag{1.10}$$

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y} = 0 \quad \text{in} \quad \Omega\times(0,\mathtt{T}) \tag{1.11}$$

$$\oplus \quad \text{suitable} \quad \text{B.C} \quad \oplus \quad \text{I.C}\ , \tag{1.12}$$

where ρ is constant and $p=\frac{p}{\rho}$.
We note that the two momentums equations can be seen as two non linear convection-diffusion equations where the non linear hyperbolic part is a convection (advection) type equation whereas the elliptic part is of Poisson type operator. Since we are in presence of a time dependent problem, it is necessary to choose a strategy in order to advance in time. This will be the object of the section 1.4 where we are giving known results.

1.4 Time advancing by finite difference

The system of equations (1.9) − (1.11) can be advance in time by suitable finite difference scheme. To start, we partition the time- interval $[0,T]$ into $\mathcal{N}$ subintervals $[t^n,t^{n+1}]$ of measure $\Delta t=\frac{T}{\mathcal{N}}$ with $t^0=0$ and $t_{\mathcal{N}}=T$. If we consider the single-step scheme [113], then at each time level $t^{n+1}=(n+1)\Delta t, n=0,1,\cdots,\mathcal{N}-1$ the system becomes :

$$\frac{\underline{u}^{n+1}-\underline{u}^n}{\Delta t}+\underline{u}_\theta^{n+1}\cdot\nabla\underline{u}_\theta^{n+1}-\mu\nabla\cdot(\nabla\underline{u}_\theta^{n+1})+\nabla p^{n+1} = \underline{f}_\theta^{n+1} \tag{1.13}$$

$$\nabla\cdot\underline{u}^{n+1} = 0 \tag{1.14}$$

with $\underline{u}_\theta^{n+1}:=\theta\underline{u}^{n+1}+(1-\theta)\underline{u}^n$, and $\underline{f}_\theta^{n+1}:=\underline{f}(\theta\underline{t}^{n+1}+(1-\theta)\underline{t}^n)$ for $0\le\theta\le 1$. This scheme is second order accurate with respect to Δt if $\theta=\frac{1}{2}$, while it is only of first order accurate for all the other values of θ. We note that accuracy is measured in H^1- norm for each velocity component and L^2- norm for pressure. When $\theta=1$, the scheme is fully implicit and require the solution of a nonlinear system at every time increment. The nonlinearity can be overcome by a second order Newton-Raphson iterative method [113] or a GMRES iterative method, see Brown and Saad [23]. We note that a second order backward differentiation scheme is obtained by setting $\theta=1$ in (1.13) and by replacing the first order backward difference $\frac{\underline{u}^{n+1}-\underline{u}^n}{\Delta t}$ by the two step second order one $\frac{3\underline{u}^{n+1}-4\underline{u}^n-\underline{u}^{n-1}}{2\Delta t}$. However this scheme requires larger storage compared with the single-step second order Crank-Nicolson method. Moreover, it needs an additional second order initialization u^1.
In order to avoid the implicit scheme, one can resort implicit methods by linearizing the momentum equations (1.13) by Newton method and performing only one iteration at each time step giving rise to a semi implicit method as :

$$\frac{\underline{u}^{n+1}-\underline{u}^n}{\Delta t}+\underline{u}^n\cdot\nabla\underline{u}^{n+1}-\mu\nabla\cdot(\nabla\underline{u}^{n+1})+\nabla p^{n+1} = \underline{f}^{n+1} \tag{1.15}$$

$$\nabla\cdot\underline{u}^{n+1} = 0 \tag{1.16}$$

In addition, we mention that when $\theta = 0$, the scheme is fully explicit and the convective term in this particular case is completely linear. Moreover, it is important to note that the incompressibility and the pressure have the same time increment level otherwise it can result to a divergence of the solution.
Other semi-implicit approaches requires a fully explicit treatment of the nonlinear term. An illustration is provided by the following second-order scheme which approximates the linear terms by the Crank-Nicolson method and the nonlinear (convective) one by the explicit Adams-Bashforth method. From [113], it is given by :

$$\begin{cases} \frac{2}{\Delta t}(\underline{u}^{n+1} - \underline{u}^n) + \frac{3}{2}\underline{u}^n \cdot \nabla \underline{u}^n - \frac{1}{2}\underline{u}^{n-1} \cdot \nabla \underline{u}^{n-1} + \mu \nabla \cdot (\nabla \underline{u}^{n+1}) + \nabla p^{n+1} + \nabla p^n = \underline{f}^{n+1} + \underline{f}^n \\ \nabla \cdot \underline{u}^{n+1} = 0 \end{cases} \tag{1.17}$$

for $n = 1, 2, \cdots \mathcal{N} - 1$, providing a suitable initialization of $\underline{u}^1$.
This discretization is second order accurate with respect to Δt providing the data and the solutions are smooth enough with respect to the variable. It can be noted that the system (1.17) is often used when the spatial approximation is based on the spectral collocation method, as the latter can take advantage of explicit evaluation of the nonlinear terms. Therefore in this case, the method is stable under the condition $\Delta t = O(N^{-2})$, N being the polynomial degree of the vector unknown field.
Others semi implicit scheme, that of Gunzberger [51], is formulated as follows:

$$\begin{aligned} \frac{1}{\Delta t}(\frac{3}{2}\underline{u}^{n+1} - 2\underline{u}^n + \frac{1}{2}\underline{u}^{n-1}) + \underline{u}^* \cdot \nabla \underline{u}^* + \nabla p^{n+1} + \mu \nabla \cdot (\nabla \underline{u}^{n+1}) &= \underline{f}^{n+1} \\ \nabla \cdot \underline{u}^{n-1} &= 0 \end{aligned}$$

where $\underline{u}^* := 2\underline{u}^n - \underline{u}^{n-1}$ and $\underline{u}^{-1} = 0$.
This scheme is based on a backward difference formula which is second order accurate in time, is conditionally stable and moreover, Baker, Dougalis and Karakashian proved in [15] that the stability is restricted to $\Delta t = O(h^{\frac{4}{5}})$.
Generally, when adopting a semi-implicit scheme, the final problem obtained is of Stokes type, hence the necessity to give well-known theorems related to the Stokes problem. This is the aim of the section 1.5.

1.5 The Stokes problem

Let us consider the Stokes problem given in this form :

$$\alpha \underline{u} - \nu \Delta \underline{u} + \nabla p = \underline{f} \quad \text{in} \quad \Omega \subset \mathbb{R}^d, \quad d = 2, 3 \tag{1.18}$$

$$\nabla \cdot \underline{u} = 0 \quad \text{in} \quad \Omega \tag{1.19}$$

$$\underline{u} = 0 \quad \text{on} \quad \partial\Omega \tag{1.20}$$

where $\alpha \geq 0$ and $\nu > 0$ are given constants, $\underline{f} : \Omega \longrightarrow \mathbb{R}^d$ is a given function, while $\underline{u} : \Omega \longrightarrow \mathbb{R}^d$ and $p : \Omega \longrightarrow \mathbb{R}$ are problem unknowns. Without loss of generality, we consider homogeneous boundary conditions. We give some definitions and results about the Sobolev spaces.

1.5.1 Some definitions and results about Sobolev spaces

We give the following definition of the Sobolev space $\mathrm{H}^m(\Omega)$.

Definition 1.1. *Given m integer ≥ 0, the Sobolev space $\mathrm{H}^m(\Omega)$ is defined as*

$$\mathrm{H}^m(\Omega) = \{v \in \mathrm{L}^2(\Omega) : D^\alpha v \in \mathrm{L}^2(\Omega), \forall |\alpha| \leq m\}$$

where $D^\alpha v$ represents a partial derivative taken in the sense of distribution,

$$D^\alpha v = \frac{\partial^\alpha v}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}, |\alpha| = \alpha_1 + \cdots + \alpha_d$$

On this space, the norm is given by

$$\|v\|^2_{\mathrm{H}^m(\Omega)} = \sum_{k \le m} |v|^2_{k,\Omega}$$

where

$$|v|^2_{k,\Omega} = \sum_{|\alpha|=k} |D^\alpha v|^2_{\mathrm{L}^2(\Omega)}$$

The space $\mathrm{L}^2(\Omega)$ is then $\mathrm{H}^0(\Omega)$ while $\mathrm{H}(div, \Omega)$ is a subset of $\mathrm{H}^1(\Omega)$. We define the space

$$\mathrm{H}_{div} := \{v \in (\mathrm{L}^2(\Omega))^d / div\, v = 0 \quad in\ \Omega, v \cdot n = 0 \quad on\ \partial\Omega\} \tag{1.21}$$

Then as norm in H_{div}, we choose the $\mathrm{H}^1(\Omega)$-semi-norm $|\cdot|$.

Definition 1.2. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. We define its support as the closure of the set $\{x \in \mathbb{R}^d \quad such\ that \quad \varphi(x) \neq 0\}$.*

Definition 1.3. *We define $\mathcal{D}(\Omega)$ the space of infinitely differentiable functions having compact support, i.e vanishing outside a bounded open subset $\Omega' \subset \Omega$ which has a positive distance from the boundary $\partial\,\Omega$ of Ω.*

We give the so-called Poincaré inequality.

Theorem 1.4. *(Poincaré inequality). Assume that Ω is bounded connected open set of $\mathbb{R}^d$ and that Σ is a (non-empty) Lipschitz continuous subset of the boundary $\partial\Omega$. Then there exists a constant $C_\Omega > 0$ such that*

$$\int_\Omega v^2(x)dx \le C_\Omega \int_\Omega |\nabla v(x)|^2 dx \tag{1.22}$$

for each $v \in \mathrm{H}^1_\Sigma(\Omega)$, where $\mathrm{H}^1_\Sigma(\Omega) := \{v \in \mathrm{H}^1(\Omega) \mid v = 0 \quad on \quad \Sigma\}$.

Under the assumption that $\partial\Omega$ is Lipschitz continuous and using the density of $\mathcal{C}^\infty(\bar{\Omega})$ in $\mathrm{H}^1(\Omega)$ it is proven that the following Green formula holds: for $w, v \in \mathrm{H}^1(\Omega)$,

$$\int_\Omega (D_j w)v dx = -\int_\Omega w D_j v dx + \int_{\partial\Omega} w v n_j d\gamma, j = 1, \cdots, d, \tag{1.23}$$

where D_j denoted the partial derivative $\frac{\partial}{\partial x_j}$. Also, if $\underline{w} \in \mathrm{H}(div; \Omega)$ and $v \in \mathrm{H}^1(\Omega)$, we have

$$\int_\Omega (div\underline{w})v dx = -\int_\Omega \underline{w} \cdot \nabla v dx + \int_{\partial\Omega} (\underline{w} \cdot n)v d\gamma. \tag{1.24}$$

Mathematical formulation and analysis

Assume $\underline{f} \in (\mathrm{L}^2(\Omega))^d$ and set $\Upsilon := \{\underline{v} \in \mathcal{D}(\Omega) \,/\, div\, \underline{v} = 0\}$ and $\mathtt{V} := (\mathrm{H}^1_0(\Omega))^d$ which is a Hilbert space. From the system (1.18)-(1.20) and using Green theorem (1.23), the weak formulation is obtained

$$\alpha(\underline{u}, \underline{v}) + \nu(\nabla\underline{u}, \nabla\underline{v}) - (p, div\underline{v}) = (\underline{f}, \underline{v}) \quad \forall\, \underline{v} \in (\mathcal{D}(\Omega))^d \tag{1.25}$$

and therefore,

$$\alpha(\underline{u}, \underline{v}) + \nu(\nabla\underline{u}, \nabla\underline{v}) = (\underline{f}, \underline{v}) \quad \forall\, \underline{v} \in \Upsilon \tag{1.26}$$

Defining the space

$$\mathtt{V}_{div} := \{\underline{v} \in \mathtt{V} \ / \ div\, \underline{v} = 0\}$$

which is a closed subspace of $\mathtt{V}$ and by virtue of Poincaré inequality [113], it is a Hilbert space for the norm

$$\|\underline{v}\| := \|\nabla \underline{v}\|_0$$

It follows that the bilinear form

$$a(\underline{u}, \underline{v}) := \alpha(\underline{u}, \underline{v}) + \nu(\nabla \underline{u}, \nabla \underline{v}), \quad \underline{u}, \underline{v} \in \mathtt{V}_{div} \tag{1.27}$$

is coercive over $\mathtt{V}_{div} \times \mathtt{V}_{div}$.
On the other hand the application $\underline{v} \to (\underline{f}, \underline{v})$ is linear and continuous over $\mathtt{V}_{div}$. Hence, by Lax-Milgram lemma, the problem

$$\text{find} \quad \underline{u} \in \mathtt{V}_{div} : a(\underline{u}, \underline{v}) = (\underline{f}, \underline{v}) \quad \forall\, \underline{v} \in \mathtt{V}_{div} \tag{1.28}$$

has a unique solution.
The two following results are proved in Girault and Raviart [56] and [113] respectively.

Lemma 1.5. *Let Ω be a bounded domain in $\mathbb{R}^d$ with a Lipschitz continuous boundary and let $\mathtt{L}$ be and element of $\mathtt{V}'$ (i.e, a linear continuous functional on $\mathtt{V}$). Then $\mathtt{L}$ vanishes identically on $\mathtt{V}_{div}$ if and only if there exists a function $p \in \mathrm{L}^2(\Omega)$ such that*

$$\mathtt{L} = (p, div\, \underline{v}) \quad \forall\, \underline{v} \in \mathtt{V} \tag{1.29}$$

moreover, (1.29) defines a unique function p up to an additive constant.

Theorem 1.6. *Let Ω be a bounded domain in $\mathbb{R}^d$ with a Lipschitz continuous boundary and for each $\underline{f} \in (\mathrm{L}^2(\Omega))^d$, let u be the solution to the system (1.18)-(1.20). Then there exists a function $p \in \mathrm{L}^2(\Omega)$ which is unique up to an additive constant such that*

$$a(\underline{u}, \underline{v}) - (p, div\, \underline{v}) = (\underline{f}, \underline{v}) \quad \forall\, \underline{v} \in \mathtt{V} \ .$$

If we define the Hilbert space $\mathtt{Q} = \mathrm{L}^2_0(\Omega)$ with

$$\mathrm{L}^2_0(\Omega) := \{q \in \mathtt{L}^2(\Omega) \ / \ \int_\Omega q = 0\} \tag{1.30}$$

and the bilinear form

$$b(\underline{v}, q) := -(q, div\, \underline{v}), \quad \underline{v} \in \mathtt{V}, q \in \mathtt{Q} \tag{1.31}$$

then, the weak formulation of the Stokes problem becomes

$$\text{find } \underline{u} \in \mathtt{V},\ p \in \mathtt{Q}, \quad a(\underline{u}, \underline{v}) + b(\underline{v}, p) = (\underline{f}, \underline{v}) \quad \forall\, \underline{v} \in \mathtt{V} \tag{1.32}$$

$$b(\underline{u}, q) = 0 \qquad \forall q \in \mathtt{Q} \tag{1.33}$$

Equation (1.33) ensures that $div\, \underline{u} = 0$ almost everywhere as $div\, \underline{u} \in \mathtt{Q}$ and it is orthogonal to all functions of $\mathtt{Q}$.

Remark 1.7. *It is proven in Teman [131] and Glowinksi and Le Tallec [54] that if we consider the Lagrangian functional*

$$\mathfrak{L}(\underline{v}, q) := \frac{1}{2} a(\underline{v}, \underline{v}) + b(\underline{v}, q) - (\underline{f}, \underline{v}), \quad \underline{v} \in \mathtt{V}, q \in \mathtt{Q}$$

then the solution $(\underline{u}, p)$ of the Stokes problem (1.32)-(1.33) is a saddle-point of the above functional i.e

$$\mathfrak{L}(\underline{u}, p) = \min_{\underline{v} \in \mathtt{V}} \max_{q \in \mathtt{Q}} \mathfrak{L}(\underline{v}, q)$$

and conversely.

Galerkin approximation of Stokes problem

We introduce two families of finite dimensional subspaces $\mathtt{V}_h \subset \mathtt{V}$ and $\mathtt{Q}_h \subset \mathtt{Q}$ depending on h. Hence the discrete problem of the system (1.32)-(1.33) is given by:

$$\text{find } \underline{u}_h \in \mathtt{V}_h,\ p_h \in \mathtt{Q}_h : a(\underline{u}_h, \underline{v}_h) + b(\underline{v}_h, p_h) = (\underline{f}, \underline{v}_h) \quad \forall\ \underline{v}_h \in \mathtt{V}_h \tag{1.34}$$

$$b(\underline{u}_h, q_h) = 0 \quad \forall\ q_h \in \mathtt{Q}_h \tag{1.35}$$

Let set $\mathfrak{Z}_h = \{\underline{v}_h \in \mathtt{V}_h : (q_h, \underline{v}_h) = 0\}$ for each $q_h \in \mathtt{Q}_h$ then $\mathfrak{Z}_h$ is the space of discrete divergence-free functions associated with the finite dimensional spaces.
Hence the bilinear form $a(\cdot\,,\cdot)$ is coercive in $\mathfrak{Z}_h$ as it is coercive in $\mathtt{V}$ and $\mathfrak{Z}_h$ is a subspace of $\mathtt{V}$ i.e there exits $C_1 > 0$ such that

$$a(\underline{u}_h, \underline{u}_h) \geq C_1 \|\underline{u}_h\|^2, \quad \forall \quad \underline{u}_h \in \mathfrak{Z}_h$$

The bilinear form $a(\cdot, \cdot)$ is also continuous by using the Poincaré inequality, there exists $C_2 > 0$ such that

$$|a(\underline{u}_h, \underline{v}_h)| \leq C_2 \|\underline{u}_h\| \|\underline{v}_h\| \quad \forall \quad \underline{v}_h \in \mathfrak{Z}_h, q_h \in \mathtt{Q}_h$$

Also the bilinear form $b(\cdot, \cdot)$ is continuous on $\mathtt{V} \times \mathtt{Q}$ i.e

$$|b(\underline{u}_h, q_h)| \leq \delta \|\underline{u}_h\| \|q_h\|$$

Remark 1.8. *The spaces $\mathtt{V}_h$ and $\mathtt{Q}_h$ should enjoy the following compatibility, Inf-Sup or Ladyzhenskaya-Babuška-Brezzi, condition :*
There exits $\beta > 0$ such that

$$\forall q_h \in \mathtt{Q}_h, \exists\ \underline{u}_h \in \mathtt{V}_h : \quad \underline{u}_h \neq 0, \quad (q_h, div\, \underline{u}_h) \geq \beta \|\underline{u}_h\| \|q_h\|$$

i.e

$$\inf_{q_h \in \mathtt{Q}_h} \sup_{\underline{u}_h \in \mathtt{V}_h} \frac{b(\underline{u}_h, q_h)}{\|\underline{u}_h\|_{\mathrm{H}^1(\Omega)} \|q_h\|_{\mathtt{L}^2(\Omega)}} \geq C_4\ .$$

Algebraic interpretation

We denote by N_h and M_h the dimension of $\mathfrak{Z}_h$ and $\mathtt{Q}_h$ respectively and by $\{\underline{\varphi}_j \ / \ j = 1, \cdots, N_h\}$ and $\{\psi_l / l = 1, \cdots M_h\}$ the bases for $\mathfrak{Z}_h$ and $\mathtt{Q}_h$ respectively.
Define $\underline{u}_h = \sum_{j=1}^{N_h} \underline{u}_j \underline{\varphi}_j$, $p_h = \sum_{l=1}^{M_h} p_l \psi_l$ Then the linear system associated to the problem (1.34)-(1.35) becomes

$$\begin{pmatrix} \underline{A} & \underline{B}^T \\ \underline{B} & \underline{O} \end{pmatrix} \begin{pmatrix} \underline{u} \\ p \end{pmatrix} = \begin{pmatrix} \underline{F} \\ \underline{O} \end{pmatrix} \tag{1.36}$$

where $\underline{A} \in \mathbb{R}^{N_h \times N_h}$ is the symmetric and positive definite matrix such that $A_{i,j} = a(\varphi_j, \varphi_i)$, $\underline{B}$ is a rectangular $M_h \times N_h$ matrix with $b_{l,i} := b(\varphi_i, \psi_l)$ and $\underline{f}_i = (\underline{f}, \underline{\varphi}_i)$.
If we set $\underline{S} = \begin{pmatrix} \underline{A} & \underline{B}^T \\ \underline{B} & \underline{O} \end{pmatrix}$ the system $\underline{SX} = \underline{T}$ whose dimension is $N_h + M_h$ can be solve with respect to the Inf-Sup condition since the global matrix $\underline{S}$ is non singular.
The matrix $\underline{A}$ derives from the bilinear form $a(\cdot\,,\cdot)$ hence it is invertible and so we obtain

$$\underline{u} = \underline{A}^{-1}(\underline{F} - \underline{B}^T p) \tag{1.37}$$

If we substitute (1.37) into the second equation of (1.36) we obtain

$$\underline{BA}^{-1}\underline{B}^T p = \underline{BA}^{-1}\underline{F} \tag{1.38}$$

Therefore p exists if and only if the matrix $\underline{BA}^{-1}\underline{B}^T$ is invertible, that means the matrix $\underline{B}$ has maximum rank; but since the Inf-Sup condition is satisfied, then the matrix $\underline{B}$ has maximum rank and so we can determined the pressure. In order words, if the base functions are well chosen, then the matrix $\underline{B}$ has maximum rank and $\underline{S}$ is invertible.

1.6 Operator splittings for a general problem

An exhaustive analysis of the operator splittings methods is formulated in many books and paper like [24, 25, 130, 113, 57] and find application in a wide class of problems.
They are often used to achieve the solution of stationary boundary value problems as steady state of corresponding time dependent problems. This underlines the assumption that the spatial differential operator can be split into a sum of two or more components of simpler structure which are successively integrated in time producing less complicated equations.
Let us consider a problem of the form

$$\frac{d\,\varphi}{d\,t} + \mathcal{L}\varphi \;=\; \psi, \quad t > 0 \tag{1.39}$$

where $\mathcal{L}$ is a matrix that arises from the spatial discretization of a differential operator, accounting also for the boundary conditions, ψ is the corresponding right hand side and φ the unknown solution.
Without loss of generality, we assume that $\mathcal{L}$ is the sum of two components only

$$\mathcal{L} \;=\; \mathcal{L}_1 + \mathcal{L}_2 \tag{1.40}$$

each of them being independent of the time variable. In literature there exists a theory for splitting methods only for the case in which

$$\text{both} \quad \mathcal{L}_1 \quad \text{and} \quad \mathcal{L}_1 \quad \text{are positive (or non-negative) definite .} \tag{1.41}$$

It is important to mention that quite often, the splitting is not operated at the algebraic level as done in (1.40), but rather at the differential stage

$$\mathrm{L} \;=\; \mathrm{L}_1 + \mathrm{L}_2 \tag{1.42}$$

that is the differential operator itself. The operator-splittings of fractional step methods is even more fascinating than the other since it is based on physical considerations : L_1 and L_2 may have specific physical meaning, such as in the case of diffusion and transport processes. The mathematical difficulty however is that (1.42) generally requires a splitting between the boundary conditions as well, in order for endowing both L_1 and L_2 with consistent boundary data.

Now set $t_0 = 0$ and $t_{n+1} = t_n + \Delta\,t$ for $n \geq 0$. We take as φ^0 a convenient approximation of the initial data u_0. Then, the Yanenko splitting is

$$\begin{cases} \dfrac{\varphi^{n+\frac{1}{2}} - \varphi^n}{\tau} + \mathcal{L}_1\varphi^{n+\frac{1}{2}} = 0 \\ \dfrac{\varphi^{n+1} - \varphi^{n+\frac{1}{2}}}{\tau} + \mathcal{L}_2\varphi^{n+1} = \psi^n \end{cases} \tag{1.43}$$

for $n \geq 0$ and $\tau = \Delta\,t$. This method is based on a couple of implicit problems, is first order accurate with respect to $\Delta\,t$ provided that the problem data is sufficiently smooth. In fact, by eliminating $\varphi^{n+\frac{1}{2}}$, we obtained

$$\frac{\varphi^{n+1} - \varphi^n}{\tau} + \mathcal{L}\varphi^{n+1} = \psi^n + \Delta\,t\mathcal{L}_1(\psi^n - \mathcal{L}_2\varphi^{n+1})$$

It has been proven in [89] under restriction (1.41), that the splitting scheme (1.43) is unconditionally stable.

Remark 1.9. *When* L *is the Laplace operator in two dimensions and $\mathcal{L}_i$ is the discrete counterpart of the second derivative $\frac{\partial^2}{\partial\,x_i^2}, i = 1, 2$, the scheme (1.43) produces a classical alternating direction method.*

A three step splitting method is given by

$$\begin{cases} \dfrac{\varphi^{n+\frac{1}{4}}-\varphi^{n}}{\tau}+\mathcal{L}_1\varphi^{n+\frac{1}{4}}=\psi^{n+\frac{1}{2}} \\ \dfrac{\varphi^{n+\frac{1}{2}}-\varphi^{n+\frac{1}{4}}}{\tau}+\mathcal{L}_2\varphi^{n+\frac{1}{2}}=0 \\ \dfrac{\varphi^{n+1}-\varphi^{n}}{2\tau}+\mathcal{L}\varphi^{n+\frac{1}{2}}=\psi^{n+\frac{1}{2}} \end{cases} \tag{1.44}$$

where $\tau=\frac{\Delta t}{2}$ and $\psi^{n+\frac{1}{2}}$ refers to the intermediate time-level $t_{n+\frac{1}{2}}=t_n+\frac{\Delta t}{2}$. This method is well-known as the predictor-corrector method.
It provides a guess $\varphi^{n+\frac{1}{2}}$ which is first order accurate for the solution at $t_{n+\frac{1}{2}}$ through two implicit equations, while the second order corrector produces the new solution φ^{n+1} explicitly.

Remark 1.10. • *This method is second order accurate in time. In fact, by eliminating $\varphi^{n+\frac{1}{4}}$ in the first equation of (1.44) and inserting it in the second equation of (1.44), we obtain*

$$\frac{\varphi^{n+\frac{1}{2}}-\varphi^{n}}{\Delta t}+\frac{1}{2}\mathcal{L}\varphi^{n+\frac{1}{2}}+\frac{\Delta t}{4}\mathcal{L}_1\mathcal{L}_2\varphi^{n+\frac{1}{2}}=\frac{1}{2}\psi^{n+\frac{1}{2}}$$

and using the third equation in (1.44) to express $\mathcal{L}\varphi^{n+\frac{1}{2}}$, it comes

$$\varphi^{n+\frac{1}{2}}=\frac{1}{2}(\varphi^{n+1}-\varphi^{n})-\frac{(\Delta t)^2}{4}\mathcal{L}_1\mathcal{L}_2\varphi^{n+\frac{1}{2}}$$

and as a consequence we obtain the relation

$$\frac{\varphi^{n+1}-\varphi^{n}}{\Delta t}+\mathcal{L}(\frac{\varphi^{n+1}+\varphi^{n}}{2})=\psi^{n+\frac{1}{2}}+\frac{(\Delta t)^2}{4}\mathcal{L}\mathcal{L}_1\mathcal{L}_2\varphi^{n+\frac{1}{2}}$$

but since

$$\begin{aligned}\varphi^{n+\frac{1}{2}} &= (I+\frac{\Delta t}{2}\mathcal{L}_2)^{-1}(I+\frac{\Delta t}{2}\mathcal{L}_1)^{-1})(\varphi^{n}+\frac{\Delta t}{2}\psi^{n+\frac{\Delta t}{2}}) \\ &= \varphi^{n}+O(\Delta t)\end{aligned}$$

we obtain finally

$$\frac{\varphi^{n+1}-\varphi^{n}}{\Delta t}+\mathcal{L}(\frac{\varphi^{n+1}+\varphi^{n}}{2})=\psi^{n+\frac{1}{2}}+\frac{(\Delta t)^2}{4}\mathcal{L}\mathcal{L}_1\mathcal{L}_2\varphi^{n}+O(\Delta t)^3$$

which is the Crank-Nicolson scheme up to $O(\Delta t)^2$.

- *Marchuk proved in [89] that the predictor-corrector scheme is unconditionally stable provided (1.41) holds .*

Another splitting method is that of Peaceman and Rachford given by

$$\begin{cases} \dfrac{\varphi^{n+\frac{1}{2}}-\varphi^{n}}{\tau}+\mathcal{L}_1\varphi^{n+\frac{1}{2}}=\psi^{n+\frac{1}{2}}-\mathcal{L}_2\varphi^{n} \\ \dfrac{\varphi^{n+1}-\varphi^{n+\frac{1}{2}}}{\tau}+\mathcal{L}_2\varphi^{n+1}=\psi^{n+\frac{1}{2}}-\mathcal{L}_1\varphi^{n+\frac{1}{2}} \end{cases} \tag{1.45}$$

with $\tau=\frac{\Delta t}{2}$.

Remark 1.11. • *This scheme is second order accurate since by eliminating $\varphi^{n+\frac{1}{2}}$, the first equation of (1.45) gives*

$$(\frac{1}{\tau}+\mathcal{L}_1)\varphi^{n+\frac{1}{2}}=\frac{1}{\tau}\varphi^n+\psi^{n+\frac{1}{2}}-\mathcal{L}_2\varphi^n \tag{1.46}$$

and the second equation of (1.45) gives

$$(\frac{1}{\tau}+\mathcal{L}_2)\varphi^{n+1}-\psi^{n+\frac{1}{2}}=(\frac{1}{\tau}-\mathcal{L}_1)\varphi^{n+\frac{1}{2}} \tag{1.47}$$

Therefore, by multiplying (1.46) by $\frac{1}{\tau}-\mathcal{L}_1$ and (1.47) by $\frac{1}{\tau}+\mathcal{L}_1$ and summing, it comes

$$\frac{\varphi^{n+1}-\varphi^n}{\Delta t}+\frac{\mathcal{L}}{2}(\varphi^{n+1}+\varphi^n)=\psi^{n+\frac{1}{2}}-\frac{(\Delta t)^2}{4}\mathcal{L}_1\mathcal{L}_2(\frac{\varphi^{n+1}-\varphi^n}{\Delta t})$$

- *If $\mathcal{L}_i$ are suitable approximations of the second order derivative $\frac{\partial^2}{\partial x_i}$, this scheme is unconditionally stable .*

A method similar to the previous one is that of Douglas and Rachford. It is given by

$$\begin{cases} \dfrac{\varphi^{n+\frac{1}{2}}-\varphi^n}{\tau}+\mathcal{L}_1\varphi^{n+\frac{1}{2}}=\psi^n-\mathcal{L}\varphi^n \\ \dfrac{\varphi^{n+1}-\varphi^{n+\frac{1}{2}}}{\tau}+\mathcal{L}_2\varphi^{n+1}=\mathcal{L}_2\varphi^n \end{cases} \tag{1.48}$$

with $\tau=\Delta t$.

Remark 1.12. • *This scheme is only first order accurate in time since by eliminating $\varphi^{n+\frac{1}{2}}$, it comes*

$$\frac{\varphi^{n+1}-\varphi^n}{\Delta t}+\mathcal{L}\varphi^{n+1}=\psi^n-(\Delta t)^2\mathcal{L}_1\mathcal{L}_2(\frac{\varphi^{n+1}-\varphi^n}{\Delta t})$$

- *It is unconditionally stable [89] even when generalized to the case $\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3$, with $\mathcal{L}_i$ a suitable approximation of $\frac{\partial^2}{\partial x_i}$.*

A generalization of the schemes (1.45) and (1.48) is given by the Il'in method

$$\begin{cases} \dfrac{\varphi^{n+\frac{1}{2}}-\varphi^n}{\tau}+\mathcal{L}_1\varphi^{n+\frac{1}{2}}=\psi^{n+\frac{1}{2}}-\mathcal{L}_2\varphi^n \\ \dfrac{\varphi^{n+1}-\varphi^{n+\frac{1}{2}}}{\tau}+\mathcal{L}_2(\varphi^{n+1}-\varphi^n)=\rho\varphi^n\dfrac{\varphi^{n+\frac{1}{2}}-\varphi^n}{\tau} \end{cases} \tag{1.49}$$

with $\tau=\frac{\Delta t}{1+\rho}$ and $\rho\in(-1,1]$ is a parameter.
When $\rho=0$ (or $\rho=1$), (1.49) is the scheme (1.48) with $\psi^{n+\frac{1}{2}}$ instead of ψ^n at the right hand side (and (1.45) respectively).
By eliminating $\varphi^{n+\frac{1}{2}}$ one obtains

$$\frac{\varphi^{n+1}-\varphi^n}{\Delta t}+\mathcal{L}(\frac{1}{1+\rho}\varphi^{n+1}+\frac{\rho}{1+\rho}\varphi^n)=\psi^{n+\frac{1}{2}}-\frac{(\Delta t)^2}{(1+\rho)^2}\mathcal{L}_1\mathcal{L}_2(\frac{\varphi^{n+1}-\varphi^n}{\Delta t}) \ .$$

Remark 1.13. *It is of second order accurate for $\rho=1$ (in this case it coincides with the Peaceman-Rachford scheme) and first other in the other cases.*

In conclusion, different kinds of operator splitting can be applied to a given problem, but for an optimal choice, it should take into account the nature of the physical processes involved in the whole problem.

1.7 Fractional step methods for Navier-Stokes problem

Let us now consider the time dependent Navier-Stokes problem that reads

$$\begin{cases} \dfrac{\partial \underline{u}}{\partial t} - \nu \Delta \underline{u} + (\underline{u} \cdot \nabla)\, \underline{u} + \nabla p = \underline{f} & \text{in} \quad Q_T := (0,T) \times \Omega \\ div\, \underline{u} = 0 & \text{in} \quad Q_T := (0,T) \times \Omega \\ \underline{u} = 0 & \text{on} \quad \Sigma_T := (0,T) \times \partial\Omega \\ \underline{u}\,|_{t=0} = \underline{u}_0 & \text{in} \quad \Omega \end{cases} \tag{1.50}$$

where $\underline{f} = f(t, \underline{x})$ and Ω is an open bounded domain of $\mathbb{R}^2$ (with d= 2) and $\partial\,\Omega$ is its boundary. For $\underline{w}, \underline{z}, \underline{v} \in V = (\mathrm{H}_0^1(\Omega))^2$, we define

$$a_j(\underline{w}, \underline{v}) := \nu \int_\Omega D_j \underline{w} D_j \underline{v}$$

$$\tilde{c}_j(\underline{w}; \underline{z}, \underline{v}) := \frac{1}{2} \sum_{i=1}^{2} \int_\Omega \underline{w}_j [\underline{v}_j D_j \underline{z}_j - \underline{z}_i D_j \underline{v}_i]$$

where $j = 1, 2$. We set $(\underline{u} \cdot \nabla)\, \underline{u} := \sum_{i=1}^{2} D_i \underline{u}$. The external force field is also split as $\underline{f} = \sum_{j=1}^{2} \underline{f}_j$, ν is the kinematic viscosity μ/ρ .
Therefore, a three step method consists for each $n = 0, 1, \cdots \mathcal{N} - 1$:

- Find $\underline{u}_h^{n+\frac{1}{3}} \in V_h$ such that

$$\frac{1}{\Delta t}(\underline{u}_h^{n+\frac{1}{3}} - \underline{u}_h^n, \underline{v}_h) + a_1(\underline{u}_h^{n+\frac{1}{3}}, \underline{v}_h) + \tilde{c}_1(\underline{u}_h^n; \underline{u}_h^{n+\frac{1}{3}}, \underline{v}_h) = (\underline{f}_{1,*}^{n+\frac{1}{3}}, \underline{v}_h) \quad \forall\, \underline{v}_h \in V_h \tag{1.51}$$

  where $(\cdot\,, \cdot\,)$ is the scalar product in $(\mathrm{L}^2(\Omega))^2$ and for each vector $\underline{g} \in (\mathrm{L}^2(Q_T)^2$ we set

$$\underline{g}_*^{n+\frac{1}{2}} := \frac{1}{\Delta t} \int_{t_n}^{t^{n+1}} \underline{g}(t)\, dt$$

- Then find $\underline{u}_h^{n+\frac{2}{3}} \in V_h$ such that

$$\frac{1}{\Delta t}(\underline{u}_h^{n+\frac{2}{3}} - \underline{u}_h^{n+\frac{1}{3}}, \underline{v}_h) + a_2(\underline{u}_h^{n+\frac{2}{3}}, \underline{v}_h) + \tilde{c}_2(\underline{u}_h^{n+\frac{1}{3}}; \underline{u}_h^{n+\frac{2}{3}}, \underline{v}_h) = (\underline{f}_{2,*}^{n+\frac{1}{3}}, \underline{v}_h) \quad \forall\, \underline{v}_h \in V_h \tag{1.52}$$

- Then find the unique solution $\underline{u}_h^{n+1} \in W_h$ to the problem by

$$(\underline{u}_h^{n+1}, \underline{w}_h) = (\underline{u}_h^{n+\frac{2}{3}}, \underline{w}_h) \quad \forall\, \underline{w}_h \in W_h \tag{1.53}$$

  where W_h is a subspace of V_h which constitutes a suitable approximation of V_{div}.

The existence and uniqueness for both (1.51) and (1.52) follow positiveness since $\tilde{c}_j(\underline{z}; \underline{v}, \underline{v}) = 0$ for each $\underline{z}, \underline{v} \in V$, $j = 1, 2$. Equation (1.53) states that $\underline{u}_h^{n+1}$ is the L^2-orthogonal projection of $\underline{u}_h^{n+\frac{2}{3}}$ onto W_h that is

$$\underline{u}_h^{n+1} = P_h \underline{u}_h^{n+\frac{2}{3}}$$

with $P_h : V_h \to W_h$ the orthogonal projection operator with respect to the scalar product of $(\mathrm{L}^2(\Omega))^2$. The fractional step (1.51)-(1.53) is well known as the projection method [113].

Remark 1.14. • *This scheme is unconditionally stable in the norm of* $(\mathrm{L}^2(\Omega))^2$

- *If the time step $\Delta t = O(h^2)$ holds, one has*

$$\Delta t \sum_{n=0}^{\mathcal{N}-1} \|\underline{u}_h^{n+1}\|_1^2 \leq C \tag{1.54}$$

- *When (1.54) holds, it is proved in [130, 131] that the scheme is convergent .*

Another approach [113] that makes use of two (rather than three) steps (and written in its continuous version) consists to define two sequences of vector functions $\underline{u}^{n+\frac{1}{2}}, \underline{u}^{n+1}$ and a sequence of scalar function q^{n+1} recursively given as:
$\underline{u}^0 := \underline{u}_0$ and for $n = 0, 1, \cdots \mathcal{N} - 1$

$$\begin{cases} \dfrac{1}{\Delta t}(\underline{u}^{n+\frac{1}{2}} - \underline{u}^n) - \nu \Delta \underline{u}^{n+\frac{1}{2}} + (\underline{u}^{n+\frac{1}{2}} \cdot \nabla)\underline{u}^{n+\frac{1}{2}} + \dfrac{1}{2}(div\, \underline{u}^{n+\frac{1}{2}})\underline{u}^{n+\frac{1}{2}} = \underline{f}_*^{n+\frac{1}{2}} \quad \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \underline{u}^{n+\frac{1}{2}} \cdot n = 0 \quad \text{on} \quad \partial\Omega \end{cases} \tag{1.55}$$

$$\begin{cases} \dfrac{1}{\Delta t}(\underline{u}^{n+1} - \underline{u}^{n+\frac{1}{2}}) + \nabla q^{n+1} = 0 \quad \text{in} \quad \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad div\, \underline{u}^{n+1} = 0 \quad \text{in} \quad \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad \underline{u}^{n+1} \cdot n = 0 \quad \text{on} \quad \partial\Omega \end{cases} \tag{1.56}$$

where n is the unit outward normal vector on $\partial\Omega$. The solution of this problem based on least squares approach has been made in [50].
The existence of a scalar q^{n+1} is a consequence of the so-called Helmholtz decomposition principle. It states that any function $\underline{v} \in (\mathrm{L}^2(\Omega))^2$ can be uniquely represented as $\underline{v} = \underline{w} + \nabla q$, where $\underline{w} \in \mathrm{H}_{div}$ and $q \in \mathrm{H}^1(\Omega)$. But $\int_\Omega \underline{z} \cdot \nabla q = 0$ for each $\underline{z} \in \mathrm{H}_{div}$ and therefore $\underline{w} = P_{div}\, \underline{v}$, where P_{div} is the orthogonal projection operator from $(\mathrm{L}^2(\Omega))^2$ onto H_{div}. Hence since $\underline{w} \in \mathrm{H}_{div}$, q turns out to be the solution of the Neumann problem

$$\begin{cases} \Delta q = div\, \underline{u} \quad \text{in} \quad \Omega \\ \dfrac{\partial q}{\partial n} = \underline{u} \cdot n \quad \text{on} \quad \partial\Omega \end{cases}$$

which defines q up to an additive constant.
It is proven in [130] that the scalar function $q^{n+1}(\underline{x})$ does approximate the pressure $p(t_{n+1}, \underline{x})$ in a weak sense.
Therefore from (1.55)-(1.56) we have

$$\frac{\partial q^{n+1}}{\partial n} = \nabla q^{n+1} \cdot n = 0 \quad \text{on} \quad \partial\Omega \tag{1.57}$$

since both $\underline{u}^{n+1} \cdot n$ and $\underline{u}^{n+\frac{1}{2}} \cdot n$ vanish on $\partial\Omega$, therefore it comes

$$\Delta q^{n+1} = \frac{1}{\Delta t} div\, \underline{u}^{n+\frac{1}{2}} \tag{1.58}$$

and

$$\frac{\partial q^{n+1}}{\partial n} = 0 \quad \text{on} \quad \partial\Omega.$$

The function q^{n+1} satisfies a Poisson problem with homogeneous Neumann boundary condition.

Remark 1.15. *This method represents correctly the velocity field in many field problems of physical interest [53].*

For $n = 0, 1, \cdots, \mathcal{N}-1$, the scheme can be rewritten in the equivalent form

$$\begin{cases} \frac{1}{\Delta t}(\underline{u}^{n+\frac{1}{2}} - \underline{u}^{n-\frac{1}{2}}) - \nu\Delta\,\underline{u}^{n+\frac{1}{2}} + (\underline{u}^{n+\frac{1}{2}}\cdot\nabla)\underline{u}^{n+\frac{1}{2}} + \frac{1}{2}(div\,\underline{u}^{n+\frac{1}{2}})\underline{u}^{n+\frac{1}{2}} + \nabla\, q^n = \underline{f}_*^{n+\frac{1}{2}} \quad \text{in } \Omega \\ \qquad\qquad \underline{u}^{n+\frac{1}{2}} = 0 \quad \text{on} \quad \partial\,\Omega \end{cases} \tag{1.59}$$

$$\begin{cases} \Delta\, q^{n+1} = \frac{1}{\Delta\, t} div\,\underline{u}^{n+\frac{1}{2}} \\ \frac{\partial\, q^{n+1}}{\partial\, n} = 0 \quad \text{on} \quad \partial\,\Omega \end{cases} \tag{1.60}$$

providing the initializations $\underline{u}^{-\frac{1}{2}} := \underline{u}_0$ and $q^0 = 0$.

Remark 1.16. *The velocity and the pressure of the scheme (1.59)-(1.60) is proven to be convergent [113] at the first order with respect to $\Delta\, t$ in the norms $(\mathrm{L}^2(\Omega))^2$ and of dual space of $\mathrm{H}^1(\Omega)\cap\mathrm{L}^2_0(\Omega)$ respectively. Moreover, with respect to the norms of $(\mathrm{H}^1(\Omega))^2$ and $\mathrm{L}^2(\Omega)$, the convergence is at the order of $(\Delta\, t)^{\frac{1}{2}}$.*

Other fractional step methods for the Navier-Stokes equations are formulated in many papers and books like [50, 115, 119].

1.8 Advancing time to be used

We will pay attention to the nonlinear convective term of the momentum equation of the problem (1.9)-(1.11). In fact we will use the semi-Lagrangian method that computes the values of the variables of interest at the foot of the trajectory.
Assume for instance that, we have the equation

$$\frac{D(F)}{Dt} = \Psi \tag{1.61}$$

where $\frac{D}{Dt}(\cdot)$ represents the total derivative, then a semi-Lagrangian off-centered scheme is :

$$F^{n+1} - F^n_{tr} = \Delta t(\alpha\Psi^{n+1} + (1-\alpha)\Psi^n_{tr}) \quad \text{with} \quad 0 \le \alpha \le 1 \tag{1.62}$$

where the foot index tr shows that the relevant variable is calculated at the foot of the trajectory by an interpolation process. The application of a similar approximation to the momentum equation (1.9) - (1.10) gives, with $\alpha = 1$

$$\underline{u}^{n+1} - \underline{u}^n(X_{tr}) = \Delta t\{\nu\Delta\underline{u} - \nabla\, p + \underline{f}\}^{n+1} \tag{1.63}$$

Therefore, by the following approach, the difficulties inherent to the non linearity are overcome and the stability is granted under mild conditions since the method of characteristics is unconditionally stable. We will review the characteristics method latter in the thesis. Now the algorithm is organized as follows:

- A provisional velocity $\underline{\tilde{u}}$ is computed by the momentum equation neglecting the divergence free constraint and considering a guessed pressure p^n and updated boundary conditions BC^{n+1} :
$$\underline{\tilde{u}} - \Delta t\Delta\underline{\tilde{u}} = \underline{u}^n(X_{tr}) + \Delta t(-\nabla p^n + \underline{f}^{n+1}) \bigoplus BC^{n+1} \tag{1.64}$$
- A Poisson equation for the pressure correction $\tilde{p}$ is written taking into account incompressibility and applying the divergence operator to the provisional velocity $\underline{\tilde{u}}$
$$-\Delta\tilde{p} = -\frac{1}{\Delta t}\nabla\cdot\underline{\tilde{u}} \tag{1.65}$$

- The adjourned values of the velocity and pressure are obtained updating the provisional values:

$$\underline{u}^{n+1} = \underline{\tilde{u}} - \Delta t \nabla \tilde{p} \quad , \quad p^{n+1} = p^n + \tilde{p} \ . \tag{1.66}$$

Since the incompressibility has been imposed at the second step, the updated velocity field $\underline{u}^{n+1}$ is divergence free.
Therefore, the most expensive computational kernels are reduced to elliptic problems. Hence, the necessity to develop a spatial approximation able to solve effectively elliptic problems and this will be the aim of the next chapter. We conclude this chapter by reviewing classical spatial discretizations that are the classical finite difference method and the finite volume method.

1.9 Spatial discretization

Previously, we recalled some approximations of the derivatives of a function by finite difference schemes (e.g. in the approximation of the temporal derivative) in the following, a brief introduction of two classical numerical methods used in computational fluid dynamics will be presented.

1.9.1 Classical finite difference (FD) method

A problem which arises often in numerical analysis is the approximation of the derivative of a function $f(x)$ on a given interval $[a, b]$.
An approach to it consists of introducing in $[a, b]$ $n+1$ nodes $\{x_k, k = 0, \cdots, n\}$, with $x_0 = a$, $x_n = b$ and $x_{k+1} = x_k + h$, $k = 0, \cdots, n-1$ where $h = (b-a)/n$. Then $f'(x_i)$ is approximated using the nodal values $f(x_k)$ as

$$f'(x_i) \simeq \sum_{k=-m'}^{m'} \beta_k f(x_{i-k}) \tag{1.67}$$

where $\{\beta_k\} \in \mathbb{R}$ are $m + m' + 1$ coefficients to be determined.An issue of the choice of the scheme (1.67) is the computational efficiency. In fact, if $m \neq 0$, determining the values $\{u_i\}$ requires the solution of a linear system.
The set of nodes which are involved in constructing the derivative of f at a certain node, is called the stencil. The band of the matrix associated with the system (1.67) will increase as the stencil gets larger.
In order to generate a formula like (1.67), the basic idea consists of resorting to the definition of the derivative. Therefore if $f'(x)$ exists, then

$$f'(x_i) = \lim_{h \to 0^+} \frac{f(x_i + h) - f(x_i)}{h} \ . \tag{1.68}$$

By replacing the limit with the finite incremental ratio, we obtain the approximation

$$u_i'^{FD} = \frac{f(x_{i+1}) - f(x_i)}{h}, \quad 0 \leq i \leq n-1 \ . \tag{1.69}$$

The relation (1.69) is a special case of (1.67) by setting $m = 0$, $\beta_{-1} = 1$, $\beta_0 = -1$, $\beta_1 = 0$.
The right side of the (1.69) is well known as the forward finite difference and the approximation that is being used corresponds to replacing $f'(x_i)$ with the slope of the straight line passing through the point $(x_i, f(x_i))$ and $(x_{i+1}, f(x_{i+1}))$ as shown in Figure 1.2. To estimate the error of the approximation, it is sufficient to expand the Taylor's series obtaining

$$f(x_{i+1}) = f(x_i) + h f'(x_i) + \frac{h^2}{2} f''(\xi_i) \quad \text{with} \quad \xi_i \in (x_i, x_{i+1}) \ .$$

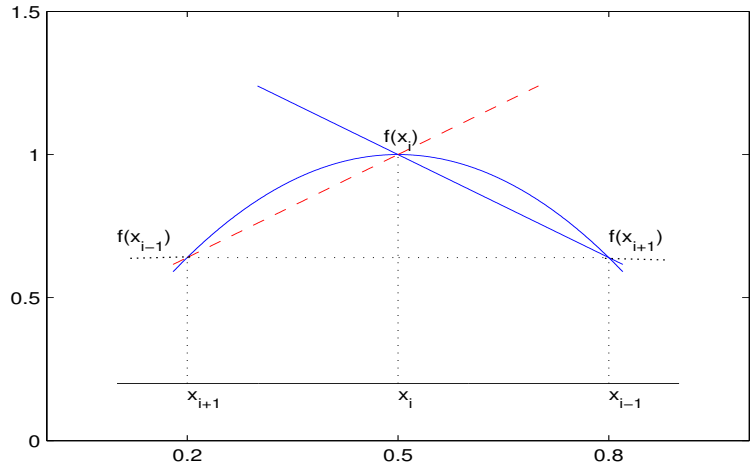

Figure 1.2: Finite difference approximation of $f'(x_i)$: backward (solid line), forward (dashed line) and centered (pointed line).

If we assume that f is regular, then

$$f'(x_i) - u_i'^{FD} = -\frac{h}{2} f''(\xi_i) \ . \tag{1.70}$$

Instead of (1.69), we can employ a centered incremental ratio, yielding the following approximation

$$u_i'^{CD} = \frac{f(x_{i+1}) - f(x_{i-1})}{2h}, \quad 1 \leq i \leq n-1 \ . \tag{1.71}$$

The scheme (1.71) is a special case of (1.67) by setting $m' = 1$, $\beta_{-1} = \frac{1}{2}$, $\beta_0 = 0$, $\beta_1 = -\frac{1}{2}$. The right side of (1.71) is called the centered finite difference and geometrically amounts to replacing $f'(x_i)$ with the slope of the straight line passing through the points $(x_{i-1}, f(x_{i-1}))$ and $(x_{i+1}, f(x_{i+1}))$ see Figure 1.2. By using the Taylor expansion, we get

$$f'(x_i) - u_i'^{CD} = -\frac{h^2}{6} f'''(\xi_i) \ . \tag{1.72}$$

Hence the formula (1.71) provides a second -order approximation to $f'(x)$ with respect to h. Using a similar procedure, we derive a backward finite difference scheme

$$u_i'^{BD} = \frac{f(x_i) - f(x_{i-1})}{h}, \quad 1 \leq i \leq n, \tag{1.73}$$

which is affected by the following error

$$f'(x_i) - u_i'^{BD} = \frac{h}{2} f''(\xi_i) \tag{1.74}$$

so that the parameters in (1.67) are $m' = 1$, $\beta_{-1} = 0$, $\beta_0 = 1$ and $\beta_1 = -1$.
Higher order schemes as well as finite difference approximations of higher order derivatives of f can be constructed using suitable Taylor's expansions.
If $f \in \mathcal{C}^4([a,b])$, the approximation of f'' is given by

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} - \frac{h^2}{24}(f^{(4)}(x_i + \theta_i h) + f^{(4)}(x_i + \omega_i h)), \quad 0 < \theta_i, \omega_i < 1 \tag{1.75}$$

Therefore, the following centered finite difference scheme can be derived

$$u_i'' = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2}, \quad 1 \leq i \leq n-1 \tag{1.76}$$

whose error is given by

$$f''(x_i) - u_i'' = -\frac{h^2}{24}(f^{(4)}(x_i + \theta_i h) + f^{(4)}(x_i + \omega_i h)) \quad , \tag{1.77}$$

thus the formula (1.76) provides a second-order approximation to $f''(x_i)$ with respect to h. In reality, the existence of $f^{(4)}$ is necessary to have an estimation of the error and not to build the finite difference formula for f''.

1.9.2 Finite volume (FV) method

The term finite volume method was invented by Jameson [65] for the discretization of the full potential equation of compressible gas dynamics, $\nabla\cdot(\rho\nabla\phi) = 0$ where $\rho = \rho(|\nabla\phi|)$ is given by Bernouilli's equation. However, the actual method was in use earlier [90] for this problem and similar schemes were widely used much earlier for modeling neutron diffusion. The basic idea is to exploit the divergence form of the equation by integrating it over a finite volume and using Gauss's theorem to convert the result into surface integral which is then discretized. In general the finite volume can be distinguished by the following criteria:

1) the geometric shape of the finite volume,

2) the position of the unknowns ("problem variables") with respect to the finite volume,

3) the approximation of the boundary (line in 2D) or surface (in 3D) integrals.

Especially, the second criterion divides the finite volume into two larges classes: the first class is the cell-centered finite volume approach pioneered by Jameson and his co-workers [68]. Here the mesh values of the unknowns are associated with the center of the cells so that to calculate the fluxes integral on the edges involves values of the unknown function in points belonging to the volumes closed to the referred volume. The second class is the cell-vertex finite volume approach associated to the name of Ni [100], the unknowns are held at the vertices, thus the trapezium rule is used along the edge to calculate the fluxes integral, see Figure 1.3. In order to apply the finite volume technique a partition of the domain Ω is necessary. There are two possibilities. The difference between them is whether the problem variables are assigned to the finite volume or, fixed some points for the problem variables, associated control volumes are then defined. Actually, there is another class which consists to associate average normal and tangential fluxes with each edge, but extra consistency relations are needed to generate a sufficient number of equations and this approach has not so far formed as practical solutions.

Assets of finite volume methods

- Flexibility with respect to the geometry of the domain Ω, eventually applying a numerical transformation to the physical domain and to the analytical equations
- Admissibility of unstructured grids (important for adaptive methods)
- Easy assembling when constructing the algebraic system
- Respect of the conservation principle for what concerns some important physical variables (like energy for instance). This is important in the numerical solution of differential problems with discontinuous coefficients or in the solution of convection-diffusion problems dominated by the convective term

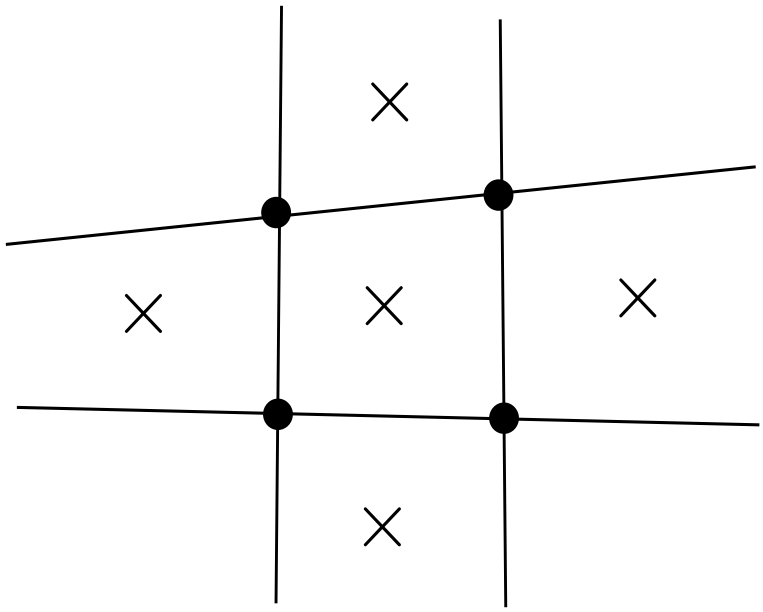

Figure 1.3: The finite volume cell, with the cell-centered × and the cell-vertex • mesh points.

- Easy linearization of nonlinear problems (Newton's method)
- Simple discretization of boundary conditions especially a "natural" treatment of Neumann or mixed boundary conditions
- No restriction of the spatial dimension d of the domain Ω.

Drawbacks of the finite volume method

- Difficulties in the design of higher order methods
- Possible complex task in the construction of some classes or type of control volumes (in particular for domains variating in time)
- Difficulties in the mathematical analysis (stability, convergence,$\cdots$).

Chapter 2

Solution of elliptic problems

2.1 The finite element (FE) method for the Poisson equation

In engineering and scientific computing, the finite element method, beginning in 1906, is one of the most important and powerful computational tools for numerical solutions of partial differential equations. It not only provides a systematic procedure for computer implementations, but also shows versatility for dealing with problems with complex geometrical shape, loads and boundary conditions. Because of the outstanding features of the finite element method and the advent of high-speed computers, the finite element method has become more and more important for numerical solution of partial differential equations by which most engineering and scientific problems are defined. The essential concept of the finite element method is first to assume an approximation of the unknown function in an appropriate $N-$dimensional space such as $\phi = \sum_{i=1}^{N} \Lambda_i B_i$ where $\Lambda_i (i = 1, 2, \cdots, N)$ are coefficients to be determined and $B_i (i = 1, 2, \cdots, N)$ are the basis functions (or shape functions) and to use an approximation method, such as the weighted residual method or the Galerkin method to determine the set $\{\Lambda_i\}_1^N$ such that a residual with respect to the approximation can be reduced to an acceptable tolerance. The common feature of the existing methods for determining $\{\Lambda_j\}_1^N$ leads to an algebraic system which can be solved by a suitable solver (direct or iterative). In what follows, we are going to review known results in finite elements.
Let Ω be a bounded domain in $\mathbb{R}^2$ with a Lipschitz continuous boundary $\partial\Omega$ and the outward normal n to the boundary. Without loss of generality, we consider the following homogeneous Dirichlet problem for the Poisson equation:
Given $f \in \mathrm{L}^2(\Omega)$, find $u \in \mathrm{H}^2(\Omega)$ such that

$$-\Delta u = f \quad \text{in} \quad \Omega \tag{2.1}$$

$$u = 0 \quad \text{on} \quad \partial\Omega \tag{2.2}$$

In order to study the problem (2.1)-(2.2), we have to give its weak formulation, that is the formulation in the sense of distributions. Let us consider an arbitrary test function v. We multiply equation (2.1) by v, integrate the result over Ω, and obtain

$$\begin{aligned} (f, v) &= \int_\Omega fv \\ &= -\int_\Omega \nabla\cdot(\nabla u)v = \int_\Omega \nabla u\cdot\nabla v - \int_{\partial\Omega} \nabla u\cdot nv = \int_\Omega \nabla u\cdot\nabla v \end{aligned} \tag{2.3}$$

where $(\cdot\,,\cdot)$ is the scalar product in $\mathrm{L}^2(\Omega)$. The equality sign at the second line of (2.3) is obtained by integration by parts using the integral theorem of Gauss. The boundary integral vanishes because $v = 0$ holds on $\partial\Omega$. Therefore, the weak formulation consists in finding $u \in \mathrm{V} = \mathrm{H}_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \qquad \forall v \in \mathrm{V} \tag{2.4}$$

with $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$.

Existence of the weak solution

The space $\mathrm{H}_0^1(\Omega)$ is a closed subspace of the Hilbert space $\mathrm{H}^1(\Omega)$, hence $\mathrm{H}_0^1(\Omega)$ is also a Hilbert space.
The application $v \mapsto \|\nabla v\|_{\mathrm{L}^2(\Omega)}$ is a norm in $\mathrm{H}_0^1(\Omega)$. In fact, $\|\nabla v\| = 0 \Rightarrow \nabla u = 0$ that is $v = C$, where C is constant but since $v \in \mathrm{H}_0^1(\Omega)$, it comes that $C = 0$ and thus $v = 0$. We have also this Poincaré inequality:

$$\text{there exists} \quad C > 0 \quad \text{such that} \quad \|v\|_{\mathrm{L}^2(\Omega)} \leq C\|\nabla v\|_{\mathrm{L}^2(\Omega)} \quad \forall v \in \mathrm{L}^2(\Omega) \tag{2.5}$$

Proof. We give the proof of the Poincaré inequality (2.5). By contradiction, assume that there exists a sequence (u_n) in $\mathrm{H}_0^1(\Omega)$ such that

$$\frac{1}{n}\|u_n\|_{\mathrm{L}^2(\Omega)} \geq \|\nabla u_n\|_{\mathrm{L}^2(\Omega)} \quad \forall\ n \geq 1 \tag{2.6}$$

Taking $v_n = \frac{u_n}{\|u_n\|_{\mathrm{L}^2(\Omega)}}$, we have that $v_n \in \mathrm{H}_0^1(\Omega)$ and $\|v_n\|_{\mathrm{L}^2(\Omega)} = 1$ and $\|v_n\|_{\mathrm{H}^1(\Omega)} \leq 1 + \frac{1}{n} \leq 2$. Hence, by Rellick-Kondrachov theorem and the fact that $\mathrm{H}_0^1(\Omega)$ is a closed subset of $\mathrm{H}^1(\Omega)$, there exists a subsequence of v_n, called again v_n and a function $v \in \mathrm{H}_0^1(\Omega)$ such that $v_n \to v$ in $\mathrm{L}^2(\Omega)$ and $\nabla v_n \to 0$ in $\mathcal{D}'(\Omega)$.
On the other hand, $v_n \to v$ in $\mathrm{L}^2(\Omega)$ implies $v_n \to v$ in $\mathcal{D}'(\Omega)$, which implies also $\nabla v_n \to \nabla v$ in $\mathcal{D}'(\Omega)$ since the derivative operator ∇ is continuous from $\mathcal{D}'(\Omega)$ to $\mathcal{D}'(\Omega)$. By uniqueness of the limit, it follows that $\nabla v = 0$, therefore $v = C$ since Ω is connected. Therefore, since $v \in \mathrm{H}_0^1(\Omega)$ lead to $v = 0$. Finally, we have $1 \leq \|v_n\|_{\mathrm{H}(\Omega)} \to 0$ as $n \to \infty$ which is a contradiction. Hence the result i.e $\|u\|_{\mathrm{H}_0^1(\Omega)} = \|\nabla u\|_{\mathrm{L}^2(\Omega)}$. □

Therefore, the bilinear form $a(\cdot\,,\cdot\,)$ is continuous in $\mathrm{H}_0^1(\Omega) \times \mathrm{H}_0^1(\Omega)$ with respect to the norm $\|\cdot\|_{\mathrm{H}_0^1(\Omega)} = \|\nabla\|_{\mathrm{L}^2(\Omega)}$. In fact

$$|a(u,v)| = |\int_\Omega \nabla u \cdot \nabla v| \leq \|\nabla u\|_{\mathrm{L}^2(\Omega)} \cdot \|\nabla u\|_{\mathrm{L}^2(\Omega)} \leq M\|u\|_{\mathrm{H}_0^1(\Omega)}\|v\|_{\mathrm{H}_0^1(\Omega)} \quad \forall u, v \in \mathrm{H}_0^1(\Omega)$$

i.e

$$M := \text{constant}: \quad |a(u,v)| \leq M\|u\|_{\mathrm{H}_0^1(\Omega)}\|v\|_{\mathrm{H}_0^1(\Omega)} \quad \forall u, v \in \mathrm{H}_0^1(\Omega) \tag{2.7}$$

The bilinear form $a(\cdot\,,\cdot\,)$ is coercive. In fact

$$a(u,u) = \int_\Omega \nabla u \cdot \nabla u \geq \|\nabla u\|^2_{\mathrm{L}^2(\Omega)} \geq \alpha\|u\|^2_{\mathrm{H}_0^1(\Omega)} \quad \forall u \in \mathrm{H}_0^1(\Omega)$$

i.e

$$a(u,u) \geq \alpha\|u\|^2_{\mathrm{H}_0^1(\Omega)} \tag{2.8}$$

The linear functional $v \mapsto \int_\Omega fv$ is continuous since

$$|\int_\Omega fv| \leq \|f\|_{\mathrm{L}^2(\Omega)}\|v\|_{\mathrm{L}^2(\Omega)} \leq C\|f\|_{\mathrm{L}^2(\Omega)}\|v\|_{\mathrm{H}_0^1(\Omega)} \leq C'\|v\|_{\mathrm{H}_0^1(\Omega)}$$

i.e

$$|\int_\Omega fv| \leq C'\|v\|_{\mathrm{H}_0^1(\Omega)}$$

We conclude by Lax-Milgram lemma that there exists a solution to the variational problem (2.4).

Uniqueness of the weak solution

Lemma 2.1. *The weak solution to the weak problem (2.4) is unique.*

Proof. Let u_1, u_2 be two weak solutions, i.e

$$a(u_1, v) = \int_\Omega fv, \quad \forall \quad v \in \mathrm{V}$$
$$a(u_2, v) = \int_\Omega fv, \quad \forall \quad v \in \mathrm{V}$$

by subtraction, it follows that

$$a(u_1 - u_2, v) = 0 \quad \text{for all} \quad v \in \mathrm{V}$$

choosing $v = u_1 - u_2$ implies $a(u_1 - u_2, u_1 - u_2) = 0$ and consequently $u_1 = u_2$ because $a(.,.)$ is positive definite. □

2.1.1 The discrete problem

Let $\mathcal{I}_h$ a partition of Ω into closed triangles K (that is including the boundary ∂K) with the following properties:

- $\bar{\Omega} = \cup_{K \in \mathcal{I}_h} K$
- For $K, K' \in \mathcal{I}_h, K \neq K'$
$$int(K) \cap int(K') = \emptyset$$
where $int(K)$ denotes the open triangle (without the boundary ∂K)
- If $K \neq K'$ but $K \cap K' \neq \emptyset$, then $K \cap K'$ is either a point or an edge common to K and K'.

We define the finite element space:

$$\mathrm{V}_h := \{v \in \mathrm{V} : v|_K \in \mathbb{P}_k(K) \quad \forall\, K \in \mathcal{I}_h\} \subset \mathrm{V} = \mathrm{H}_0^1(\Omega) \tag{2.9}$$

where $\mathbb{P}_k$ are polynomials of degree $\leq k$.
The discrete problem consists in:

$$\text{finding} \quad u_h \in \mathrm{V}_h : \int_\Omega \nabla u_h \cdot \nabla v = \int_\Omega fv \quad \forall v_h \in \mathrm{V}_h$$

or in a more general setting

$$\text{find} \quad u_h \in \mathrm{V}_h : a(u_h, v) = (f, v) \quad \forall v \in \mathrm{V}_h \ . \tag{2.10}$$

Convergence property

The study of the convergence requires to consider the variational formulation (2.4) and the discrete problem (2.10). Let $a(.,.)$ be the bilinear form. It is symmetric and, if we denotes $e := u - u_h$ the error, then the important relation

$$a(e, v) = 0 \quad \text{for all} \quad v \in \mathrm{V}_h \tag{2.11}$$

is satisfied. To obtain this relation it is sufficient to consider (2.4) for $v \in \mathrm{V}_h \subset \mathrm{V}$ and then to subtract the result from the discrete problem (2.10).
Since $a(.,.)$ is symmetric and positive definite, i.e $a(u, v) = a(v, u)$, $a(u, u) \geq 0$, $a(u, u) = 0 \Longleftrightarrow u = 0$ then the error is orthogonal to the space V_h with respect to the scalar product a.
The relation (2.11) is often called the orthogonality of the error, in fact the element $u_h \in \mathrm{V}_h$ with minimal distance to $u \in \mathrm{V}$ with respect to the induced norm $\|\cdot\|_a$ is characterized by (2.11).

Lemma 2.2. *The discrete solution u_h according to (2.10) is stable in the following sense:*

$$\|u_h\| \leq \frac{1}{\alpha}\|f\| \quad \textit{independently of } h, \textit{ where} \quad \|f\| := \sup\{\ \frac{(f,v)}{\|v\|} \quad | \quad v \in \mathrm{V},\ v \neq 0\ \}\,. \quad (2.12)$$

Proof. If $u_h = 0$, there is nothing to prove. If $u_h \neq 0$, from $a(u_h, v) = (f, v)$ for all $v \in \mathrm{V}_h$, it follows that

$$\alpha\|u_h\|^2 \leq a(u_h, u_h) = (f, u_h) \leq \frac{|(f, u_h)|}{\|u_h\|}\|u_h\| \leq \|f\|\|u_h\|$$

Then, dividing this relation by $\alpha\|u_h\|$, we get the desired result. We note also that the approximation property (2.12) holds up to a constant . □

We give also Céa's lemma.

Lemma 2.3. *Assume (2.7)-(2.8), then the following error estimate for the discrete problem holds:*

$$\|u - u_h\| \leq \frac{M}{\alpha}\min\{\ \|u - v\| \quad | \quad v \in \mathrm{V}_h\ \}\,.$$

Proof. If $\|u - u_h\| = 0$, then the assertion is true. Otherwise, let $v \in \mathrm{V}_h$ be arbitrary, by using error equation (2.11) and $u_h - v \in \mathrm{V}_h$, $a(u - u_h, u_h - v) = 0$.
Therefore, using (2.8) we have

$$\begin{aligned}\alpha\|u - u_h\|^2 \leq a(u - u_h, u - u_h) &= a(u - u_h, u - u_h) + a(u - u_h, u_h - v)\\ &= a(u - u_h, u - v)\end{aligned}$$

Furthermore, by means of (2.7), we obtain

$$\alpha\|u - u_h\|^2 \leq a(u - u_h, u - v) \leq M\|u - u_h\|\|u - v\| \quad \text{for arbitrary} \quad v \in \mathrm{V}_h$$

Thus, by dividing by $\alpha\|u - u_h\|$ we obtain the result. In general, in order to get an asymptotic error estimate in h, it is sufficient to estimate the best approximation error of V_h that is $\min\{\ \|u - v\| \mid v \in \mathrm{V}_h\}$. □

Remark 2.4. *By means of the Céa's lemma, it is possible to define the strong consistency property of the finite element schemes. In particular we can affirm that the methods with this property guarantee the optimal approximation (having fixed the partition $\mathcal{I}_h$ and the polynomial degree k) .*

We give the following approximation property.

Property 2.5. *Let $u \in \mathrm{H}_0^1(\Omega)$ be the exact solution of the weak formulation (2.4) and $u_h \in \mathrm{V}_h$ its finite element approximation using continuous piecewise polynomials of degree $k \geq 1$. Assume also that $u \in \mathrm{H}^s(\Omega)$ for some $s \geq 2$. Then the following error estimate holds*

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq \frac{M}{\alpha}Ch^l|u|_{\mathrm{H}^{l+1}(\Omega)} \quad (2.13)$$

where $l = \min(k, s - 1)$. Under the same assumptions, we have also

$$\|u - u_h\|_{\mathrm{L}^2(\Omega)} \leq Ch^{l+1}|u|_{\mathrm{H}^{l+1}(\Omega)}\ . \quad (2.14)$$

The proof of these approximations is referred to [110]. The estimate (2.13) shows that the method is convergent i.e., approximation error tends to zero as $h \to 0$ and the order of convergence is l in $\mathrm{H}^1(\Omega)$ norm and $l + 1$ in $\mathrm{L}^2(\Omega)$ norm respectively. We also see that there is no convenience in increasing the degree k of the finite element approximation if the solution u is not sufficiently smooth. In this case l is called a regularity threshold. An alternative to

gain accuracy in any case is to reduced the step size h. When the exact solution u has the minimum regularity ($s = 1$), the Cèa's lemma ensures that the finite element method is still convergent since as $h \to 0$ the subspace V_h is dense into V. However the estimate (2.13) is no longer valid so that it is not possible to establish the order of convergence of the numerical method $\mathrm{H}^1(\Omega)$ norm. We summarize in the table 2.1 the orders of convergence of the finite element for $k = 1, \cdots, 4$ and $s = 1, \cdots, 5$.

k	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
1	only convergence	h^1	h^1	h^1	h^1
2	only convergence	h^1	h^2	h^2	h^2
3	only convergence	h^1	h^2	h^3	h^3
4	only convergence	h^1	h^2	h^3	h^4

Table 2.1: Order of convergence with respect to $\|\cdot\|_{\mathrm{H}^1(\Omega)}$ of the FE method as a function of k (degree of polynomials) and s (the Sobolev regularity of the solution u).

2.1.2 Interpolation operator and error

In order to define the interpolation operator, it is primordial to identify the degrees of freedom and the shape functions. In fact it is an operator defined on the space of continuous functions and valued in the finite elements spaces. We consider the finite element space X_h^k which is the space of continuous functions defined in Ω such that the restriction in each element K of the triangulations $\mathcal{I}_h$ are polynomials of degree $\leq k$. For the space X_h^k and for each $v \in \mathcal{C}^0(\overline{\Omega})$ we can set

$$I_h^k(v) := \sum_{i=1}^{N_h} v(a_i)\varphi_i \tag{2.15}$$

where a_i are the nodes on $\overline{\Omega}$, φ_i are the corresponding shape functions. The interpolant $I_h^k(v)$ is the unique function in X_h^k which takes the same values of the given function v at all the nodes a_i.
We have the following interpolation error from [113]:

Theorem 2.6. *Let $\mathcal{I}_h$ be a family of triangulations and assume that $m = 0$ or 1, $l = \min(k, s-1) \geq 1$. Then there exists a constant C, independent of h, such that*

$$|v - I_h^k(v)|_{m,\Omega} \leq C\, h^{l+1-m} |v|_{l+1,\Omega} \quad \forall\ v \in \mathrm{H}^s(\Omega)\ . \tag{2.16}$$

The proof of this result can be found in [113].
In the next section 2.1.3, we assume $\Omega = [a, b]$ and in this case X_h^k is the space of continuous function over $[a, b]$ whose restriction on each subdivisions I_j are polynomials of degree $\leq k$.

2.1.3 Generation of the shape functions in the 1D case

We now focus to an important issue that consists to generate suitable basis functions φ_j for the finite element space X_h^k in the special cases $k = 1$ and $k = 2$. The point consists to choose appropriately a set of degrees of freedom for each element I_j of the partition $\mathcal{I}_h$ i.e, the parameters which permit uniquely identifying a function X_h^k. Therefore, the generic function u_h can be written as

$$u_h(x) = \sum_{i=0}^{n} u_i \varphi_i(x)$$

where $\{u_i\}$ denotes the set of the degrees of freedom of u_h and the basis functions φ_i (also called shape functions) are assumed to satisfy the Lagrange interpolation property
$\varphi_i(x_j) = \delta_{i,j}$ for $i, j = 0, \cdots, n$ where $\delta_{i,j}$ is the Kronecker symbol.

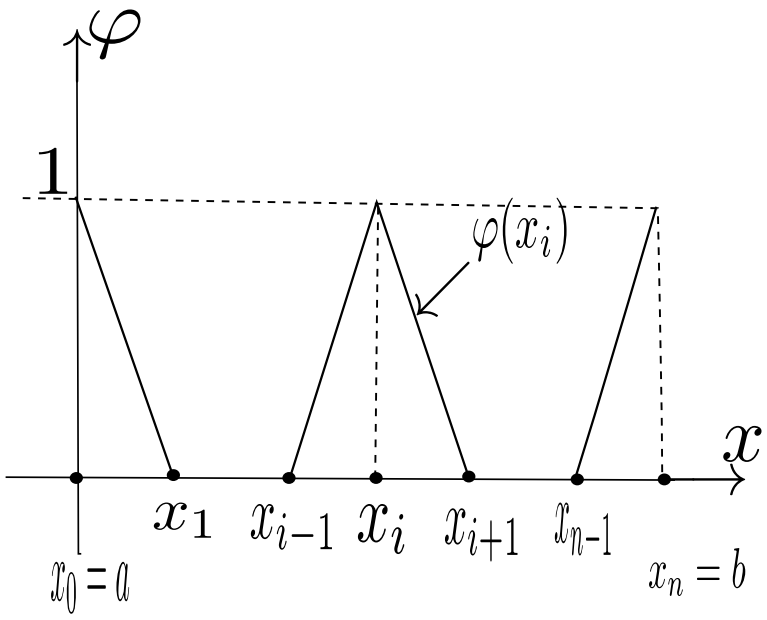

Figure 2.1: Basis functions of X_h^1 associated with internal and boundary nodes.

Shape functions of the space X_h^1

The space X_h^k consists of continuous and piecewise linear functions over the partition $\mathcal{I}_h$. Since a unique straight line passes through two distinct nodes, the number of degrees of freedom for u_h is equal to the number $n+1$ of nodes in the partition. Thus, $n+1$ shape functions $\varphi_i, i = 0, \cdots, n$ are needed to completely span the space X_h^1. The most natural choice for $\varphi_i, i = 1, \cdots, n-1$ is

$$\varphi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}}, & \text{for} \quad x_{i-1} \le x \le x_i \\ \\ \frac{x_{i+1}-x}{x_{i+1}-x_i}, & \text{for} \quad x_i \le x \le x_{i+1} \\ \\ 0 & \text{elsewhere.} \end{cases} \tag{2.17}$$

Thus the basis function φ_i is piecewise linear over $\mathcal{I}_h$, its value is 1 at the node x_i and 0 at all the other nodes of the partition. Its support (i.e, the subset of $[a,b]$ where φ_i is not vanishing) consists of the union of the intervals I_i and I_{i+1} if $1 \le i \le n-1$, while it coincides with the intervals I_1 (respectively I_n) if $i = 0$ (resp $i = n$). The Figure 2.1 shows the plots of φ_i, φ_0 and φ_n. For any interval $I_i = [x_{i-1}, x_i]$, $i = 1, \cdots, n$, the two basis functions φ_i and φ_{i-1} can be regarded as the images of two "reference" shape functions $\widehat{\varphi}_0$ and $\widehat{\varphi}_1$ (defined over the reference interval $[0,1]$) through the linear affine mapping $\phi : [0,1] \to I_i$.

$$x = \phi(\xi) = x_{i-1} + \xi(x_i - x_{i-1}), \ i = 1, \cdots, n \tag{2.18}$$

Defining $\widehat{\varphi}_0(\xi) = 1-\xi$, $\widehat{\varphi}_1(\xi) = \xi$, the two shape functions φ_i and φ_{i+1} can be constructed over the interval I_i as (see Figure 2.2)

$$\varphi_{i-1}(x) = \widehat{\varphi}_0(\xi(x)), \ \ \varphi_i(x) = \widehat{\varphi}_1(\xi(x))$$

where $\xi(x) = \frac{x-x_{i-1}}{x_i-x_{i-1}}$

Remark 2.7.

- *The elements $I_i, i = 1, \cdots, n$ do not need to have constant length h (this is different in respect to what happens for the classical FD method)*
- *The basis function $\varphi_0(x)$ and $\varphi_n(x)$ have analytical expressions like $(2.17)_1$ and $(2.17)_2$ respectively .*

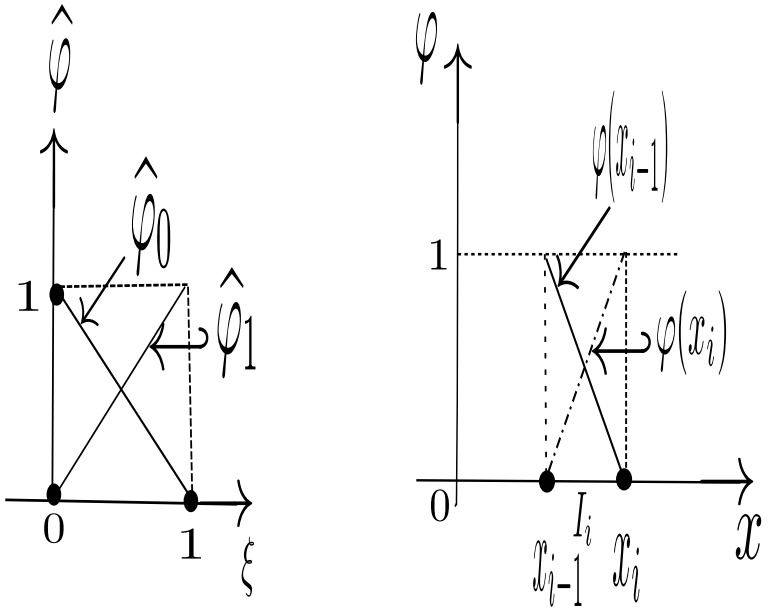

Figure 2.2: Linear affine mapping φ from the reference interval $[0,1]$ to the generic interval I_i of the partition .

Shape functions of the space X_h^2

The generic function $u_h \in \mathrm{X}_h^2$ is a piecewise polynomial of degree two over each interval I_i. It can be uniquely determined when three of its values at three distinct points of I_i are assigned and traditionally the third node is the middle point of the element I_i. In order to ensure continuity of u_h over $[a,b]$ the degrees of freedom are thus equal to $2n+1$. In this case, it is convenient to label the degrees of freedom and the corresponding nodes in the partition starting from $x_0 = 0$ until $x_{2n} = 1$ in such a way that the midpoints of each interval correspond to the nodes with odd index while the endpoints of each interval correspond to the nodes with even index.
The explicit expression of the single shape function is

$$(i \text{ even})\ \varphi_i(x) = \begin{cases} \frac{(x-x_{i-1})(x-x_{i-2})}{(x_i-x_{i-1})(x_i-x_{i-2})} & \text{for} \quad x_{i-2} \le x \le x_i, \\ \frac{(x_{i+1}-x)(x_{i+2}-x)}{(x_{i+1}-x_i)(x_{i+2}-x_i)} & \text{for} \quad x_i \le x \le x_{i+2}, \\ 0 \qquad\qquad \text{elsewhere} \end{cases} \tag{2.19}$$

$$(i \text{ odd})\ \varphi_i(x) = \begin{cases} \frac{(x_{i+1}-x)(x-x_{i-1})}{(x_{i+1}-x_i)(x_i-x_{i-1})} & \text{for} \quad x_{i-1} \le x \le x_{i+1}, \\ 0 \qquad\qquad \text{elsewhere} \end{cases} \tag{2.20}$$

Each basis function enjoys the property that $\varphi_i(x_j) = \delta_{i,j},\ i,j = 0,\cdots,2n$.
The shape functions for X_h^2 on the reference interval $[0,1]$ are

$$\widehat{\varphi}_0(\xi) = (1-\xi)(1-2\xi), \quad \widehat{\varphi}_1(\xi) = 4(1-\xi)\xi, \quad \widehat{\varphi}_2(\xi) = \xi(2\xi-1) \tag{2.21}$$

and they are shown in Figure 2.3.
As in the case of piecewise linear finite elements of X_h^1, the shape functions (2.19) and (2.20) are the images of (2.21) through the affine mapping (2.18).
We notice that the support of the basis function φ_{2i+1} associated with the midpoint x_{2i+1} coincides with the interval to which the midpoint belongs. Due to its shape, φ_{2i+1} is usually

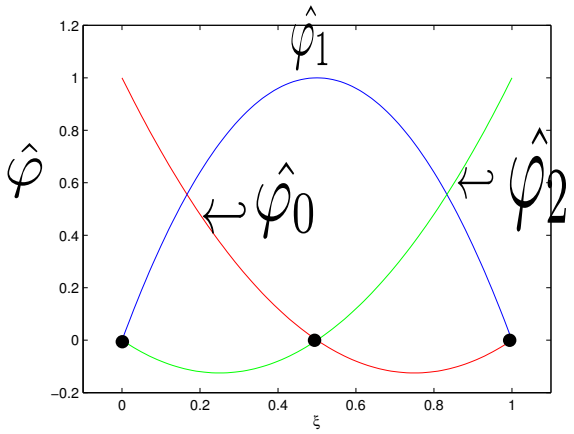

Figure 2.3: Shape functions of X_h^2 on the reference interval .

referred to as bubble function. A procedure analogous to that examined in this section can be used in principle to construct a basis for every subspace X_h^k with k being arbitrary. However, it is important to remember that an increase in the degree k of the polynomial gives rise to an increase of the number of the degrees of freedom and, consequently, of the computational cost required for the solution of the final linear system.

2.1.4 Generation of the shape functions in 2D case

Of course, the representation of the shape functions in the 2D case is not as easy as in the 1D case, thus we choose a graphic representation of the space X_h^1 and a more accurate definition of the X_h^2 space (in fact this space will be used later for carrying out numerical experiences). The extension in 2D of the representation of a shape function of X_h^1 can be seen properly in

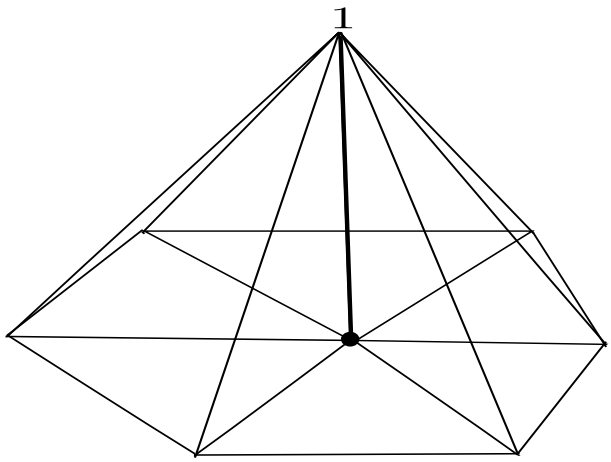

Figure 2.4: Shape function of X_h^1 and its support .

the Figure 2.4.

Shape functions of the space X_h^2

Suppose we have six nodes P_i of coordinates (x_i, y_i), $i = 1, \cdots, 6$ on the boundary of each triangle K, eventually also with curved edges, belonging to the partition of the domain Ω in the physical space (x, y) and that with each node P_i is associated a basis function $\varphi_i(x, y)$,

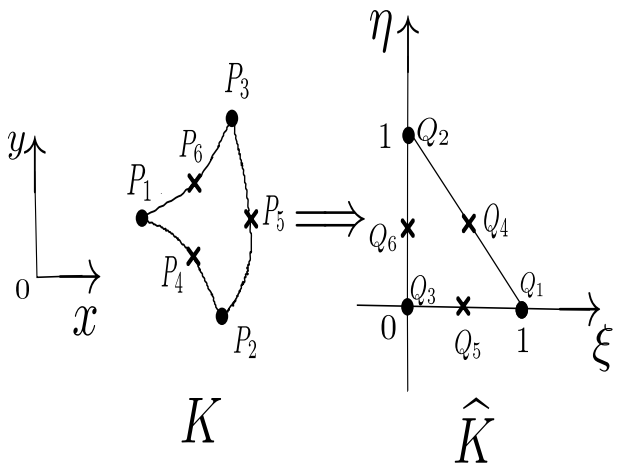

Figure 2.5: Element K of the physical space transformed in reference element $\widehat{K}$ of the transformed space .

polynomial of second degree. Moreover we suppose to define a correspondence, given by

$$x = \sum_{i=1}^{6} \varphi_i x_i$$
$$y = \sum_{i=1}^{6} \varphi_i y_i$$

between K and $\widehat{K}$, the reference element belonging to the transformed space (ξ, η) (see Figure 2.5). Then we can define the six basis functions with respect to each node in the reference space and they are:

$$\varphi_1(\xi,\eta) = 2\,\xi^2 - \xi$$
$$\varphi_2(\xi,\eta) = 2\,\eta^2 - \eta$$
$$\varphi_3(\xi,\eta) = 2\,z^2 - z$$
$$\varphi_4(\xi,\eta) = 4\,\xi\,\eta$$
$$\varphi_5(\xi,\eta) = 4\,\eta\,z$$
$$\varphi_6(\xi,\eta) = 4\,\xi\,z$$

with the relations

$$\sum_{i=1}^{6} \varphi_i = 1$$
$$z = 1 - \xi - \eta \ .$$

In order to build some terms necessary to the numerical integration process, we compute the constants

$$a_1 = 4(x_1 + x_3 - 2\,x_6), \quad a_2 = 4(x_3 + x_4 - x_5 - x_6), \quad a_3 = -x_1 - 3\,x_3 + 4\,x_6$$
$$b_1 = 4(y_1 + y_3 - 2\,y_6), \quad b_2 = 4(y_3 + y_4 - y_5 - y_6), \quad b_3 = -y_1 - 3\,y_3 + 4\,y_6$$
$$c_1 = 4(x_3 + x_4 - x_5 - x_6), \quad c_2 = 4(x_2 + x_3 - 2\,x_5), \quad c_3 = -x_2 - 3\,x_3 + 4\,x_5$$
$$d_1 = 4(y_3 + y_4 - y_5 - y_6), \quad d_2 = 4(y_2 + y_3 - 2\,y_5), \quad d_3 = -y_2 - 3\,y_3 + 4\,y_5 \ ,$$

then the Jacobian of the isoparametric transformation is :

$$J = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{bmatrix}$$

where

$$\frac{\partial x}{\partial \xi} = a_1\,\xi + a_2\,\eta + a_3$$
$$\frac{\partial y}{\partial \xi} = b_1\,\xi + b_2\,\eta + b_3$$
$$\frac{\partial x}{\partial \eta} = c_1\,\xi + c_2\,\eta + c_3$$
$$\frac{\partial y}{\partial \eta} = d_1\,\xi + d_2\,\eta + d_3 \ .$$

An exhaustive analysis of the finite element methods is referred to books such as [28, 71, 110, 137].

2.2 Discontinuous Galerkin (DG) finite elements methods

Reed and Hill [116] in 1973 introduced the first discontinuous Galerkin method for hyperbolic equations. It was used for second order elliptic problems by Douglas-Dupont [36] and fourth order problems by Baker [10]. Later on, the method has been abandoned because of the large matrix of its algebraic system, but has got a great revival some ten years ago mainly by Cockburn-Shu [31, 32, 33, 34] and also for an application to problems where the elliptic part is not dominant such as strongly advection-dominated equations and very thin Reissner-Mindlin plates [7]. Our interest in this section is the application of the DG method for problems where the elliptic part is important. We suggest [3] for an unified and complete presentation of the DG method for elliptic problems.
Among the most important aims that motivated the used of DG to partial differential equations, we could quote:

- Local conservation properties
- Mesh adaptation on irregular grids
- hp-Adaptivity with locally varying polynomial degrees
- Multi-domain approaches
- High-order extension of the finite volume methods
- Ease of implementing h-refinement with hanging nodes
- Ease of implementing p-refinement with locally varying p

For elliptic problems, the DG methods were formulated by using penalty methods.

2.2.1 Enforcing Dirichlet boundary condition through penalties

In 1968, Lions [82] used the penalty formulation for enforcing the Dirichlet boundary conditions; for instance, by considering $-\Delta\omega = f$ in Ω and $\omega = g$ on $\partial\Omega$ where f is taken in $\mathrm{L}^2(\Omega)$ and $g \in \mathrm{H}^{-\frac{1}{2}}(\partial\Omega)$. He regularized the above problem by replacing the Dirichlet boundary condition by the approximate boundary condition $\omega + \mu^{-1}\dfrac{\partial\omega}{\partial n} = g$ where $\mu \ggg 1$ is a penalty parameter. He also proved that for each $\mu > 0$, there is a unique solution u of the problem so that when μ goes to infinity, this solution converges to the solution ω of the original problem. The weak regularized form of the above problem is to find $\omega \in \mathrm{H}^1(\Omega)$ such that

$$\int_\Omega \nabla\omega\cdot\nabla v dx + \int_{\partial\Omega} \mu(\omega - g) ds = \int_\Omega f v dx \quad \forall v \in \mathrm{H}^1(\Omega) \ .$$

Aubin [6] for the problem above, in the finite difference framework, proved convergence to the exact solution provided the penalty parameter μ goes to infinity as the discretization parameter h goes to zero, that is when μ is of the order of $h^{-1+\epsilon}$ for arbitrary small $\epsilon > 0$.
In the finite element context, Babuška [9], for the same problem, with homogeneous boundary conditions ($g = 0$), obtained a convergence rate of order $h^{\frac{2k+1}{3}}$ in the energy norm, when the penalty parameter μ is taken to be of the order of $h^{-\frac{(2k+1)}{3}}$, where k represents the degree of the polynomials considered. The lack of optimality in the order of convergence is due to the lack of consistency of the weak regularized formulation. In fact, the exact solution ω does not satisfy the weak regularized formulation, it rather satisfies

$$\int_\Omega \nabla\omega\cdot\nabla v dx - \int_{\partial\Omega} \frac{\partial\omega}{\partial n} v ds + \int_{\partial\Omega} \mu(\omega - g) v ds = \int_\Omega f v dx \quad \forall \quad v \in \mathrm{H}^1(\Omega) \ .$$

Nitsche's formulation

The penalty parameter has been included also in the Nitsche's formulation [99] but without introducing any consistency error. His formulation is:

$$\int_\Omega \nabla\omega\cdot\nabla v dx - \int_{\partial\Omega} \frac{\partial\omega}{\partial n} v ds - \int_{\partial\Omega} \omega\frac{\partial v}{\partial n} ds + \int_{\partial\Omega} \mu(\omega - g) v ds = \int_\Omega f v dx - \int_{\partial\Omega} g\frac{\partial v}{\partial n} ds \quad \forall v \in \mathrm{H}^1(\Omega)$$

for any weighting value μ. He proved that when μ is taken as αh^{-1} where h is the element size and α is a sufficiently large constant, the discrete solution converges to the exact solution with optimal order in H^1 (i.e $O(h^k)$) and L^2 (i.e $O(h^{k+1})$ where k is the degree of the polynomial used. Let's give the primal and the flux formulation of the DG methods.

2.2.2 The flux and the primal formulation

We consider the model problem

$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad \Omega \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{2.22}$$

where Ω is assumed to be convex polygonal domain and f is given in $\mathrm{L}^2(\Omega)$.
In order to obtain the DG methods, we first rewrite the problems as a first order system

$$\begin{aligned} \underline{\sigma} &= -\nabla u \\ \nabla\cdot\underline{\sigma} &= f \quad \text{in} \quad \Omega \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned}$$

Multiplying these equations by test function $\underline{\tau}$ and v respectively and integrating by parts on a subset K of Ω, the weak formulation is given:

$$\begin{aligned} \int_K \underline{\sigma}\cdot\underline{\tau}\ dx &= \int_K u\nabla\cdot\underline{\tau}\ dx - \int_{\partial K} u\underline{\tau}\cdot n_K\ ds \\ -\int_K \underline{\sigma}\cdot\nabla v\ dx &= \int_K f v\, dx - \int_{\partial K} \underline{\sigma}\cdot n_K v\ ds \end{aligned} \tag{2.23}$$

where n_K is the unit outward normal to ∂K.
Let us assume that a triangulation $\mathcal{I}_h = \{K\}$ of Ω. We define
$\mathsf{V}_h := \{v \in \mathrm{L}^2(K) : v|_K \in \mathbb{P}(K) \quad \forall\ K \in \mathcal{I}_h\}$
$\Sigma_h := \{\underline{\tau} \in [\mathrm{L}^2(K)]^2 : \underline{\tau}|_K \in \Sigma(K) \quad \forall\ K \in \mathcal{I}_h\}$
where $\mathbb{P}(K) = \mathbb{P}_k(K)$ is the space of polynomial functions of degree $\leq k$ on K and $\Sigma(K) =$

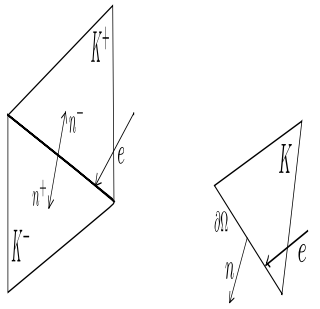

Figure 2.6: Adjacent elements K^-, K^+ and interior edge e.

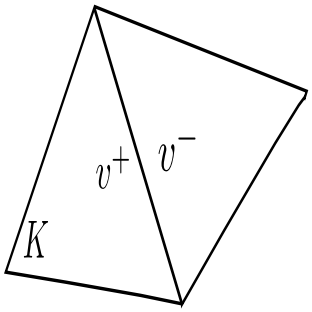

Figure 2.7: trace of v on ∂K.

$[\mathbb{P}_k(K)]^2$.
The discrete formulation becomes : find $u_h \in \mathsf{V}_h$ and $\underline{\sigma}_h \in \Sigma_h$ such that for all $K \in \mathcal{I}_h$ we have

$$\int_K \underline{\sigma}_h \cdot \underline{\tau}\, dx = \int_K u_h \nabla \cdot \underline{\tau}\, dx - \int_{\partial K} \widehat{u}_K\, \underline{\tau} \cdot n_K\, ds \quad \forall\, \underline{\tau} \in \Sigma_h(K) \tag{2.24}$$

$$-\int_K \underline{\sigma}_h \cdot \nabla v\, dx = \int_K f v\, dx - \int_{\partial K} \widehat{\underline{\sigma}}_K \cdot n_K v\, ds \quad \forall\, v \in \mathsf{V}_h(K) \tag{2.25}$$

where the numerical fluxes $\widehat{\underline{\sigma}}_K$ and $\widehat{u}_K$ are approximations of $\underline{\sigma} = -\nabla u$ and of u respectively on the boundary K. Equations (2.24)-(2.25) is the flux formulation that must be completed by expressing the numerical fluxes $\widehat{\underline{\sigma}}_K$ and $\widehat{u}_K$ in terms of $\underline{\sigma}_h$, u_h and of the boundary conditions. In order to obtain the primal formulation, we need to introduce some notations. Let e be an interior edge that is shared by elements K^+ and K^-. We define the unit normal vectors n^+ and n^- on e pointing exterior to K^+ and K^- respectively, see Figure 2.6. On Figure 2.7, v^+ is the interior trace of v on ∂K taken from within K and v^- is the exterior trace of v on ∂K taken from outside K. We have the following operator:
jumps operator $[q] := q^+ n^+ + q^- n^-$ for scalar function q and $[\underline{\varphi}] := \underline{\varphi}^+ n^+ + \underline{\varphi}^- n^-$ for vector function $\underline{\varphi}$ $e \in \xi_h^0$, where ξ_h^0 is the set of interior edges e.
Average operator:$\{q\} = \frac{1}{2}(q^+ + q^-)$, for scalar function q and $\{\underline{\varphi}\} = \frac{1}{2}(\underline{\varphi}^+ + \underline{\varphi}^-)$ for vector function $\underline{\varphi}$ on $e \in \xi_h^0$.

Remark 2.8. *The jump $[q]$ of the scalar function q is a vector parallel to the normal and the jump $[\underline{\varphi}]$ of the function $\underline{\varphi}$ is a scalar quantity.*

Remark 2.9. *The functional spaces $\mathsf{V}_h(K)$ and $\Sigma_h(K)$ have to respect a compatibility condition similar to that seen for Stokes problem.*

Writing the flux formulation (2.24)-(2.25) over all the elements, we obtain

$$\int_\Omega \underline{\sigma}_h \cdot \underline{\tau}\ dx = \int_\Omega u_h \nabla_h \cdot \underline{\tau}\ dx - \sum_{K\in\mathcal{I}_h} \int_{\partial K} \widehat{u}_K\, \underline{\tau}\cdot n_K\ ds \quad \forall\ \underline{\tau} \in \Sigma_h \tag{2.26}$$

$$-\int_\Omega \underline{\sigma}_h \cdot \nabla_h v\ dx = \int_\Omega f v\ dx - \sum_{K\in\mathcal{I}_h} \int_{\partial K} \widehat{\underline{\sigma}}_K \cdot n_K v\ ds \quad \forall\ v \in \mathsf{V}_h \tag{2.27}$$

where $\nabla_h v$ and $\nabla_h \cdot \underline{\tau}$ are functions whose restriction to each element $K \in \mathcal{I}_h$ are equal to ∇v and $\nabla \cdot \underline{\tau}$ respectively and $\widehat{u} = \widehat{u}(u_h)$, $\widehat{\underline{\sigma}} = \widehat{\underline{\sigma}}(u_h, \underline{\sigma}_h)$ are numerical fluxes chosen in a suitable way (see later).
The following crucial formula, relating the functions q and $\underline{\varphi}$ with the jump functions $[q]$ and $[\underline{\varphi}]$ and the average functions $\{q\}$ and $\{\underline{\varphi}\}$, is valid:

$$\sum_{K\in\mathcal{I}_h} \int_{\partial K} q\, \underline{\varphi} n_K\ ds = \int_\Gamma [q]\{\underline{\varphi}\}\ ds + \int_{\Gamma^0} \{q\}\,[\underline{\varphi}]\ ds \tag{2.28}$$

where Γ is the union of all the edges of K and Γ^0 the union of edges of K internal to Ω.
We give the proof of the crucial formula (2.28) for $q = v$ and $\varphi = \tau$.

Proof. Let ξ_h the set of all edge e of K. We have

$$\underbrace{\sum_{K\in\mathcal{I}_h} \int_{\partial K} v\underline{\tau}\cdot n_K\ ds}_{I} = \underbrace{\sum_{e\in\xi_h} \int_e [v]\{\underline{\tau}\}\ ds + \sum_{e\in\xi_h^0} \int_e \{v\}[\underline{\tau}]\ ds}_{II}$$

but

$$II = \underbrace{\sum_{e\in\xi_h^0} \int_e ([v]\cdot\{\underline{\tau}\} + \{v\}\cdot[\underline{\tau}])\ ds}_{III} + \sum_{e\in\partial\Omega} \int_e v\underline{\tau}\cdot n\ ds$$

on internal edges $e \in \xi_h^0$; since $n^+ = -n^-$ moreover we have

$$\begin{aligned}\int_e ([v]\cdot\{\underline{\tau}\} + \{v\}\cdot[\underline{\tau}])\ ds &= \frac{1}{2}\int_e (v^+ - v^-)(\tau^+ + \tau^-)\cdot n^+ + (v^+ + v^-)(\tau^+ - \tau^-)\cdot n^+\ ds \\ &= \frac{1}{2}\int_e (2v^+\tau^+\cdot n^+ - 2v^-\tau^-\cdot n^+)\ ds \\ &= \int_e (v^+\tau^+\cdot n^+ + v^-\tau^- n^-)\ ds\end{aligned}$$

Summing with respect to all the edges $e \in \xi_h^0$, we obtain the desired result. □

Applying the identity (2.28) to (2.26) and (2.27) , it comes

$$\int_\Omega \underline{\sigma}_h \cdot \underline{\tau}\ dx = \int_\Omega u_h \nabla_h \cdot \underline{\tau}\ dx - \int_\Gamma [\widehat{u}]\cdot\{\underline{\tau}\}\ ds - \int_{\Gamma^0} \{\widehat{u}\}[\underline{\tau}]\ ds \quad \forall\ \underline{\tau} \in \Sigma_h \tag{2.29}$$

$$-\int_\Omega \underline{\sigma}_h \cdot \nabla_h v\ dx + \int_\Gamma \{\widehat{\underline{\sigma}}\}\cdot[v]\ ds + \int_{\Gamma^0} [\widehat{\underline{\sigma}}]\{v\}\ ds = \int_\Omega f v\ dx \quad \forall\ v \in \mathsf{V}_h \tag{2.30}$$

Using the classical Green formula

$$\int_\Omega u_h \nabla_h \cdot \underline{\tau}\ dx = -\int_\Omega \nabla_h u_h \cdot \underline{\tau}\ dx + \sum_K \int_{\partial K} u_h \underline{\tau}\cdot n\ ds$$

Note: From here on, we will indicate by $\sum_e$ both the expressions $\sum_{e\in\xi_h}$ and $\sum_{e\in\xi_h^0}$ assigning to the reader the work to suitably fix the right choice.
The discrete problem (2.29)- (2.30) can be rewritten as: find $u_h \in \mathsf{V}_h$, $\sigma_h \in \Sigma_h$ such that $\forall\, \underline{\tau} \in \Sigma_h$, $\forall\, v \in \mathsf{V}_h$

$$\int_\Omega \underline{\sigma}_h\cdot \underline{\tau}\, dx = \int_\Omega \nabla_h u_h\cdot \underline{\tau}\, dx - \sum_e \int_e [\widehat{u} - u_h]\cdot \{\underline{\tau}\}\, ds - \sum_e \int_e \{\widehat{u} - u_h\}[\underline{\tau}]\, ds \tag{2.31}$$

$$-\int_\Omega \underline{\sigma}_h\cdot \nabla_h v\, dx = \int_\Omega f v\, dx - \sum_e \int_e [v]\cdot \{\underline{\widehat{\sigma}}\}\, ds - \sum_e \int_e \{\widehat{v}\}[\underline{\widehat{\sigma}}]\, ds \tag{2.32}$$

providing the assumption that $\nabla(\mathsf{V}_h) \subset \Sigma_h$.
Taking $\underline{\tau} = -\nabla_h v$ in (2.31) and summing the equations (2.31) and (2.32), we obtain

$$\begin{aligned}\int_\Omega \nabla_h u_h\cdot \nabla_h v\ dx + \sum_e \int_e [\widehat{u} - u_h]\cdot \{\nabla_h v\}\ ds + \sum_e \int_e \{\widehat{u} - u_h\}[\nabla_h v]\ ds = \int_\Omega f v\ dx - \\ \sum_e \int_e [v]\cdot \{\underline{\widehat{\sigma}}\}\ ds - \sum_e \int_e \{\widehat{v}\}[\underline{\widehat{\sigma}}]\ ds\end{aligned} \tag{2.33}$$

An important issue consists to make a choice of the numerical fluxes. If we express the numerical fluxes $\widehat{u}$ and $\widehat{\sigma}$, more precisely, taking $\widehat{u} = \widehat{u}(u_h)$ and $\underline{\widehat{\sigma}} = \underline{\widehat{\sigma}}(u_h, \nabla_h u_h)$, we obtain

$$B_h(u_h, v_h) = \int_\Omega f v\ dx \quad \forall\ v \in \mathsf{V}_h \tag{2.34}$$

with

$$\begin{aligned}B_h(u_h, v_h) := \int_\Omega \nabla_h u_h\cdot \nabla_h v\ dx + \sum_e \int_e ([\widehat{u} - u_h]\cdot \{\nabla_h v\} + \{\underline{\widehat{\sigma}}\}\cdot [v])\ ds + \\ \sum_e \int_e (\{\widehat{u} - u_h\}[\nabla_h v] + [\underline{\widehat{\sigma}}]\{v\})\ ds.\end{aligned} \tag{2.35}$$

(2.34) is called the primal formulation of the method and the bilinear form $B_h(\cdot\,,\cdot\,)$ the primal form.

2.2.3 Choices of numerical fluxes

Some choices of the fluxes take the form $\widehat{u} = \widehat{u}(u_h)$ and $\underline{\widehat{\sigma}} = \underline{\widehat{\sigma}}(u_h, \underline{\sigma}_h)$, but they require some considerations.

The non-stabilized interior penalty method

In this case, taking $\widehat{u} = \{u_h\}$ on $e \in \xi_h^0$, and $\widehat{u} = 0$ on $e \subset \partial\Omega$, we have $[\widehat{u} - u_h] = -[u_h]$, and $\{\widehat{u} - u_h\} = 0$.
On the other hand, by taking on every edge $\underline{\widehat{\sigma}} = -\{\nabla_h u_h\}$, we have $[\underline{\widehat{\sigma}}] = 0$, and $\{\underline{\widehat{\sigma}}\} = -\{\nabla_h u_h\}$.
Therefore, inserting these into (2.35) it comes

$$\int_\Omega \nabla_h u_h\cdot \nabla_h v\ dx - \sum_e \int_e [u_h]\cdot \{\nabla_h v\}\ ds - \sum_e \int_e [v]\cdot \{\nabla_h u_h\}\ ds = \int_\Omega f v\ dx \tag{2.36}$$

which is the non stabilized version of the interior penalty method see [5, 36, 135]. The discrete problem is given as

$$\text{find} \quad u_h \in \mathsf{V}_h : B_h(u_h, v) = \int_\Omega f v\ dx \quad \forall\ v \in \mathsf{V}_h$$

where the bilinear form $B_h(\cdot,\cdot)$ is the left hand side of (2.36).

Remark 2.10. *This formulation is symmetric and the final algebraic system obtained is also symmetric.*

The stabilized interior penalty method

If we take the numerical flux $\widehat{u}$ as $\widehat{u} = \{u_h\}$ on $e \in \xi_h^0$ and $\widehat{u} = 0$ on $e \in \partial\Omega$, the jump operator $[\widehat{u} - u_h] = -[u_h]$ and its average operator $\{\widehat{u} - u_h\} = 0$; moreover, if we take the numerical flux $\underline{\widehat{\sigma}}$ as $\underline{\widehat{\sigma}} = -\{\underline{\sigma}_h u_h\} + c|e|^{-1}[u_h]$ on every edge, then the jump operator $[\underline{\widehat{\sigma}}] = 0$ and its average operator $\{\underline{\widehat{\sigma}}\} = -\{\nabla_h u_h\} + c|e|^{-1}[u_h]$. Inserting this choice of the numerical fluxes into (2.35), it comes

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v \, dx - \sum_e \int_e [v] \cdot \{\nabla_h u_h\} \, ds - \sum_e \int_e [u_h] \cdot \{\nabla_h v\} \, ds + c \sum_e |e|^{-1} \int_e [u_h] \cdot [v] \, ds = \int_\Omega f v \, dx \tag{2.37}$$

which is the the stabilized interior penalty method see [5, 36, 135].

Remark 2.11.

- *The last term of left hand side of (2.37) is the penalty term where c is a constant "big enough"*
- *This formulation is symmetric and the final algebraic matrix obtained is also symmetric.*

The discrete problem is given

$$\text{find} \quad u_h \in \mathsf{V}_h : \quad B_h(u_h, v) = \int_\Omega f v \, dx \quad \forall \, v \in \mathsf{V}_h$$

where the bilinear form $B_h(\cdot\,,\cdot\,)$ is the left hand side of (2.37).

The Baumann-Oden method

Another choice of the fluxes are the following: take on the boundary of each element $\widehat{u} = \{u_h\} + n_K[u_h]$ on $e \in \xi_h^0$, $\widehat{u} = 0$ on $e \subset \partial\Omega$. Then from this choice of the numerical flux we have $[\widehat{u} - u_h] = 2[u_h] - [u_h] = [u_h]$ and $\{\widehat{u} - u_h\} = 0$; on the same time, taking on every edge $\underline{\widehat{\sigma}} = -\{\nabla_h u_h\}$ for this choice of $\underline{\widehat{\sigma}}$, we have $[\underline{\widehat{\sigma}}] = 0$ and $\{\underline{\widehat{\sigma}}\} = -\{\nabla_h u_h\}$. Therefore, the relations (2.34) and (2.35) become

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v \, dx - \sum_e \int_e [v] \cdot \{\nabla_h u_h\} \, ds + \sum_e \int_e [u_h] \cdot \{\nabla_h v\} \, ds = \int_\Omega f v \, dx \tag{2.38}$$

that is the Baumann-Oden method [18].
The discrete form can be written:

$$\text{find} \quad u_h \in \mathsf{V}_h : B_h(u_h, v) = \int_\Omega f v \, dx \quad \forall \, v \in \mathsf{V}_h$$

where the bilinear form $B_h(\cdot\,,\cdot\,)$ is the left hand side of (2.38).

Remark 2.12. *This scheme is not symmetric and so give a stiffness matrix non symmetric.*

Remark 2.13. *If conforming finite elements are used, then there are some terms of (2.38) which automatically annihilate; due to the simple formulation of (2.38), we decided to use it to define a finite element approach with conservation of numerical fluxes.*

The stabilized Baumann-Oden method

By taking on the boundary of each element K the numerical flux $\widehat{u} = \{u_h\} + n_K[u_h]$ on $e \in \xi_h^0$, $\widehat{u} = 0$ on $e \in \partial\Omega$ then $[\widehat{u} - u_h] = 2[u_h] - [u_h]$, and $\{\widehat{u} - u_h\} = 0$; if we take the numerical flux $\underline{\widehat{\sigma}} = -\{\nabla_h u_h\} + c|e|^{-1}[u_h]$, then $[\widehat{\sigma}] = 0$ and $\{\widehat{\sigma}\} = -\{\nabla_h u_h\} + c|e|^{-1}[u_h]$. Inserting these fluxes into (2.35), we obtain

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v \, dx - \sum_e \int_e [v] \cdot \{\nabla_h u_h\} \, ds + \sum_e \int_e [u_h] \cdot \{\nabla_h v\} \, ds + \sum_e c|e|^{-1} \int_e [u_h] \cdot [v] \, ds = \int_\Omega f v \, dx \tag{2.39}$$

which is the stabilized version of the Baumann-Oden method, see Rivière-Wheeler-Girault [117].
The discrete problem is given:

$$\text{find } u_h \in \mathsf{V}_h : B_h(u_h, v) = \int_\Omega f v \, dx \quad \forall\, v \in \mathsf{V}_h$$

where the bilinear form $B_h(\cdot\,,\cdot)$ is the left hand side of (2.39).

Remark 2.14. *The last term in the bilinear form of (2.39) is the penalty term that stabilizes the formulation with the penalty constant c "big enough".*

The incomplete interior penalty Galerkin (IIPG) method

If we take on the boundary of each element K then numerical flux $\widehat{u} = \{u_h\} + \frac{1}{2} n_K[u_h]$ on $e \in \xi_h^0$ and $\widehat{u} = 0$ on $e \subset \partial\Omega$, then the jump $[\widehat{u} - u_h] = [u_h] - [u_h] = 0$ and the average $\{\widehat{u} - u_h\} = 0$; meantime, taking on every edge the numerical flux $\underline{\widehat{\sigma}} = -\{\nabla_h u_h\} + c|e|^{-1}[u_h]$, then the jump operator $[\underline{\widehat{\sigma}}] = 0$ and the average operator $\{\underline{\widehat{\sigma}}\} = -\{\nabla_h u_h\} + c|e|^{-1}[u_h]$. Thus the relation (2.35) take the form

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v \, dx - \sum_e \int_e [v] \cdot \{\nabla_h u_h\} \, ds + \sum_e c|e|^{-1} \int_e [u_h] \cdot [v] \, ds = \int_\Omega f v \, dx \tag{2.40}$$

which is the incomplete interior penalty Galerkin method of Sun-Wheeler [129].
The discrete form consists to:

$$\text{find} \quad u_h \in \mathsf{V}_h : B_h(u_h, v) = \int_\Omega f v \, dx \quad \forall\, v \in \mathsf{V}_h$$

where the bilinear form $B_h(\cdot\,,\cdot)$ is the left hand side of (2.40).

The method of Heinrich

If we set $\{v\}_\beta = \beta v^+ + (1-\beta) v^-$ where v^+ and v^- correspond to a given choice of an orientation on e, and choosing the numerical flux $\widehat{u} = \{u_h\}_{(1-\beta)}$ on $e \in \xi_h^0$, and $\widehat{u} = 0$ on $e \in \partial\Omega$, then the jump operator $[\widehat{u} - u_h] = -[u_h]$ and its average operator $\{\widehat{u} - u_h\} = \{u_h\}_{(1-\beta)} - \{u_h\}$; moreover, if we take the numerical flux $\widehat{\sigma} = -\{\nabla_h u_h\}_\beta + c|e|^{-1}[u_h]$ on every edge, then the jump operator $[\widehat{\sigma}] = 0$ and its average operator $\{\widehat{\sigma}\} = -\{\nabla_h u_h\} + c|e|^{-1}[u_h]$. Inserting these numerical fluxes into the relation (2.35), it comes

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v \, dx - \sum_e \int_e [v] \cdot \{\nabla_h u_h\} \, ds - \sum_e \int_e [u_h] \cdot \{\nabla_h v\} \, ds + c \sum_e |e|^{-1} \int_e [u_h] \cdot [v] \, ds = \int_\Omega f v \, dx \tag{2.41}$$

which is the Heinrich method [63]. The discrete form is written:

$$\text{find}\, u_h \in \mathsf{V}_h : B_h(u_h, v) = \int_\Omega f v \; dx \quad \forall\, v \in \mathsf{V}_h$$

where the bilinear form $B_h(\cdot\,,\cdot\,)$ is the left hand side of the (2.41).

Remark 2.15. *The last term in the bilinear form of (2.41) is the stabilizing term where c is a penalty constant "big enough".*

2.2.4 Other choices of the numerical fluxes

There are some formulations of $\underline{\widehat{\sigma}}$ that depends only on $\underline{\sigma_h}$. When one of these choices is made, we have to obtain $\underline{\sigma_h}$ from the equation

$$\int_\Omega \sigma_h \cdot \underline{\tau} \; dx = -\int_\Omega \nabla_h u_h \cdot \underline{\tau} \; dx - \sum_e \int_e [\widehat{u} - u_h] \cdot \{\underline{\tau}\} \; ds - \sum_e \int_e \{\widehat{u} - u_h\} [\underline{\tau}] \; ds \tag{2.42}$$

and substitute it into

$$-\int_\Omega \underline{\sigma_h} \cdot \nabla_h v \; dx = \int_\Omega f v \; dx - \sum_e \int_e [v] \cdot \{\underline{\widehat{\sigma}}\} \; ds - \sum_e \int_e \{\widehat{v}\} [\underline{\widehat{\sigma}}] \; ds \tag{2.43}$$

But some definitions are needed in order to proceed.

Definition 2.16. *We define the lifting R, l such that $\forall\, v \in \mathsf{V}_h$, $R([v]) \in \Sigma_h$, $l(\{v\}) \in \Sigma_h$. Then we have*

$$\int_\Omega R([v]) \cdot \underline{\tau} \; dx = -\sum_e \int_e [v] \cdot \{\underline{\tau}\} \; ds \quad \forall\, \underline{\tau} \in \Sigma_h$$

$$\int_\Omega l(\{v\}) \cdot \underline{\tau} \; dx = -\sum_e \int_e \{v\} \cdot [\underline{\tau}] \; ds \quad \forall\, \underline{\tau} \in \Sigma_h$$

By means of this definition we can write $\underline{\sigma_h} = -\nabla_h u_h + R([\widehat{u}] - u_h) + l(\{\widehat{u} - u_h\})$. Substituting this expression of $\underline{\sigma_h}$ into the equation (2.43), we obtain the discrete formulation:

$$\text{find} \quad u_h \in \mathsf{V}_h : \; B_h(u_h, v) = \int_\Omega f v \; dx \quad \forall\, v \in \mathsf{V}_h \tag{2.44}$$

where

$$\begin{aligned} B_h(u_h, v) := & \int_\Omega \nabla_h u_h \cdot \nabla_h v \; dx - \int_\Omega R([\widehat{u} - u_h]) \cdot \nabla_h v \; dx - \\ & \int_\Omega l(\{\widehat{u} - u_h\}) \cdot \nabla_h v \; dx + \sum_e \int_e \{\underline{\widehat{\sigma}}\} \cdot [v] \; ds + \\ & \sum_e \int_e [\underline{\widehat{\sigma}}] \{v\} \; ds \\ \equiv & \int_\Omega \nabla_h u_h \cdot \nabla_h v \; dx + \sum_e \int_e [\widehat{u} - u_h] \cdot \{\nabla_h v\} \; ds + \\ & \sum_e \int_e \{\widehat{u} - u_h\} [\nabla_h v] \; ds + \sum_e \int_e \{\underline{\widehat{\sigma}}\} \cdot [v] \; ds + \\ & \sum_e \int_e [\underline{\widehat{\sigma}}] \{v\} \; ds \end{aligned} \tag{2.45}$$

The formulation (2.44) with (2.45) is the base of the method of Bassi-Rebay.

The first Bassi-Rebay method

If the numerical flux $\widehat{u}$ is taken such that $\widehat{u} = \{u_h\}$ on $e \in \xi_h^0$, and $\widehat{u} = 0$ on $\partial\Omega$, then the jump operator $[\widehat{u} - u_h] = [u_h]$ and the average operator $\{\widehat{u} - u_h\} = 0$; on the other hand, if the numerical flux $\underline{\widehat{\sigma}}$ on every edge is taken as $\underline{\widehat{\sigma}} = \{\underline{\sigma}_h\}$, then the jump operator $[\underline{\widehat{\sigma}}] = 0$ and its average operator $\{\underline{\widehat{\sigma}}\} = \{\underline{\sigma}_h\}$. By consequence, $\underline{\sigma}_h = -\nabla_h u_h - R([u_h])$. But we have

$$\begin{aligned}\sum_e \int_e \{\underline{\widehat{\sigma}}\}\cdot[v]\ ds &= -\sum_e \int_e \{\nabla_h u_h\}\cdot[v]\ ds - \sum_e \int_e \{R([u_h])\}\cdot[v]\ ds \\ &= \int_\Omega \nabla_h u_h \cdot R([v])\ dx + \int_\Omega R([u_h])\cdot R([v])\ dx\end{aligned}$$

and substituting in the bilinear form (2.45) gives this first Bassi-Rebay formulation [21]:

$$\text{find } u_h \in \mathsf{V}_h : \quad \int_\Omega [\nabla_h u_h + R([u_h])]\cdot[\nabla_h v + R([v])]\ dx = \int_\Omega f v\ dx \quad \forall\, v \in \mathsf{V}_h \tag{2.46}$$

or equivalently

$$\begin{aligned}\int_\Omega \nabla_h u_h \cdot \nabla_h v\ dx - \sum_e \int_e \{\nabla_h u_h\}\cdot[v]\ ds - \sum_e \int_e [u_h]\cdot\{\nabla_h v\}\ ds + \\ \int_\Omega R([u_h])\cdot R([v])\ dx = \int_\Omega f v\ dx \quad \forall\, v \in \mathsf{V}_h\ .\end{aligned} \tag{2.47}$$

Remark 2.17. *The first Bassi-Rebay formulation (2.46) is unstable.*

The second Bassi-Rebay method

The second Bassi-Rebay formulation is the stabilized version of the first Bassi-Rebay formulation. It requires to define a strain correction r_e on the edges e of $\mathcal{I}_h$.

Definition 2.18. *The strain correction r_e is such that $\forall\, v \in \mathsf{V}_h$, $r_e \in \Sigma_h$ and*

$$\int_\Omega r_e([v])\cdot\underline{\tau} + \sum_e \int_e [v]\cdot\{\underline{\tau}\} = 0 \quad \forall\, \underline{\tau} \in \Sigma_h\ .$$

From this definition we have $r_e([v]) = R([v])$. If now we take for the numerical flux $\widehat{u}$ as $\widehat{u} = \{u_h\}$ on $e \in \xi_h^0$ and $\widehat{u} = 0$ on $e \subset \partial\Omega$, then the jump operator $[\widehat{u} - u_h] = -[u_h]$ and the average operator $\{\widehat{u} - u_h\} = 0$; on the other hand, taking on every edge the numerical flux $\widehat{\sigma}$ as $\widehat{\sigma} = -\{\nabla_h u_h\} + c\, r_e([u_h])$, the jump operator $[\widehat{\sigma}] = 0$ and its average operator $\widehat{\sigma} = -\{\nabla_h u_h\} + c\{r_e([u_h])\}$. Thus the second Bassi-Rebay method is given [22]:

$$\begin{aligned}\text{find} \quad u_h \in \mathsf{V}_h : \int_\Omega \nabla_h u_h \cdot \nabla_h v\ dx - \sum_e \int_e [v]\cdot\{\nabla_h u_h\}\ ds - \sum_e \int_e [u_h]\cdot\{\nabla_h u\}\ ds + \\ c\sum_e \int_e r_e([u_h])\cdot r_e([v])\ ds = \int_\Omega f v\ dx \quad \forall\, v \in \mathsf{V}_h\ ,\end{aligned} \tag{2.48}$$

where the last integral in the left hand side is the stabilization term with the penalty weighting function c.

Remark 2.19. *The difference between this second Bassi-Rebay method and the interior penalty method is the choice of the stabilizing term. More frequently, the weighting function c is taken as $\eta_e h^{-1}$ on each $e \in \xi_h$ with η_e positive number.*

2.2.5 Convergence of the DG

The studies of the convergence requires two useful tools that are the trace and the inverse inequalities.

Lemma 2.20. *Trace inequality (Agmon [4], Arnold [5]) .*
Let D be a polygonal and let e be an edge of D. For a function $\varphi \in \mathrm{H}^1(D)$ it holds: $\exists\, C_t > 0$, only depending on the minimum angle of D, such that

$$\|\varphi\|_{0,e} \leq C_t(|e|^{-1}\|\varphi\|_{0,D}^2 + |e||\varphi|_{1,D}^2)^{\frac{1}{2}} \ .$$

We give the proof of this trace inequality in 1D.

Proof. Let's take $D = [0, h]$. Then by the first fundamental theorem of calculus, we have

$$\varphi^2(x) = \varphi^2(0) + \int_0^x d\varphi^2(t) = \varphi^2(0) + 2\int_0^x \varphi(t)\varphi^{'}(t)dt$$

hence $\varphi^2(0) = \varphi^2(x) - 2\int_0^x \varphi(t)\varphi^{'}(t)dt$. Integrating both term of the equality from 0 to h, it comes

$$\begin{aligned} h\varphi^2(0) &= \int_0^h \varphi^2(x)\ dx - 2\int_0^h\int_0^x \varphi(t)\varphi^{'}(t)\ dtdx \\ &\leq \|\varphi\|_{0,D}^2 + 2h\|\varphi\|_{0,D}|\varphi|_{1,D} \\ &\leq \|\varphi\|_{0,D}^2 + \|\varphi\|_{0,D}^2 + h^2|\varphi|_{1,D}^2 \quad \text{since} \quad 2ab \leq a^2 + b^2 \end{aligned}$$

then dividing by h , we have

$$\varphi^2(0) \leq 2h^{-1}\|\varphi\|_{0,D}^2 + h|\varphi|_{1,D}^2$$

hence the result. □

Lemma 2.21. *Inverse inequality (Nitsche [99]).*
Let D be a polygon of diameter h_D and let P be a polynomial on D. Then $\exists\ C_{inv} > 0$, only depending on the minimum angle of D and the degree of P, such that:

$$|P|_{1,D} \leq C_{inv}h_D^{-1}\|P\|_{0,D} \ . \tag{2.49}$$

We give the proof in 1D .

Proof. Let us consider the intervals $D = [0, h]$ and $\widehat{D} = [0, 1]$. We set $\widehat{\varphi}(\widehat{x}) := \varphi(h\widehat{x}) \equiv \varphi(x)$ $\forall \widehat{x} \in \widehat{D}$ with $x = h\widehat{x}$. Then $dx = hd\widehat{x}$ and $\varphi^{'}(x) = h^{-1}\widehat{\varphi}^{'}(\widehat{x})$ Therefore

$$\begin{aligned} \int_0^h (\varphi^{'}(x))^2 dx &= h^{-2}\int_0^h (\widehat{\varphi}^{'}(\widehat{x}))^2 dx \\ &= h^{-1}\int_0^1 (\widehat{\varphi}^{'}(\widehat{x}))^2 d\widehat{x} \end{aligned}$$

On the other hand, for a polynomial $\widehat{\varphi}$, we have

$$\int_0^1 (\widehat{\varphi}^{'}(\widehat{x}))^2 d\widehat{x} \leq C_k \int_0^1 (\widehat{\varphi}(\widehat{x}))^2 d\widehat{x}$$

where C_k is a constant depending on the degree k of the polynomial, hence

$$\int_0^h (\varphi'(x))^2 dx \leq h^{-1} C_k \int_0^1 (\widehat{\varphi}(\widehat{x}))^2 d\widehat{x} \leq h^{-1} C_k h^{-1} \int_0^h (\varphi(x))^2 dx$$

and so

$$|\varphi|^2_{1,D} \leq C_k h^{-2} \|\varphi\|^2_{0,D} \ .$$

□

We will give the proof of the continuity and the coercivity of the stabilized interior penalty method with respect to a suitable norm.
We define the space: $\mathsf{V}(h) = \mathsf{V}_h + \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega) \subset \mathrm{H}^2(\mathcal{I}_h)$ endowed with the norm

$$|||v|||^2 = |v|^2_{1,h} + \sum_{K \in \mathcal{I}_h} h_K^2 |v|^2_{2,h} + \sum_{e \in \xi_h} \|[v]\|^2_{0,e} \quad \forall\ v \in \mathsf{V}(h) \tag{2.50}$$

where

$$|v|^2_{1,h} = \sum_{K \in \mathcal{I}_h} |v|^2_{1,K} \ .$$

We have seen that the bilinear form of the stabilized Interior Penalty method is (see 2.37):

$$B_h(u,v) := \int_\Omega \nabla_h u \cdot \nabla_h v \ dx - \sum_{e \in \xi_h} \int_e [v] \cdot \{\nabla_h u\} \ ds - \sum_{e \in \xi_h^0} \int_e [u] \cdot \{\nabla_h v\} \ ds + \sum_{e \in \xi_h} c|e|^{-1} \int_e [u] \cdot [v] \ ds$$

We want to show that this bilinear form is continuous, that is :

$$\exists\ C_b > 0 \quad \text{such that} \quad \forall\ u, v \in \mathsf{V}(h), B_h(u,v) \leq C_b |||u|||\ |||v|||$$

Proof. We have $\forall\ u, v \in \mathsf{V}(h)$

$$\int_\Omega \nabla_h u \cdot \nabla_h v \ dx \leq |u|_{1,h} \cdot |v|_{1,h} \leq |||u||||||v|||$$

using, the Cauchy-Schwarz inequality ($\sum a_i b_i \leq (\sum a_i^2)^{\frac{1}{2}} (\sum b_i^2)^{\frac{1}{2}}$) in the second term of the bilinear form, we have

$$\begin{aligned} \sum_e \int_e [u] \cdot \{\nabla_h v\} \ ds &= \sum_e \int_e |e|^{-\frac{1}{2}} [u] \cdot |e|^{\frac{1}{2}} \{\nabla_h v\} \ ds \\ &\leq (\sum_e |e|^{-1} \|[u]\|^2_{0,e})^{\frac{1}{2}} (\sum_e |e| \ \|\{\nabla_h v\}\|^2_{0,e})^{\frac{1}{2}} \end{aligned}$$

similarly, the third term in the bilinear form gives

$$\begin{aligned} \sum_e \int_e [v] \cdot \{\nabla_h u\} \ ds &= \sum_e \int_e |e|^{-\frac{1}{2}} [v] \cdot |e|^{\frac{1}{2}} \ \{\nabla_h u\} \ ds \\ &\leq (\sum_e |e|^{-1} \|[v]\|^2_{0,e})^{\frac{1}{2}} (\sum_e |e| \ \|\{\nabla_h u\}\|^2_{0,e})^{\frac{1}{2}} \end{aligned}$$

the last term of the bilinear form gives

$$\sum_e \frac{1}{|e|} \int_e [u][v] \ ds \leq (\sum_e \frac{1}{|e|} \|[u]\|^2_{0,e})^{\frac{1}{2}} (\sum_e \frac{1}{|e|} \|[v]\|^2_{0,e})^{\frac{1}{2}}$$

now, using the trace inequality, we have

$$\|\{\nabla_h u\}\|_{0,e} \leq C_t(|e|^{-1}|u|^2_{1,K} + |e|\ |u|^2_{2,K})^{\frac{1}{2}}$$

Therefore, the third term of the bilinear form becomes

$$\begin{aligned}\sum_e \int_e [v]\cdot\{\nabla_h u\}\ ds &\leq (\sum_e |e|^{-1}\|[v]\|^2_{0,e})^{\frac{1}{2}}(\sum_e |e| C_t(|e|^{-1}|u|^2_{1,K} + |e|\,|u|^2_{2,K}))^{\frac{1}{2}} \\ &\leq (\sum_e |e|^{-1}\|[v]\|^2_{0,e})^{\frac{1}{2}}(\sum_e C_t|u|^2_{1,K} + \sum_e C_t|e|^2\,|u|^2_{2,K})^{\frac{1}{2}} \\ &\leq C|||v|||||u|||\end{aligned}$$

similarly the second term of the bilinear form using the same trace inequality gives

$$\sum_e \int_e [u]\cdot\{\nabla_h v\}\ ds \leq C|||u|||||v|||$$

Therefore, putting everything together, we have

$$B_h(u,v) \leq C_b|||u|||||v||| \quad \forall\, u, v, \in \mathsf{V}(h)$$

□

We also show the coercivity of the bilinear form (2.37). Before we give the definition:

Definition 2.22. *$v \to (|v|^2_{1,h} + \sum_{e\in\xi_h}\|[v]\|^2_{0,e})^{\frac{1}{2}}$ be a norm for analyzing the coercivity of the bilinear form restricted to $v \in \mathsf{V}_h$. This norm on V_h and the one of (2.50) are equivalent on $\mathsf{V}(h)$.*

Proof of the coercivity:

Proof. $\forall\ v \in \mathsf{V}_h$,

$$B_h(v,v) = |v|^2_{1,h} - 2\sum_e \int_e [v]\cdot\{\nabla_h v\} + c\sum_e \frac{1}{|e|}\|[v]\|^2_{,e}$$

By using the trace and the inverse inequality, we have: for every $v \in \mathsf{V}_h$

$$\begin{aligned}\|\nabla_h v\|_{0,e} &\leq C_t(|e|^{-1}|v|^2_{1,K} + |e|\,|v|^2_{2,K})^{\frac{1}{2}} \\ &\leq C_t(|e|^{-1}|v|^2_{1,K} + |e|\,|e|^{-2}C_k^2|v|^2_{1,K})^{\frac{1}{2}} \\ &\leq C_t(1+C_k^2)^{\frac{1}{2}}|e|^{-\frac{1}{2}}|v|_{1,K} \\ &\leq C_t C_{inv}|e|^{-\frac{1}{2}}|v|_{1,K} \quad \text{with} \quad C_{inv} = (1+C_k^2)^{\frac{1}{2}}\end{aligned}$$

applying the Cauchy Schwarz and the arithmetic-geometric mean inequality for every $\epsilon > 0$, it comes

$$\begin{aligned}-2\sum_e \int_e [v]\cdot\{\nabla_h v\} &= -2\sum_e \int_e |e|^{-\frac{1}{2}}[v]\,|e|^{\frac{1}{2}}\{\nabla_h v\} \\ &\geq -2C_*(\sum_e \frac{1}{2}\|[v]\|^2_{0,e})^{\frac{1}{2}}\,|v|_{1,h} \\ &\geq -C_*\epsilon\sum_e \frac{1}{e}\|[v]\|^2_{0,e} - \frac{C_*}{\epsilon}|v|^2_{1,h}\end{aligned}$$

Table 2.2: Properties of some DG methods

Method	consistency	stability	condition	H^1	L^2
Interior penalty	Yes	Yes	$c_0 > c^*$	h^p	h^{k+1}
Baumann-Oden(k=1)	Yes	×	-	×	×
Baumann-Oden(k=2)	Yes	×	-	h^k	h^k
BAssi-Rebay	Yes	Yes	-	$[h^k]$	$[h^{k+1}]$

Therefore the bilinear form becomes

$$B_h(v,v) = |v|^2_{1,h} - 2\sum_e \int_e [v]\cdot\{\nabla_h v\} + c\sum_e \frac{1}{|e|}\|[v]\|^2_{,e}$$
$$\geq (1-\frac{C_*}{\epsilon})|v|^2_{1,h} + (C - C_*\epsilon)\sum_e \frac{1}{|e|}\|[v]\|^2_{0,e} \quad \forall\, v \in \mathsf{V}_h\ .$$

Hence, it is sufficient to take $C > C_*^2$. □

For the analysis for the consistency and the approximation of stabilized interior penalty method (and other methods), we referred to [3].
In Table 2.2 are summarized some properties of the most important methods previuosly analysed: consistency, stability, theoretical requirement on $c_0 = \inf_e c$ for stability and rates of convergences on H^1 and in L^2 norms.

Another important method that ensures the local conservation properties is the finite volume element method which is the object of the next section. We will give a short overview of this method.

2.3 The finite volume-element (FVE) method

Introduced in [88], the finite volume-element method is a discretization technique for partial differential equations posed in divergence form. After having partitioned the domain Ω in a finite set of volumes, it uses in each volume an integral formulation of the problem restricting the approximating functions to a finite element space. The FVE is closely related to the so-called control volume finite element method (CVFEM) introduced at the beginning of the eighties in mechanical engineering literature [19] (tailored especially to composite grid applications). Initially the FVE was developed to provide an effective discretization scheme in the context of multilevel adaptative methods based on local and global uniform grids [96]. We note that these methods are especially well suited for use in the framework of fluid flows because the resulting stencils are simple, in fact a local subproblem defined on a local grid can be treated as if it was separated from the the original problem using suitables boundary conditions. As a combination of the FE and the FV methods, the FVE uses the flexibility of the FE techniques and the conservative properties of the FV methods. Therefore, the two basic choices for FVE are the finite element space $\mathcal{S}^h$ and the finite set of volumes $\mathcal{V}^h$. There are two different possibilities to build $\mathcal{V}^h$ volumes. In the first one, after having chosen some points (the futures vertexes of the triangles) inside Ω and on $\partial\Omega$ and after having made a Delaunay triangulation, by these points, the related Voronoï tessellation [113] gives the desired partition of Ω. We remember that the boundary of each volume $\mathcal{V}^h$ is made by connecting with a straight line the circumcenter of the triangles having common vertex P and requiring that each edge of the polygonal boundary is orthogonal to an edge of a pair of adjacent triangles (see Figure 2.8). We note that these circumcenters do not lie outside the triangle by virtue of the assumption of the definition of the triangulation that no triangle has an internal angle greater that 90 degree. In the second approach of Donald type, in every triangle the centroid substitutes the circumcenter so that

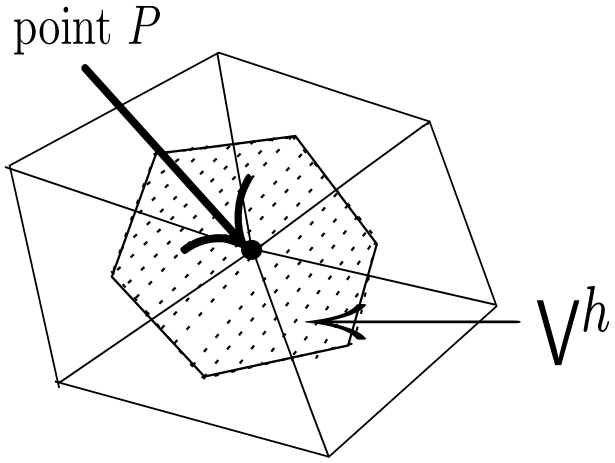

Figure 2.8: Example of control volume V^h of Voronoï type (shadow region) relevant to the point P ($\bullet$) .

the edges of the polygonal boundary are no more orthogonal to the triangle edges (see Figure 2.9). We referred to these books and papers [44, 29, 30] for more development and more details

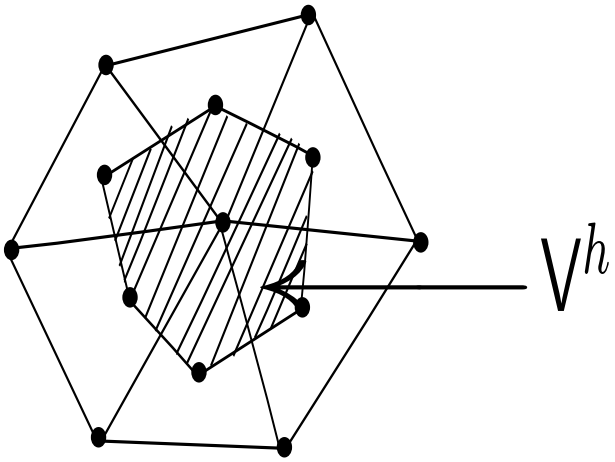

Figure 2.9: Example of control volume V^h of Donald type (shadow region) .

on the FVE method. We are now presenting a new conservative method for elliptic problems in computational fluids dynamics resembling the discontinuous Galerkin finite element method [69].

2.4 New conservative finite element method

As our final aim is to solve the 2D Navier-Stokes equations, the conservation of the fluxes is a very important property that the numerical scheme should satisfy. This property becomes fundamental if we think of a possible extension to 3D problems [38]. The key idea of this method is that we wish to apply conservation of the mathematical fluxes with respect to the original elements generated and not with respect to the dual mesh (that one composed by Voronoï or Donald polygonals). In order to present the method, we will assume considering the typical Poisson problem of finding u, solution of

$$\begin{cases} -\Delta u = f & \text{in} \quad \Omega \\ \quad u = 0 & \text{on} \quad \partial\Omega \end{cases} \tag{2.51}$$

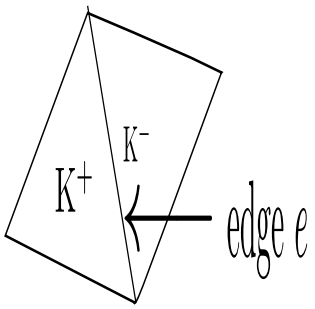

Figure 2.10: Pair of adjacent triangles with common edge e.

where Ω is a two dimensional domain with Lipschitz continuous boundary $\partial\Omega$. Let $\mathcal{I}_h$ be a suitable triangulation of Ω i.e $\mathcal{I}_h = \{K\}$.
We wish to find the solution u_h in the functional space $\mathrm{S}_h \subset \mathrm{H}^1(\Omega)$ and wish also the test functions to belong to the same functional space. Therefore, the weak formulation consists to find $u_h \in \mathrm{S}_h$ such that

$$a(u_h, s_h) = (f, s_h) \quad \forall\ s_h \in \mathrm{S}_h \tag{2.52}$$

where

$$a(u_h, s_h) = \sum_K (\nabla u_h, \nabla s_h)_K - \sum_{\partial K} (\frac{\partial u_h}{\partial n}, s_h)_{\partial K}$$

,

$$(f, s_h) = \sum_K \int_K f s_h dK$$

and $(\cdot,\cdot)_K$ denotes the inner scalar product.
Keeping in mind the local conservation of the mathematical fluxes at the edges of the K elements, we are led into looking for a solution $u_h \in \mathrm{S}_h$ to the following system

$$\sum_{K\in\mathcal{I}_h} \int_K \nabla u_h \cdot \nabla s_h dK - \sum_{e\in\partial K} \int_e \frac{\partial u_h}{\partial n} s_h|_K de = \sum_{K\in\mathcal{I}_h} \int_K f s_h dK, \quad \forall\ s_h \in \mathrm{S}_h \tag{2.53}$$

$$\sum_{e\in\partial K} (q\cdot n|_{K^+} + q\cdot n|_{K^-}, s_h|_{K^+})_e = 0 \quad \forall\, q \in \mathrm{H}(div, K) \tag{2.54}$$

where K^+ and K^- are adjacent triangle, see Figure (2.10) and $\mathrm{H}(div, K) = \{v \in \mathrm{L}^2(K) \mid \ div\, v \in \mathrm{L}^2(K)\}$. For more specifications of this functional space, we referred to [87].

Remark 2.23. *The equations (2.53) can be interpreted like a hybrid saddle point approach see Quarteroni-Valli [113].*

Remark 2.24. *The equations (2.54) enforces the flux conservation on inter-element edges. This formulation differs from the hybrid method by the conservation property and also by the fact that we are seeking only the function u_h, while the hybrid method is mainly interested on the function u_h and the gradient vector ∇u_h.*

Remark 2.25. *The space $\mathrm{H}^1(\Omega)$ does not take into account that the assigned Dirichlet boundary conditions are homogeneous. Actually this choices makes possible the approach (2.54) that guarantees the conservation property of mathematical fluxes, also for other kind of boundary conditions.*

Remark 2.26. *If we use for the u_h approximation the space X_h^2 (i.e. 2D polynomials of degree two), then the continuity of the mathematical fluxes (represented in (2.52) by the Neumann term $\frac{\partial u_h}{\partial n}|_{\partial K}$) on the ∂K edges internal to Ω, can be imposed. This continuity property allows a good approximation of the numerical fluxes (see section 2.4.1), an easy imposition of the conservation property and a consistent assignment of all kinds of boundary conditions.*

2.4.1 Algebraic description

In order to describe the structure of the algebraic system arising from the discretization of system (2.53)-(2.54) we have considered the standard Galerkin finite element formulation by taking basis functions φ as traditional $2D$ polynomials of second degree with six nodes on the boundary (see Figure 2.11) and considering the macro elements (see Figure 2.12).

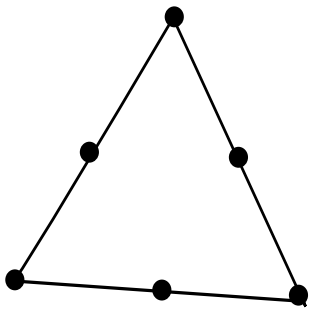

Figure 2.11: Element with six nodes ($2D$ second degree polynomial) .

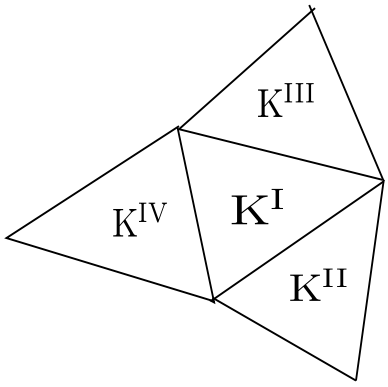

Figure 2.12: A macroelement.

In fact, for each pair of elements of the macroelement (2.53) is written as

$$\sum_K [\, \nabla u^I \cdot \nabla \varphi^I \, dK^I \, - \int_{\partial K^I} (\frac{\partial u}{\partial n})^I \varphi^I \, d\partial K^I] = \sum_{K^I} \int_{K^I} f \varphi^I \, dK^I \quad \forall \, \varphi$$

i.e

$$\sum_K [\, \sum_{i=1}^6 u_i^I \int_{K^I} \nabla \varphi_i^I \cdot \nabla \varphi_j^I, dK^I \, - \sum_{i=1}^6 u_i^I \int_{\partial K^I} (\frac{\partial \varphi_i^I}{\partial x} \cdot n_x^I + \frac{\partial \varphi_i^I}{\partial y} \cdot n_y^I) \varphi_j^I \, d\, \partial K^I \,] =$$

$$\sum_K [\, \sum_{i=1}^6 f_i^I \int_{K^I} \varphi_i^I \varphi_j^I \, dK^I \,], \quad \forall \, j = 1, \cdots, 6$$

i.e.,

$$\sum_K [\, \sum_{i=1}^6 u_i^I \int_{K^I} \nabla \varphi_i^I \cdot \nabla \varphi_j^I \, d\, K^I \, + \sum_{i=1}^6 u_i^{II} \int_{\partial K^I} (\frac{\partial \varphi_i^{II}}{\partial x} \cdot n_x^{II} + \frac{\partial \varphi_i^{II}}{\partial y} \cdot n_y^{II}) \varphi_j^I \, d\, \partial K^I \,] =$$

$$\sum_K [\, \sum_{i=1}^6 f_i^I \int_{K^I} \varphi_i^I \varphi_j^I \, dK^I \,], \quad \forall \, j = 1, \cdots, 6$$

(by applying the local conservation properties (2.53)), thus

$$\underline{\mathrm{A}}_K \underline{u} + \underline{\mathrm{B}}_{\partial K} \underline{u} = \mathrm{M}_K \underline{f} \, . \tag{2.55}$$

In (2.55),

$$\underline{\mathrm{A}}_K := \sum_K \int_K \nabla\varphi_i \cdot \nabla\varphi_j \, dK \quad \text{is the stiffness matrix},$$

$$\underline{\mathrm{M}}_K := \sum_K \int_K \varphi_i \varphi_j \, dK \quad \text{is the mass matrix},$$

and

$$\underline{\mathrm{B}}_{\partial K} := \sum_K \int_{\partial K^I} (\frac{\partial \varphi_i^{II}}{\partial x} \cdot n_x^{II} + \frac{\partial \varphi_i^{II}}{\partial y} \cdot n_y^{II}) \varphi_j^I \, d\partial K^I$$

is the final matrix obtained by assembling each 6×9 local matrix of each pair of elements resulting from the term $\int_{\partial K^I} (\frac{\partial \varphi_i^{II}}{\partial x} \cdot n_x^{II} + \frac{\partial \varphi_i^{II}}{\partial y} \cdot n_y^{II}) \varphi_j^I \, d\partial K^I$ due to the application of the conservation of the mathematical fluxes. $\underline{n} = (n_x, n_y)$ are the components of the outward normal at the edges of each element considered.
The matrices $\underline{\mathrm{A}}_K$, $\underline{\mathrm{M}}_K$ and $\underline{\mathrm{B}}_{\partial K}$ are computed using the Gauss quadrature formula in the reference element evaluated at seven integration points (vertices, mid-edges and centroid points). The transformation from the general triangle to the reference element is an isoparametric one; this choice of course is not compulsory in particular if we use triangles with straight edges (like those used in the numerical tests). The assembling of local matrices can be easily understood if one considers the macro element formed by a triangle K^I and its three adjacent triangles K^{II}, K^{III}, K^{IV}. Each central element of the macro element has a contribution due to the local Neumann term of a 6×9 local matrix repeated three times corresponding to its three edges. Performing this strategy element by element, we obtain a balance of the number of equations with respect to the number of unknowns. Denoting by N_h the total number of degrees of freedom in the triangulation $\mathcal{I}_h$, the system of equations (2.53)-(2.54) can be written as

$$\mathrm{A}_h \underline{u}_h = \underline{\mathrm{b}}_h \tag{2.56}$$

where $\mathrm{A}_h = \underline{\mathrm{A}}_K + \underline{\mathrm{B}}_{\partial K}$ is the global non singular and non symmetric stiffness matrix, $\underline{u}_h$ the point value of the numerical solution at the nodes and $\underline{\mathrm{b}}_h = \mathrm{M}_K \underline{f}$ is the suitable final right hand side. We note that the final global matrix A_h and the final right hand side $\underline{\mathrm{b}}_h$ does not take into account boundary conditions. In section 2.4.2, we explain the insertion of the boundary conditions.

Remark 2.27. *It is the matrix* $\underline{\mathrm{B}}_{\partial K}$ *that leads to the non-symmetry of the global matrix* A_h

The algebraic system (2.56) can be solved by a Gauss direct solver or by a Bi-CGSTAB iterative solver.

2.4.2 Treatment of boundary conditions

Dirichlet boundary conditions

The Dirichlet boundary conditions are treated by means of this process: after constructing the global matrix A_h that involves all the nodes of the domain, put 1 in the diagonal position of all the nodes belonging to the Dirichlet boundary and 0 in the remaining position in the same rows, then assign the Dirichlet values in the corresponding positions of the right hand side vector b_h.

Neumann boundary conditions

Let e_N be an edge that belongs to the Neumann boundary and $\frac{\partial u}{\partial n}|_{node1}$, $\frac{\partial u}{\partial n}|_{node2}$ $\frac{\partial u}{\partial n}|_{node3}$ the three Neumann values at the three nodes of the edge e_N (two vertices and their mid-edge). We make this approximations:

$$\int_{e_N} \frac{\partial u}{\partial n}\varphi \, de_N \cong \int_{e_N} (\frac{\partial u}{\partial n}|_{node1} \times h_1(x,y) + \frac{\partial u}{\partial n}|node2 \times h_2(x,y) + \frac{\partial u}{\partial n}|_{node3} \times h_3(x,y))\varphi \, de_N \quad (2.57)$$

where $h_1(x,y)$, $h_2(x,y)$, $h_3(x,y)$ are suitable 2nd degree polynomials passing through the nodes of edge e_N; then we evaluate the integral (2.57) in the segment $[-1,1]$ by a seven points Gauss formula. The results of this integral is a local vector of six positions corresponding to the six test functions that have to be added in the right positions of the final right hand side vector b_h.

Remark 2.28. *A suitable code has been written (in Fortran 90) in order to verify the correctness and efficiency of our numerical approach. More information about the code will given later.*

2.4.3 Numerical results

In order to check the accuracy of this method and the correctness of the computer code, we chose the analytical solution of the Poisson problem (2.51) given by

$$u(x,y) = xy(1-x)(1-y)$$
$$f(x,y) = 2(x - x^2 + y - y^2)$$

All the computations were performed in $\Omega = [0,1]^2$. We computed the L^∞ and L^2 error norms generated by the new conservative approach and by a traditional finite element approach. Some comparisons among these norms are in the Tables 2.3 and 2.4.
Two different kinds of boundary conditions have been considered. The homogeneous Dirichlet boundary conditions for the Test 2.4.3.1 and the mixed boundary conditions (Dirichlet on the vertical sides and Neumann on horizontal sides) for the Test 2.4.3.2. We use a Gauss direct solver for the solution of the algebraic system (2.56). In the following Tables, as aforementioned, N_h denotes the total number of nodes with respect to each domain considered. Four different discretizations have been considered.

Test 2.4.3.1

Table 2.3: Homogeneous Dirichlet boundary condition

	Conservative FE		Traditional FE	
N_h	L^∞-norm	L^2-norm	L^∞-norm	L^2-norm
49	$1.431E-2$	$1.561E-2$	$3.720E-4$	$7.119E-4$
81	$7.073E-3$	$1.210E-2$	$3.855E-4$	$6.269E-4$
137	$5.186E-3$	$4.080E-3$	$8.120E-5$	$1.945E-4$
169	$5.687E-3$	$7.610E-3$	$5.230E-5$	$1.498E-4$

Test 2.4.3.2

Table 2.4: Mixed boundary conditions

	Conservative FE		Traditional FE	
N_h	L^∞-norm	L^2-norm	L^∞-norm	L^2-norm
49	$2.095E-2$	$2.585E-2$	$7.949E-3$	$1.406E-2$
81	$7.932E-3$	$1.509E-2$	$6.734E-3$	$1.675E-2$
137	$6.563E-3$	$1.460E-2$	$5.903E-3$	$8.170E-3$
169	$6.461E-3$	$9.297E-3$	$8.832E-4$	$1.059E-3$

2.5 Conclusions of the chapter

Checking Tables 2.3 and 2.4 and making the comparison with the error norms, it appears that the conservative finite element method is convergent but, unfortunately, the new scheme is not conservative according to the classical definition; however it could be generalized so that a scheme genuinely conservative, like those named finite volume-elements, could be obtained. We have verified heuristically the conservation property by solving some numerical tests. We underline that the conservation property of the fluxes is respected by construction and that a 2D FE able to guarantee the continuity of the fluxes is the eighteen quintic $\mathcal{C}^1$ finite element.

If the algebraic system (2.56) is not preconditioned, its solution could be time consuming and since we are aiming to solve the 2D Navier-Stokes equations, an optimal sophisticated preconditioner is recommended. We choose the additive Schwarz overlapping domain decompositions approach as preconditioner for the algebraic system. This is the object of the next chapter.

Chapter 3

Domain decomposition methods

3.1 Introduction

Historically they were used for solving problems with PDEs defined on domains of general form and had a renewed interest with parallel computing (in particular for distributed memory computers). The idea on which they are based is the subdivision of the original problem defined in Ω in some subproblems, each defined in Ω_i, such that $\bigcup_i \Omega_i = \Omega$ and $\bigcup_i u_i = u$. The most important properties of the methods are:

- The flexibility that allows the solution of the original problem taking into account also the eventual local physical situations
- The possibility to solve problems of dimension smaller than that of the original one reducing the computing time
- The possibility to use iterative solvers for the solutions of the algebraic systems using multi-domains techniques in order to generate preconditioning (in particular with scalable processes).

3.2 Multi-domain techniques

In the literature, there are three fundamental paradigms for multi-domain techniques :

3.2.1 Fictitious domain (or domain embedding) methods

This is one of the earliest ideas closely related to multi-domain. The leading motivation is that whenever a problem needs to be solved on domain Ω having complicated boundary, it may be useful to embed it into a larger domain Ω' of simpler shape, say for instance a rectangle, then solving a problem of similar type therein (see Figure 3.1). For the theory of the method and other developments, we suggest these references [13, 70, 91, 92, 94].

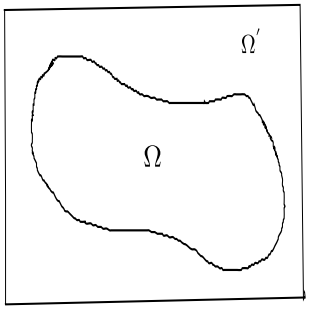

Figure 3.1: Fictitious domain.

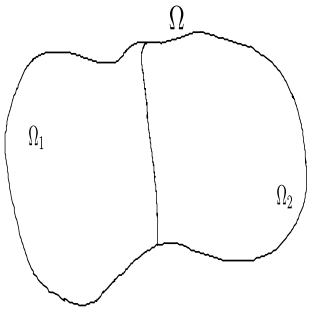

Figure 3.2: Disjoint partition.

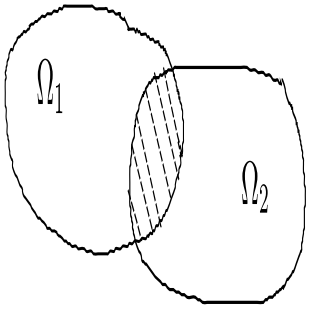

Figure 3.3: Overlapping sub-domain.

3.2.2 Disjoint partitions

In this case the domain Ω is partitioned into two non-overlapping sub-domains Ω_1 and Ω_2 see Figure 3.2. This multi-domain technique is based on the iteration by sub-domain methods based on transmission conditions at the interface, with various interface conditions like: Dirichlet-Neumann, Neumann-Neumann or Robin ones. An application of this technique to the generalized Stokes problem will be seen in section 3.7.1.

3.2.3 Overlapping sub-domains

The Schwarz approach is undoubtedly the earliest example of a domain decomposition approach for partial differential equations. It was introduced in 1869 by Schwarz [125]; however, among others, let us also mention the early contributions of Sobolev [127], Mikhlin [93] and Matsokin and Nepomnyaschikh [95]. In this case the domain Ω is subdivided in two or more overlapping sub-domains Ω_1 and Ω_2 see Figure 3.3. For some reasons that will be explained further, we will use the overlapping Schwarz method as our multi-domain technique. Other material can be found in the following books [111, 114] and papers [107, 112].

3.3 Usefulness of multi-domain techniques

Multi-domain techniques are very useful for some important reasons including the ability to deal with problems with a partial differential equation, with a system of PDEs, with different PDEs in different regions of the domain Ω. This is the case for example of the viscous-inviscid flow interactions in boundary layers, molecular-continuous state of flow in the upper atmosphere, etc. They allow also the use of various kinds of spatial approximation methods, for example the finite difference method, the finite element method, the finite volume method and the spectral collocation method. The next section recalls some important tools and objects we will regularly deal with.

3.4 Solution of algebraic systems

3.4.1 Direct and iterative methods

Direct methods

Given an $n \times n$ real matrix A and a real n-vector $\underline{b}$, the problem considered consists to find:

$$\underline{x} \quad \text{belonging to } \mathbb{R}^n \quad \text{such that} \quad A\underline{x} = \underline{b} \tag{3.1}$$

Equation (3.1) is a linear system, A is the coefficient matrix, b is the right-hand side vector and $\underline{x}$ is the unknowns vector. Then, the existence and uniqueness of the solution to (3.1) is ensured if one of the following conditions holds:

1. A is invertible
2. $rank(A) = n$
3. The homogeneous system $A\underline{x} = \underline{0}$ admits only the null solution.

The solution of (3.1) is formally provided by Cramer's rule

$$x_j = \frac{\Delta_j}{\det(A)}, \quad j = 1, 2, \cdots, n \tag{3.2}$$

where Δ_j is the determinant of the matrix obtained by substituting the $j-th$ column of A with the right hand side $\underline{b}$. But this formula is of little practical use because, if the determinants are evaluated by the following recursive relation

$$\det(A) = \begin{cases} a_{11} & \text{if} \quad n = 1 \\ \sum_{j=1}^{n} \Delta_{ij} a_{ij} & \text{for} \quad n > 1 \end{cases} \tag{3.3}$$

known as the Laplace rule, the computational effort of Cramer's rule is of order $(n+1)!$ flops and therefore turns out to be unacceptable even for small dimensions of A (a computer able to perform 10^9 flops per second would take 9.6×10^{47} years to solve a linear system of only 50 equations). This is why alternatives to Cramer's rule have been developed. They are called direct methods if they yield the solution of the system in a finite number of steps and iterative methods if they require (theoretically) an infinite number of steps. Iterative methods will be addressed in the next sub-subsection. Solving a linear system by a numerical method invariably leads to the introduction of rounding errors. Only stable methods can keep away the propagation of such errors from polluting the accuracy of the solution. Informations relevant to this can be given by the condition number of a matrix $A \in \mathbb{C}^{n \times n}$ defined as

$$\mathcal{K}(A) = \|A\|\|A^{-1}\| \tag{3.4}$$

where $\|\cdot\|$ is an induced matrix norm.
We have these definitions.

Definition 3.1. *Let* A *be a square matrix of order* n *of real or complex entries: the number* $\lambda \in \mathbb{C}$ *is called an eigenvalue of* A *if there exits a non null vector* $\underline{x} \in \mathbb{C}^n$ *such that* $A\underline{x} = \lambda\underline{x}$. *The vector* $\underline{x}$ *is the eigenvector associated with the eigenvalue* λ *and the set of the eigenvalues of* A *is called the spectrum of* A, *denotes by* $\sigma(A)$. *The maximum module of the eigenvalues of* A *is called the spectral radius of* A *and is denoted by*

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

In general, $\mathcal{K}(\mathrm{A})$ depends on the choice of the norm; if $\mathcal{K}_p(\mathrm{A})$ denotes the condition number of A in the $p-$ norm, then for $p = 2$, it can be proven that

$$\mathcal{K}_2(\mathrm{A}) = \|\mathrm{A}\|_2\|\mathrm{A}^{-1}\|_2 = \frac{\sigma_1(A)}{\sigma_n(A)} \tag{3.5}$$

where $\sigma_1(A)$(resp $\sigma_n(A)$) are the maximum (resp the minimum) singular values of A. As a consequence of relation 3.5, in the case the matrices are symmetric and positives definite, we have

$$\mathcal{K}_2(\mathrm{A}) = \frac{\lambda_{max}}{\lambda_{min}} = \rho(\mathrm{A})\rho(\mathrm{A}^{-1})$$

and λ_{max}(resp $\lambda_{min}(A)$)are the maximum (resp minimum) eigenvalues of A. An increase in the condition number produces a higher sensitivity of the solution of the linear system to change in the data. Due to rounding errors, a numerical method for solving (3.1) does not provide the exact solution but only an approximate one which satisfies a perturbed system; this means that a numerical method yields an exact solution $x + \delta x$ of the perturbed system

$$(\mathrm{A} + \delta\mathrm{A})(\underline{x} + \delta\underline{x}) = \underline{\mathrm{b}} + \delta\underline{\mathrm{b}} \tag{3.6}$$

The most popular direct methods are the Gaussian elimination method (GEM) and the Lower-upper (LU) factorization. The Gaussian elimination method aims at reducing the system (3.1) to an equivalent system (that is having the same solution) of the form $\mathrm{U}\underline{x} = \widehat{\underline{\mathrm{b}}}$, where U is an upper triangular matrix and $\widehat{\underline{\mathrm{b}}}$ is an updated of right side vector. The latter system can then be solved by the backward substitution process. Moreover, the GEM is equivalent to performing a factorization of the matrix A into a product of two matrices $\mathrm{A} = \mathrm{LU}$ with $\mathrm{U} = \mathrm{A}^{(n)}$ where $\mathrm{A}^{(n)}$ is the $n - th$ transformation of the original matrix. The computational effort, about $\frac{2n^3}{3}$ flops, is spent overall in the elimination procedure. A thorough discussion on this method can be found in the books [111, 123].

Iterative methods

The basic idea of iterative methods is to construct a sequence of vectors $\underline{x}^{(k)}$ that enjoy the property of convergence

$$\underline{x} = \lim_{k\to\infty} \underline{x}^k \tag{3.7}$$

and $\underline{x}$ the solution to (3.1). In practice, one of the stopping criteria of the iterative process is the minimum value of n such that $\|\underline{x}^n - \underline{x}\| < \epsilon$, where ϵ is a fixed tolerance and $\|\cdot\|$ is any convenient vector norm. To start with, we consider an iterative method of the form:

$$\text{given } \underline{x}^{(0)}, \quad \underline{x}^{k+1} = \mathrm{B}\underline{x}^k + \underline{\mathrm{f}}, k \geq 0 \tag{3.8}$$

where B is an $n \times n$ square matrix called the iteration matrix and $\underline{\mathrm{f}}$ a vector depending on the right-hand side and $\underline{x}^0$ the initial guess.

Definition 3.2. *An iterative method of the form (3.8) is said to be consistent with (3.1) if* $\underline{\mathrm{f}}$ *and* B *are such that* $\underline{x} = \mathrm{B}\underline{x} + \underline{\mathrm{f}}$ *equivalently,* $\underline{\mathrm{f}} = (\mathrm{I} - \mathrm{B})\mathrm{A}^{-1}\underline{\mathrm{b}}$.

Theorem 3.3. *Let (3.8) be a consistent method. Then the sequence of vector* $\underline{x}^{(k)}$ *converges to the solution of (3.1) for any choice of* $\underline{x}^{(0)}$ *if and only if* $\rho(\mathrm{B}) < 1$.

A general technique to devise consistent iterative methods is based on an additive splitting of the matrix A of the form $\mathrm{A} = \mathrm{P} - \mathrm{N}$ where P and N are two suitable matrices and P is non singular. Later we will see that P is called preconditioning matrix or preconditioner. Precisely, given $\underline{x}^{(0)}$, one can compute $\underline{x}^{(k)}$ for $k \geq 1$, solving the system

$$\mathrm{P}\underline{x}^{(k+1)} = \mathrm{N}\underline{x}^{(k)} + \underline{\mathrm{b}}, \quad k \geq 0 \tag{3.9}$$

The iteration matrix of method (3.9) is $\mathrm{B} = \mathrm{P}^{-1}\mathrm{N}$, while $\underline{\mathrm{f}} = \mathrm{P}^{-1}\underline{\mathrm{b}}$. Alternatively (3.9) can be written in the form

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \mathrm{P}^{-1}\underline{r}^{(k)}, \tag{3.10}$$

where

$$\underline{r}^{(k)} = \underline{\mathrm{b}} - \mathrm{A}\underline{x}^{(k)}, \tag{3.11}$$

denotes the residual vector at step k.
The Jacobi and Gauss-Seidel iterations are both of the form $\mathrm{P}\underline{x}^{(k+1)} = \mathrm{N}\underline{x}^{(k)} + \underline{\mathrm{b}} = (\mathrm{P} - \mathrm{A})\underline{x}^{(k)} + \underline{\mathrm{b}}$ in which $\mathrm{A} = \mathrm{P} - \mathrm{N}$ is the splitting of A, $\mathrm{P} = \mathrm{D}$ for the Jacobi and $\mathrm{P} = \mathrm{D} - \mathrm{E}$ for the Gauss-Seidel, where D is the diagonal counterpart of A and E being the lower triangular counterpart of A. The iteration matrix of the Jacobi method is given by $\mathrm{B}_J = \mathrm{I} - \mathrm{D}^{-1}\mathrm{A}$. In the Gauss-Seidel method, the associated iteration matrix is $\mathrm{B}_{\mathrm{GS}} = (\mathrm{D} - \mathrm{E})^{-1}\mathrm{F}$ in which F is the upper triangular counterpart of the original matrix A. Another kind of iterative method is the stationary Richardson method. It is defined as :

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha\mathrm{P}^{-1}\underline{r}^{(k)},\ k \geq 0 \tag{3.12}$$

α being a relaxation (or acceleration) parameter. The iteration matrix in this case is $\mathrm{B}_R(\alpha) = \mathrm{I} - \alpha\mathrm{P}^{-1}\mathrm{A}$. We note that the Jacobi and Gauss-Seidel methods can be regarded as stationary Richardson methods with P taking the same value as previously mentioned and $\alpha = 1$ in both cases.

3.5 Convergence estimates

The following results from [111] give some convergence estimates of previously mentioned methods.

Theorem 3.4. *For any nonsingular matrix* P*, the Richardson method (3.12) converges if and only if*

$$\frac{2\Re_e\lambda_i}{\alpha|\lambda_i|^2} > 1, \quad \forall i = 1, \cdots, n \tag{3.13}$$

where $\lambda_i \in \mathbb{C}$ *are the eigenvalues of* $\mathrm{P}^{-1}\mathrm{A}$ *.*

Theorem 3.5. *Assume* P *is a nonsingular matrix and that* $\mathrm{P}^{-1}\mathrm{A}$ *has positive real eigenvalues, ordered in such a way that* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$*. Then, the stationary Richardson method (3.12) is convergent if and only if* $0 < \alpha < \frac{2}{\lambda_1}$*. Moreover, letting*

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n}$$

the spectral radius of the iteration matrix $\mathrm{B}_R(\alpha)$ *is minimum if* $\alpha = \alpha_{opt}$*, with*

$$\rho_{opt} = \min_{\alpha}[\rho(\mathrm{B}_R(\alpha))] = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \quad .$$

3.5.1 Preconditioner

Lack of robustness is a widely recognized weakness of the traditional iterative solvers. This drawback alters the acceptance of iterative methods in industrial application despite their intrinsic appeal for very large linear systems. Both the efficiency and the robustness of iterative techniques can be improved by using preconditioning. Introduced in the Section 3.4.1, roughly speaking, a preconditioner is any form of implicit or explicit modification of the original linear system which make it "easier"to solve by a given iterative method. For example, scaling all rows of a linear system to make the diagonal elements equal to one is an explicit form of

preconditioning. The resulting system can be solved by a Krylov subspace method [120] and may require fewer steps to converge than that of the original system (although this is not guaranteed). All the methods we have seen before can be written (cast) in the form (3.8) so that they can be regarded as being methods for solving the system

$$(\mathrm{I} - \mathrm{B})\underline{x} = \underline{\mathrm{f}} = \mathrm{P}^{-1}\underline{\mathrm{b}}. \tag{3.14}$$

On the other hand since $\mathrm{B} = \mathrm{P}^{-1}\mathrm{N}$, system (3.1) can be reformulated as

$$\mathrm{P}^{-1}\mathrm{A}\underline{x} = \mathrm{P}^{-1}\underline{\mathrm{b}}. \tag{3.15}$$

(3.15) being the preconditioned system and P being the preconditioning matrix or left preconditioner. There are point preconditioners and block preconditioners depending on whether they are applied to the single entries of A or to a blocks of a partition of A. Since the preconditioner acts on the spectral radius of the iteration matrix, it is useful to pick-up for a given linear system, an optimal preconditioner; this could be defined as a preconditioner which is able to make the number of iterations required for convergence independent of the size of the system and also, able to make preconditioning operations inexpensive to apply to an arbitrary vector. In order to fix the ideas, in what follows, it is convenient to consider that the problem at hands is solved by means of a FE method.

3.5.2 Restriction and prolongation operators

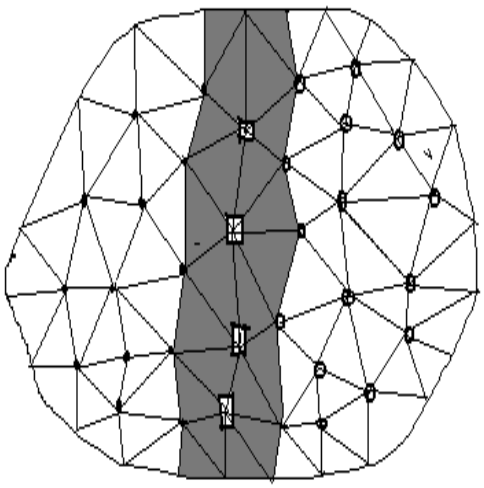

Figure 3.4: Overlapping subdivision of the domain Ω with more than one layer of overlap.

We assume the domain Ω is subdivided into two sub-domains Ω_1 and Ω_2. Let I_1 and I_2 be the indices of the nodes in the interior of Ω_1 and Ω_2 respectively. Obviously, if N_h is the number of internal nodes of Ω and I the set of all indices from 1 to N_h, then we have that I_1 and I_2 form an overlapping subdivision of I, i.e $I_1 \cup I_2 = I$, $I_1 \cap I_2 \neq \emptyset$. If n_1 and n_2 indicate the number in I_1 and I_2 respectively, due to overlap $n_1 + n_2 > N_h$. Now, order the indices in

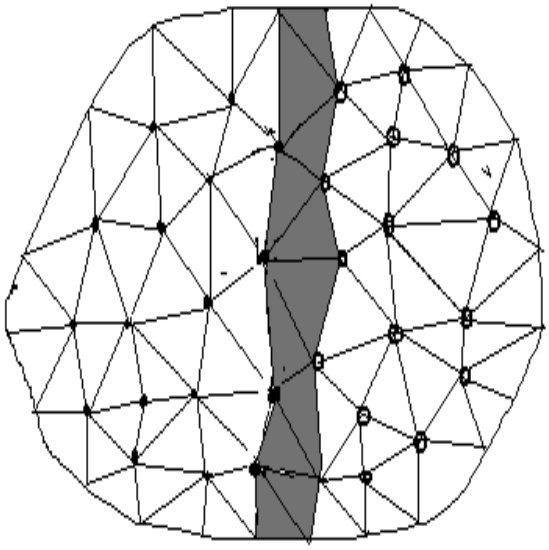

Figure 3.5: Overlapping subdivision of the domain with only one layer of overlap.

such a way that those corresponding to the nodes internal to Ω_1 but not internal to Ω_2 are first memorized and then take the remaining ones. Let A_1 and A_2 denote the principal sub-matrices of A formed by the first n_1 rows and columns and the last n_2 rows and columns, respectively see Figure 3.6. Then A_1 is the stiffness matrix for the sub-domain Ω_1 and A_2 that of Ω_2. They

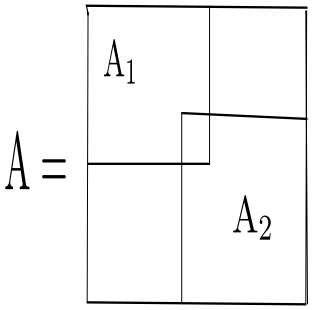

Figure 3.6: Subdivision of the stiffness matrix.

are related to the global stiffness matrix A of Ω by the algebraic relation

$$\mathrm{A}_1 = \mathrm{R}_1 \mathrm{A} \mathrm{R}_1^{\mathrm{T}}, \qquad \mathrm{A}_2 = \mathrm{R}_2 \mathrm{A} \mathrm{R}_2^{\mathrm{T}} \tag{3.16}$$

where $\mathrm{R}_i^{\mathrm{T}}$ and R_i, $i = 1, 2$ are the extension and the restriction matrices respectively. More precisely, $\mathrm{R}_i^{\mathrm{T}}$ is a $N_h \times n_i$ matrix whose action extends by 0 a vector of nodal values in Ω_i; that means that, given a sub-vector v^i of length n_i of nodal values, we have

$$(\mathrm{R}_i^{\mathrm{T}} v^i)_j = \begin{cases} v_j^i & \text{for } j \in I_i \\ 0 & \text{for } j \in I \setminus I_i \end{cases} \tag{3.17}$$

In other words, $\mathrm{R}_1^{\mathrm{T}}$ is a matrix in which n_1 rows and columns form the identity matrix, whereas the entries of the last $N_h - n_1$ rows are all 0 see Figure 3.7.

The transpose R_i of $\mathrm{R}_i^{\mathrm{T}}$ is a restriction matrix whose action restricts a vector v of dimension N_h to a vector of length n_i by preserving the entries with indice belonging to I_i. Thus finally, $\mathrm{R}_i v$ is the sub-vector of nodal values of v in the interior of Ω_i

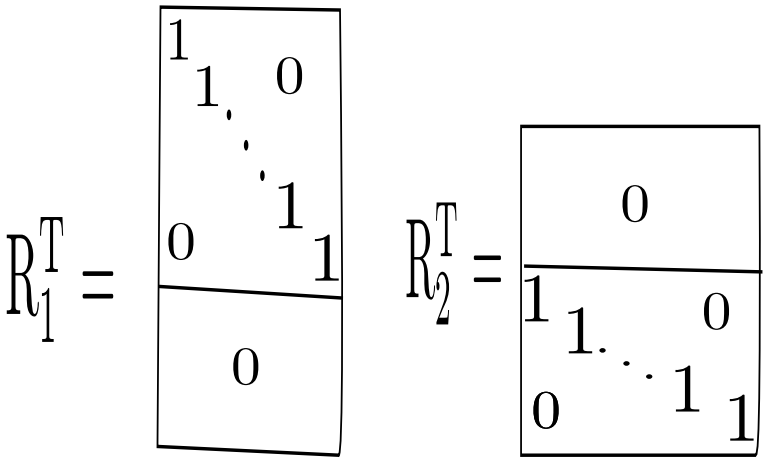

Figure 3.7: Extension matrices.

3.6 Solution of a Poisson equation

3.6.1 The Schwarz approach

To start, we consider the 2D domain Ω with a Lipschitz boundary $\partial\Omega$ as decomposed in two overlapping sub-domains Ω_1 and Ω_2 as shown for example in Figure 3.8. In Ω, we wish to solve by FE technique the linear elliptic PDE

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{3.18}$$

where f is a given function in $\mathrm{L}^2(\Omega)$, $\Delta := \sum_{j=1}^{2} D_j D_j$ is the Laplace operator and D_j

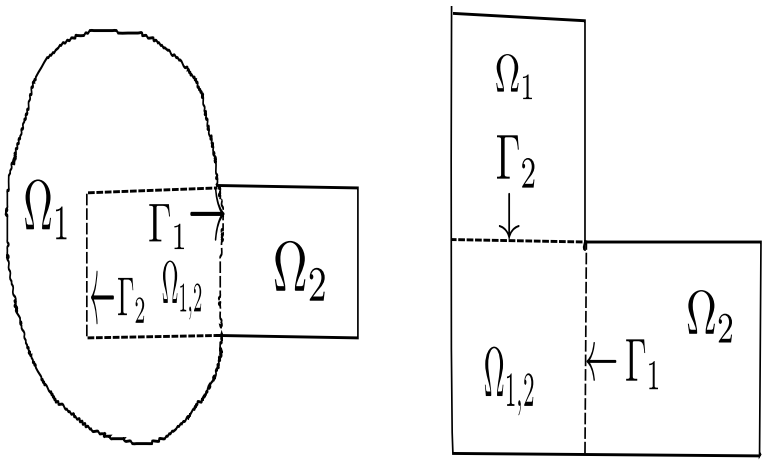

Figure 3.8: Example of overlapping subdivisions where $\Omega_{1,2} = \Omega_1 \cap \Omega_2$.

denotes the partial derivative with respect to x_j, $j = 1, 2$. Let note that Ω, Ω_1 and Ω_2 do not include their boundaries. Also let $\overline{\Omega} = \Omega \cup \partial\Omega$ denote the closure of the domain. The artificial boundaries Γ_i are the part of the Ω_i boundary that is interior to Ω (see Figure 3.8). The rest of the sub-domain boundaries are denoted by $\partial\Omega_i \setminus \Gamma_i$. We have $\Gamma_1 := \partial\Omega_1 \cap \Omega_2$, $\Gamma_2 := \partial\Omega_2 \cap \Omega_1$,

$\Omega_{1,2} := \Omega_1 \cap \Omega_2$ (see Figure 3.8). In order to describe the classical alternating (or the original form of) Schwarz method, we introduce the notations: u_i^k denotes the approximate solution on $\overline{\Omega}_i$ after k iterations and ${u_1}^k|_{\Gamma_2}$ is the restriction of u_1^k to Γ_2; similarly ${u_2}^k|_{\Gamma_1}$ is the restriction of u_2^k to Γ_1 (more precisely, since u_i^k are not necessarily continuous functions, ${u_i}^k|_{\Gamma_j}$ is the trace of u_i^k on Γ_j). Now let u^0 be an initialization function defined in Ω and vanishing on $\partial\Omega$ and set $\widehat{u}_2^0 := u^0|_{\Gamma_1}$. Therefore, the classical Schwarz approach consists to define two sequences $\widehat{u}_1^{k+1}$ and $\widehat{u}_2^{k+1}$ for $k \geq 0$ and to solve iteratively for $k = 0, 1, \cdots$, the boundary value problem

$$\begin{cases} -\Delta \widehat{u}_1^{k+1} = f \quad \text{in} \quad \Omega_1 \\ \quad \widehat{u}_1^{k+1} = \widehat{u}_2^k \quad \text{on} \quad \Gamma_1 \\ \quad \widehat{u}_1^{k+1} = 0 \quad \text{on} \quad \partial\Omega_1 \backslash \Gamma_1 \end{cases} \tag{3.19}$$

for $\widehat{u}_1^{k+1}$. This is followed by the solution of the boundary value problem

$$\begin{cases} -\Delta \widehat{u}_2^{k+1} = f \ \text{ in } \ \Omega_2 \\ \quad \widehat{u}_2^{k+1} = \widehat{u}_1^{k+1} \ \text{ on } \ \Gamma_2 \\ \quad \widehat{u}_2^{k+1} = 0 \ \text{ on } \ \partial\Omega_2 \backslash \Gamma_2 \end{cases} \tag{3.20}$$

for $\widehat{u}_2^{k+1}$. For the reasons that will be specified next, this approach is named of multiplicative kind.
On the other hand, if we set $\widehat{u}_1^0 := u^0|_{\Gamma_2}$ and $\widehat{u}_2^0 := u^0|_{\Gamma_1}$, we could make the two steps independent of each other by solving

$$\begin{cases} -\Delta \widehat{u}_1^{k+1} = f \quad \text{in} \quad \Omega_1 \\ \quad \widehat{u}_1^{k+1} = \widehat{u}_2^k \quad \text{on} \quad \Gamma_1 \\ \quad \widehat{u}_1^{k+1} = 0 \quad \text{on} \quad \partial\Omega_1 \backslash \Gamma_1 \end{cases} \tag{3.21}$$

and

$$\begin{cases} -\Delta \widehat{u}_2^{k+1} = f \quad \text{in} \quad \Omega_2 \\ \quad \widehat{u}_2^{k+1} = \widehat{u}_1^k \quad \text{on} \quad \Gamma_2 \\ \quad \widehat{u}_2^{k+1} = 0 \quad \text{on} \quad \partial\Omega_2 \backslash \Gamma_2 \end{cases} \tag{3.22}$$

Thus, in each half-step of the classical Schwarz method, we solve an elliptic boundary value problem on the sub-domain Ω_i with the given homogeneous boundary value on the true boundary $\partial\Omega_i \cap \partial\Omega$, and the previous approximate solution on the interior boundary Γ_i. This approach is called of additive kind.
The alternating Schwarz method (3.19)-(3.20) and (3.21)-(3.22) converges to the solution u of (3.18) provided some mild assumptions on the sub-domains Ω_1 and Ω_2 are satisfied. Precisely, there exist $C_1, C_2 \in (0, 1)$ such that for all $k \geq 0$,

$$\|u|_{\Omega_1} - \widehat{u}_1^{k+1}\|_{\mathrm{L}^\infty(\Omega_1)} \leq C_1 C_2 \|u - \widehat{u}^0\|_{\mathrm{L}^\infty(\Gamma_1)} \tag{3.23}$$

$$\|u|_{\Omega_2} - \widehat{u}_2^{k+1}\|_{\mathrm{L}^\infty(\Omega_2)} \leq C_1 C_2 \|u - \widehat{u}^0\|_{\mathrm{L}^\infty(\Gamma_2)} \tag{3.24}$$

The error reduction constants C_1 and C_2 can be quite close to one if the overlapping region $\Omega_{1,2}$ is thin. The proof of estimates (3.23),(3.24) can be obtained via the maximum principle (see for example Kantorovich and Krylov [72] and Lions [84]).

Weak formulation of the Schwarz approach

In the following, we formulate the Schwarz approach in a weak way. This requires the introduction of Sobolev spaces and to take into account their inherent properties that are very important in order to analyze the convergence behavior of the domain decomposition algorithms. We will

not dwell here on this argument and we refer the interested reader to the comprehensive presentation of this theory that can be found, for example, in Lions and Magenes [87]. By integrating by parts in Ω, the weak formulation of (3.18) consists to find $u \in \mathrm{V}$ such that

$$a(u,v) = (f,v), \qquad \forall v \in \mathrm{V} \tag{3.25}$$

with

$$\begin{aligned}
(f,v) &:= \int_\Omega fv \\
a(u,v) &:= (\nabla u, \nabla v) \\
\mathrm{H}^1(\Omega) &:= \{v \in \mathrm{L}^2(\Omega) \mid D_j v \in \mathrm{L}^2(\Omega), j = 1,2\} \\
\mathrm{H}^1_0(\Omega) &:= \{v \in \mathrm{H}^1(\Omega) \mid v|_{\partial\Omega} = 0\} \\
\mathrm{V} &:= \mathrm{H}^1_0(\Omega)
\end{aligned}$$

and $v|_{\partial\Omega}$ denotes the trace of v (that is its restriction) on $\partial\Omega$. The norm of $\mathrm{H}^1(\Omega)$ will be denoted by $\|\cdot\|_{1,\Omega}$ while $\|\cdot\|_{0,\Omega}$ will indicate the norm of $\mathrm{L}^2(\Omega)$. We recall that $\|v\|_{0,\Omega} = (v,v)^{\frac{1}{2}}$, while $\|v\|_{1,\Omega} = (\|v\|^2_{0,\Omega} + \sum_{j=1}^2 \|D_j v\|^2_{0,\Omega})^{\frac{1}{2}}$ for each $v \in \mathrm{H}^1(\Omega)$. The Poincaré inequality states that there exists a constant $C_\Omega > 0$ such that

$$\int_\Omega v^2 \le C_\Omega \int_\Omega \sum_{j=1}^2 (D_j v)^2 \quad \forall v \in \mathrm{H}^1_0(\Omega).$$

Therefore, the norm $\|v\|_{1,\Omega}$ is equivalent to the norm $\|\nabla v\|_{0,\Omega}$ for all $v \in \mathrm{H}^1_0(\Omega)$. We can not say that the same result is true for functions that vanish only on an open and non-empty subset Σ of $\partial\Omega$. We also recall that the trace space of $\mathrm{H}^1(\Omega)$ on the boundary $\partial\Omega$ is denoted by $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$. In an analogous way, the trace on an open and non-empty subset $\Sigma \subset \partial\Omega$ is indicated by $\mathrm{H}^{\frac{1}{2}}(\Sigma)$. The trace operator from $\mathrm{H}^1(\Omega)$ to $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$ is surjective and continuous; that is there exists C^*_Ω

$$\|v|_{\partial\Omega}\|_{\frac{1}{2},\partial\Omega} \le C^*_\Omega \|v\|_{1,\Omega}, \qquad \forall v \in \mathrm{H}^1(\Omega)$$

where $\|\cdot\|_{\frac{1}{2},\partial\Omega}$ denotes the norm in $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$. Finally, it can be shown that there exist injective, linear and continuous extension operators from $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$ to $\mathrm{H}^1(\Omega)$.
Now, the weak formulation of the Schwarz method for the homogeneous Dirichlet boundary value problem associated with it can be stated as follows : set as before, $\mathrm{V} := \mathrm{H}^1_0(\Omega)$, $\mathrm{V}^0_i := \mathrm{H}^1_0(\Omega_i \setminus \Gamma_i)$, $i = 1,2$, and $V^*_1 = \Omega \setminus \overline{\Omega}_1$ and $V^*_2 = \Omega \setminus \Omega_2$ respectively.

Therefore, (3.19), (3.20) reads : given $u^0 \in \mathrm{V}$, solve for each $k \ge 0$

$$\begin{aligned}
w_1^k \in \mathrm{V}^0_1 \; &: a_1({w_1}^k, v_1) = (f,v_1)_{\Omega_1} - a_1(u^k, v_1), \quad \forall v_1 \in \mathrm{V}^0_1 \\
u^{k+\frac{1}{2}} &= u^k + \widetilde{w_1^k} \\
&\text{and} \\
w_2^k \in \mathrm{V}^0_2 \; &: a_2(w_2^k, v_2) = (f,v_2)_{\Omega_2} - a_2(u^{k+\frac{1}{2}}, v_2), \quad \forall v_2 \in \mathrm{V}^0_2 \\
u^{k+1} &= u^{k+\frac{1}{2}} + \widetilde{w_2^k}
\end{aligned} \tag{3.26}$$

where $\widetilde{w_i^k}$ denotes the extension of w_i^k by 0 in V^*_i. Similarly, method (3.21),(3.22) is obtained by solving

$$\begin{aligned}
w_1^k \in \mathrm{V}^0_1 \; &: a_1({w_1}^k, v_1) = (f,v_1)_{\Omega_1} - a_1(u^k, v_1), \quad \forall v_1 \in \mathrm{V}^0_1 \\
w_2^k \in \mathrm{V}^0_2 \; &: a_2(w_2^k, v_2) = (f,v_2)_{\Omega_2} - a_2(u^k, v_2), \quad \forall v_2 \in \mathrm{V}^0_2 \\
u^{k+1} &= u^k + \widetilde{w_1^k} + \widetilde{w_2^k}
\end{aligned} \tag{3.27}$$

The proof that these weak formulations are equivalent to the original ones (3.19),(3.20) and (3.21),(3.22) is obtained via the verification of the following relations :

$$u^{k+\frac{1}{2}} = \begin{cases} \widehat{u}_1^{k+1} & \text{in} \quad \Omega_1 \\ \widehat{u}_2^{k} & \text{in} \quad \Omega \setminus \Omega_1 \end{cases} \quad , \quad u^{k+1} = \begin{cases} \widehat{u}_2^{k+1} & \text{in} \quad \Omega_2 \\ \widehat{u}_1^{k+1} & \text{in } \Omega \setminus \Omega_2 \end{cases} \tag{3.28}$$

and

$$u^{k+1} = \begin{cases} \widehat{u}_1^{k+1} & \text{in} \quad \Omega \setminus \Omega_2 \\ \widehat{u}_1^{k+1} + \widehat{u}_2^{k+1} - \widehat{u}^{k} & \text{in} \quad \Omega_{1,2} \\ \widehat{u}_2^{k+1} & \text{in} \quad \Omega \setminus \Omega_1 \end{cases} \tag{3.29}$$

The Schwarz method as a projection

The concept of projection is important for developing an understanding of domain decomposition methods. Before defining projection, we introduce the objects we are dealing with and the way the angles and distances are measured. Assume we are given a vector space V with an inner product $a(\cdot,\cdot)$. For instance, V may be the usual Euclidean space $\mathbb{R}^N$ and $a(u,v) = u^{\mathrm{T}}\mathrm{A}v$, where A is a symmetric, positive definite matrix; that is, all of its eigenvalues are real positives. Associated with the inner product $a(\cdot,\cdot)$ is the a-norm defined by $\|u\|_a = \sqrt{a(u,u)}$. The norm measures in some sense, the length of the vector u. The norm $\|u-v\|_a$ measures the distance between the two vectors u and v. The standard Euclidean inner product is simply $u^{\mathrm{T}}v$, that is A = I, the identity operator. Also, the standard Euclidean norm is given by $\|u\|_2 = \sqrt{u^{\mathrm{T}}u} = \sqrt{\sum_{i=1}^{n} u_i^2}$. Let V_1 be a subspace of V and e be an element of V. A very natural question is to find the element in V_1 that is "closest " to e ; that is which element $e_1 \in \mathrm{V}_1$ minimizes the distances between e and e_1. We formally define the projection of e onto the subspace V_1, in the inner product $a(\cdot,\cdot)$ by

$$e_1 = \mathrm{P}e = \arg \inf_{v \in \mathrm{V}_1} \|e - v\|_a \tag{3.30}$$

There is a very convenient alternative definition of $e_1 = \mathrm{P}e$ given by the following :

$$\text{find } e_1 \in \mathrm{V}_1 \text{ so that } a(e_1, v) = a(e, v), \ \forall v \in \mathrm{V}_1 \tag{3.31}$$

or equivalently, $a(e_1 - e, v) = 0, \quad \forall v \in \mathrm{V}_1$. This is another way of saying that $e_1 - e$ is orthogonal, in the $a(\cdot,\cdot)$ inner product, to all elements of V_1.

Theorem 3.6. *The solution of (3.31) is the minimizer of (3.30).*

Proof.

$$\begin{aligned} \|e_1 - e\|_a^2 &= a(e_1 - e, e_1 - e) \\ &= a(e_1 - e, v - e) \quad \forall v \in \mathrm{V}_1 \\ &\leq \|e_1 - e\|_a \cdot \|v - e\|_a \quad \forall v \in \mathrm{V}_1 \ . \end{aligned}$$

Therefore, by dividing through $\|e_1 - e\|_a$, we obtain

$$\|e_1 - e\|_a \leq \|v - e\|_a \quad \forall v \in \mathrm{V}_1$$

. □

In the special case that $e \in \mathrm{V}_1$ the projection of e is exactly e, that is , $e_1 = pe = e$.

Theorem 3.7. *When* $\mathrm{V} = \mathbb{R}^N$ *and* V_1 *is the span of columns of* R^{T}*, then the projection may be written as the matrix*

$$\mathrm{P} = \mathrm{R}^{\mathrm{T}}(\mathrm{R}\mathrm{A}\mathrm{R}^{\mathrm{T}})^{-1}\mathrm{R}\mathrm{A}$$

.

Proof. Since any element in V_1 is a linear combination of the columns of R^{T}, $e_1 = \mathrm{R}^{\mathrm{T}}\widetilde{e}_1$ for some $\widetilde{e}_1$, similarly, $v = \mathrm{R}^{\mathrm{T}}\widetilde{v}$. Inserting this into (3.31) gives

$$\begin{aligned} a(e_1, v) &= e_1^{\mathrm{T}}\mathrm{A}v \\ &= (\mathrm{R}^{\mathrm{T}}\widetilde{e}_1)^{\mathrm{T}}\mathrm{A}\mathrm{R}^{\mathrm{T}}\widetilde{v} \\ &= \widetilde{e}_1^{\mathrm{T}}\mathrm{R}\mathrm{A}\mathrm{R}^{\mathrm{T}}\widetilde{v} \end{aligned}$$

and

$$a(e, v) = e^{\mathrm{T}}\mathrm{A}\mathrm{R}^{\mathrm{T}}\widetilde{v} \quad \forall\, \widetilde{v} \in span(\mathrm{R}^{\mathrm{T}})$$

thus from (3.31)

$$\widetilde{e}_1^{\mathrm{T}}\mathrm{R}\mathrm{A}\mathrm{R}^{\mathrm{T}}\widetilde{v} = e^{\mathrm{T}}\mathrm{A}\mathrm{R}^{\mathrm{T}}\widetilde{v}$$

and applying the transpose operator and the symmetric property of A we obtain

$$(\mathrm{R}\mathrm{A}\mathrm{R}^{\mathrm{T}})\widetilde{e}_1 = \mathrm{R}\mathrm{A}e$$

or $\widetilde{e}_1 = (\mathrm{R}\mathrm{A}\mathrm{R}^{\mathrm{T}})^{-1}\mathrm{R}\mathrm{A}e$ which implies $\mathrm{P}e = e_1 = \mathrm{R}^{\mathrm{T}}\widetilde{e}_1 = \mathrm{R}^{\mathrm{T}}(\mathrm{R}\mathrm{A}\mathrm{R}^{\mathrm{T}})^{-1}\mathrm{R}\mathrm{A}e$. □

We recall that our aim is to express the projection formulation of the Schwarz methods.

Theorem 3.8. *The classical Schwarz method (3.26) can be written*

$$u^{k+1} = u^k + \mathrm{Q}_m\mathrm{G}(f - \mathrm{L}u^k) \ ,$$

where Q_m, G *and* L *are suitable operators.*

Proof. For each $v \in \mathrm{V}_1^0$, we have

$$\begin{aligned} a(u^{k+\frac{1}{2}} - u^k, v) = a(\widetilde{w_1^k}, v) &= a(w_1^k, v|_{\Omega_1}) \\ &= (f, v|_{\Omega_1})_{\Omega_1} - a_1(u^k, v|_{\Omega_1}) \\ &= (f, v) - a(u^k, v) \\ &= a(u - u^k, v) \end{aligned}$$

and for the second equation

$$a(u^{k+1} - u^{k+\frac{1}{2}}, v) = a(u - u^{k+\frac{1}{2}}, v), \quad \forall v \in \mathrm{V}_2^0$$

Therefore, the sequences $u^{k+\frac{1}{2}}$ and u^{k+1} satisfy :

$$\begin{aligned} u^{k+\frac{1}{2}} - u^k &= \mathrm{P}_1^0(u - u^k) \\ u^{k+1} - u^{k+\frac{1}{2}} &= \mathrm{P}_2^0(u - u^{k+\frac{1}{2}}) \end{aligned} \tag{3.32}$$

where P_i^0, $i = 1, 2$ is the orthogonal projection of V onto V_i^0 with respect to the scalar product induced by the bilinear form $a(\cdot, \cdot)$; that is, for any $w \in \mathrm{V}$, it holds that

$$\mathrm{P}_i^0 w \in \mathrm{V}_i^0 : a(\mathrm{P}_i^0 w - w, v) = 0, \quad \forall\, v \in \mathrm{V}_i^0$$

At each half-step, the corrections calculated are $u^{k+\frac{1}{2}} - u^k$ and $u^{k+1} - u^{k+\frac{1}{2}}$, and thus the projections of the error onto the subspace $\mathrm{H}_0^1(\Omega_1)$ and $\mathrm{H}_0^1(\Omega_2)$. Let denote by I the identical operator, by S_i, $i = 1, 2$ the (non-dense) immersion of V_i^0 into V (that is, $\mathrm{S}_i v = v$ for each

$v \in \mathrm{V}_i^0$) and by $\mathrm{S}_i^{\mathrm{T}}$ its transpose operator; that is, the (non injective) map from V'(the dual space of V) into $(\mathrm{V}_i^0)'$ (the dual space of $\mathrm{V}_{\mathrm{i}}^0$) by

$$< \mathrm{S}_i^T F, v >=< F, \mathrm{S}_i v >, \quad \forall F \in \mathrm{V}',\ v \in \mathrm{V}_i^0 \tag{3.33}$$

If we set $\mathrm{P}_i := \mathrm{S}_i \mathrm{P}_i^0 : \mathrm{V} \to \mathrm{V}$, then from (3.32) it follows that

$$\begin{aligned} u^{k+\frac{1}{2}} &= (\mathrm{I} - \mathrm{P}_1)u^k + \mathrm{P}_1 \mathrm{G} f \\ &= u^k + \mathrm{P}_1 \mathrm{G}(f - \mathrm{L}u^k) \end{aligned} \tag{3.34}$$

and

$$\begin{aligned} u^{k+1} &= (\mathrm{I} - \mathrm{P}_2)u^{k+\frac{1}{2}} + \mathrm{P}_2 \mathrm{G} f \\ &= u^{k+\frac{1}{2}} + \mathrm{P}_2 \mathrm{G}(f - \mathrm{L}u^{k+\frac{1}{2}}) \end{aligned} \tag{3.35}$$

where G is the resolvent operator associated with the Poisson problem (3.18); that is $\mathrm{G} = (\mathrm{L})^{-1} = (-\Delta)^{-1}$, L the Laplace operator and in particular $\mathrm{G}f = u$. Therefore, the alternating multiplicative Schwarz method (3.26) becomes

$$\begin{aligned} u^{k+1} &= (\mathrm{I} - \mathrm{P}_2)[(\mathrm{I} - \mathrm{P}_1)u^k + \mathrm{P}_1 \mathrm{G} f] + \mathrm{P}_2 \mathrm{G} f \\ &= (\mathrm{I} - \mathrm{P}_1)(\mathrm{I} - \mathrm{P}_2)u^k + (\mathrm{I} - \mathrm{P}_2)\mathrm{P}_1 \mathrm{G} f + \mathrm{P}_2 \mathrm{G} f \\ &= u^k + \mathrm{Q}_m(\mathrm{G} f - u^k) = u^k + \mathrm{Q}_m \mathrm{G}(f - \mathrm{L}u^k), \end{aligned} \tag{3.36}$$

where

$$\mathrm{Q}_m := \mathrm{P}_1 + \mathrm{P}_2 - \mathrm{P}_1 \mathrm{P}_2. \tag{3.37}$$

□

Remark 3.9. *The definition of multiplicative Schwarz method for (3.26) is due to the presence of the term* $\mathrm{P}_1\mathrm{P}_2$ *in (3.37).*

Theorem 3.10. *The alternating additive Schwarz method (3.27) can be written as*

$$u^{k+1} = u^k + \mathrm{Q}_a \mathrm{G}(f - \mathrm{L}u^k) \ ,$$

where Q_a, G *and* L *are suitable operators.*

Proof. As previously seen $\widetilde{w_1^k} = \mathrm{P}_1^0(u - u^k)$, $\widetilde{w_2^k} = \mathrm{P}_2^0(u - u^k)$. Therefore

$$\begin{aligned} u^{k+1} &= (\mathrm{I} - \mathrm{P}_1 - \mathrm{P}_2)u^k + (\mathrm{P}_1 + \mathrm{P}_2)\mathrm{G} f \\ &= u^k + \mathrm{Q}_a(\mathrm{G} f - u^k) = u^k + \mathrm{Q}_a \mathrm{G}(f - \mathrm{L}u^k) \end{aligned} \tag{3.38}$$

where

$$\mathrm{Q}_a := \mathrm{P}_1 + \mathrm{P}_2 \ . \tag{3.39}$$

□

Now concerning the error equations for the Schwarz method (3.26), it results from (3.34) - (3.35) that

$$\begin{aligned} u - u^{k+\frac{1}{2}} &= (\mathrm{I} - \mathrm{P}_1)(u - u^k) \\ u - u^{k+1} &= (\mathrm{I} - \mathrm{P}_2)(u - u^{k+\frac{1}{2}}) \end{aligned} \tag{3.40}$$

Introducing the error $e^k := u - u^k$, the previous relations yield the recursion formula :

$$e^{k+1} = (\mathrm{I} - \mathrm{P}_2)(\mathrm{I} - \mathrm{P}_1)e^k \quad \forall\ k \geq 0 \tag{3.41}$$

This relation is the basis of the convergence proof of u^k to u in $\mathrm{H}^1(\Omega)$ that will be seen in the next section. In the same way, using (3.38) for the Schwarz method (3.27), it holds that

$$e^{k+1} = (\mathrm{I} - \mathrm{P}_1 - \mathrm{P}_2)e^k, \quad \forall k \geq 0 \tag{3.42}$$

3.6.2 One level Schwarz

We assume the domain Ω is discretized by conforming finite elements methods (see [27]). Let V_h denote a finite dimensional subspace of $\mathrm{H}_0^1(\Omega)$. A Galerkin finite element approximation to (3.25) is defined as follows:

$$\text{find} \quad u_h \in \mathrm{V}_h \ : \ a(u_h, v_h) = (f, v_h), \quad \forall v_h \in \mathrm{V}_h. \tag{3.43}$$

The unknowns of the finite dimensional problem (3.43) are given by the point values of u_h at the finite element nodes a_j. In fact, denoting by N_h, the total number of nodes and by φ_j the basis functions of V_h, there is a unique function in V_h satisfying $\varphi_j(a_i) = \delta_{ij}$ for each $i, j = 1, \cdots, N_h$ and the function $u_h \in \mathrm{V}_h$ can be represented through

$$u_h(\underline{x}) = \sum_{j=1}^{N_h} u_h(a_j)\varphi_j(\underline{x}) \ . \tag{3.44}$$

Introducing the notation

$$\underline{u} := \{u_h(a_j)\}_{j=1,\cdots,N_h} \tag{3.45}$$

and

$$\underline{f} := \{(f, \varphi_j)\}_{j=1,\cdots,N_h} \tag{3.46}$$

problem (3.43) can be written as

$$\mathrm{A}\underline{u} = \underline{f} \ . \tag{3.47}$$

The matrix A is called the finite element stiffness matrix and is given by

$$\mathrm{A}_{lj} := a(\varphi_j, \varphi_l), \quad l, j := 1, \cdots, N_h \tag{3.48}$$

The stiffness matrix A is positive definite; that is, for any $v_h \in \mathbb{R}^{N_h}$, $v_h \neq 0$, $(\mathrm{A}v_h, v_h) > 0$, where $(\cdot, \cdot)$ denotes the Euclidean scalar product. Indeed, let $v_h \in \mathrm{V}_h$ be the function defined as $v_h(\underline{x}) = \sum_{j=1}^{N_h} v_j\varphi_j(\underline{x})$ then

$$\begin{aligned}(\mathrm{A}v_h, v_h) &= \sum_{l,j=1}^{N_h} v_l a(\varphi_j, \varphi_l)v_j \\ &= a(v_h, v_h) \geq 0\end{aligned}$$

and $(\mathrm{A}v_h, v_h) = 0$ if and only if $v_h = 0$. In particular, it follows that any eigenvalue of A has a positive real part.

Remark 3.11. *Since the bilinear form $a(\cdot, \cdot)$ is symmetric, it follows that* A *is also symmetric.*

Remark 3.12. *Another important remark concerns the condition number*

$$\mathcal{K}_2(\mathrm{A}) := \|\mathrm{A}\|_2\|\mathrm{A}^{-1}\|_2 = \frac{\sqrt{\lambda_{\max(\mathrm{A}^{\mathrm{T}}\mathrm{A})}}}{\sqrt{\lambda_{\min(\mathrm{A}^{\mathrm{T}}\mathrm{A})}}} \ . \tag{3.49}$$

In the symmetric case, we have the simplified relation

$$\mathcal{K}_2(\mathrm{A}) = \frac{\lambda_{\max(\mathrm{A})}}{\lambda_{\min(\mathrm{A})}}$$

and it can be proved that $\mathcal{K}_2(\mathrm{A}) = 0(h^{-2})$, being h the characteristic parameter of the partition of Ω.

Following (3.26), the alternating (multiplicative) Schwarz method at the discrete level is: given $u_h^0 \in \mathrm{V}_h$, solve for each $k \geq 0$

$$\begin{aligned} w_{1,h}^k \in \mathrm{V}_{1,h}^0 \ &: a_1(w_{1,h}^k, v_{1,h}) = (f, v_{1,h})_{\Omega_1} - a_1(u_h^k, v_{1,h}), \quad \forall v_{1,h} \in \mathrm{V}_{1,h}^0 \\ u_h^{k+\frac{1}{2}} &= u_h^k + \widetilde{w_{1,h}^k} \\ w_{2,h}^k \in \mathrm{V}_{2,h}^0 \ &: a_2(w_{2,h}^k, v_{2,h}) = (f, v_{2,h})_{\Omega_2} - a_2(u_h^{k+\frac{1}{2}}, v_{2,h}) \quad \forall v_{2,h} \in \mathrm{V}_{2,h}^0 \end{aligned} \tag{3.50}$$

where $\widetilde{w_{i,h}^k}$ is the finite element function that extends $w_{i,h}^k$ by 0 in $\Omega \setminus \Omega_i$, and a_i denotes the restriction of a to Ω_i. $\mathrm{V}_{i,h}^0$ is the finite dimensional subspace of $\mathrm{H}_0^1(\Omega_i)$. In the same way can be define the additive Schwarz method (3.27) at the finite element level.
The algebraic form of the alternating (multiplicative) Schwarz method follows immediately from (3.34) and (3.35) upon replacing the operators with the corresponding matrices.

$$\begin{aligned} u^{k+\frac{1}{2}} &= u^k + \mathrm{R}_1^{\mathrm{T}} \mathrm{A}_1^{-1} \mathrm{R}_1 (f - \mathrm{A}u^k) \\ u^{k+1} &= u^{k+\frac{1}{2}} + \mathrm{R}_2^{\mathrm{T}} \mathrm{A}_2^{-1} \mathrm{R}_2 (f - \mathrm{A}u^{k+\frac{1}{2}}) \quad . \end{aligned} \tag{3.51}$$

Replacing f with $\mathrm{A}u$, these equations can be written as

$$\begin{aligned} u^{k+\frac{1}{2}} &= u^k + \mathrm{P}_{1,h}(u - u^k) = (\mathrm{I} - \mathrm{P}_{1,h})u^k + \mathrm{P}_{1,h}u \\ u^{k+1} &= u^{k+\frac{1}{2}} + \mathrm{P}_{2,h}(u - u^{k+\frac{1}{2}}) = (\mathrm{I} - \mathrm{P}_{2,h})u^{k+\frac{1}{2}} + \mathrm{P}_{2,h}u \quad , \end{aligned} \tag{3.52}$$

where we have introduced the discrete projection operators $\mathrm{P}_{i,h} := \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i \mathrm{A}$, $i = 1, 2$.

Lemma 3.13. *The matrices* $\mathrm{P}_{i,h}$ *are symmetric and non negative definite with respect to the* A− *scalar product*

$$(w, v)_{\mathrm{A}} := (\mathrm{A}w, v), \quad \forall\ w, v \in \mathbb{R}^{N_h} \tag{3.53}$$

which is induced by the symmetric and positive definite stiffness matrix A. *Moreover,* $\mathrm{P}_{i,h}$ *is the orthogonal projection in the* A−*scalar product onto the subspace spanned by the rows of* $\mathrm{R}_i^{\mathrm{T}}$.

Proof. If we set

$$\mathrm{Q}_i := \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i = \mathrm{P}_{i,h} \mathrm{A}^{-1}, \ i = 1, 2 \tag{3.54}$$

then the correction $c_i^k := \mathrm{Q}_i(f - \mathrm{A}u^k)$ is such that $\mathrm{P}_{i,h}(u - u^k) = c_i^k$; that is c_i^k is the closest vector to the error $u - u^k$ in the subspace spanned by the rows of $\mathrm{R}_i^{\mathrm{T}}$. □

In a compact form, the multiplicative Schwarz method reads

$$\begin{aligned} u^{k+1} &= u^k + (\mathrm{Q}_1 + \mathrm{Q}_2 - \mathrm{Q}_2 \mathrm{A} \mathrm{Q}_1)(f - \mathrm{A}u^k) \\ &= u^k + [\mathrm{I} - (\mathrm{I} - \mathrm{P}_{2,h})(\mathrm{I} - \mathrm{P}_{1,h})] \mathrm{A}^{-1}(f - \mathrm{A}u^k) \quad . \end{aligned} \tag{3.55}$$

Similarly, the additive Schwarz method becomes

$$u^{k+1} = u^k + (\mathrm{Q}_1 + \mathrm{Q}_2)(f - \mathrm{A}u^k) \quad . \tag{3.56}$$

The generalization to M sub-domains, $M > 2$, is straightforward. Setting $\mathrm{P}_{i,h} := \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i \mathrm{A}$, $\mathrm{Q}_i := \mathrm{P}_{i,h} \mathrm{A}^{-1}$, $i = 1, \cdots, M$, the multiplicative Schwarz method becomes

$$\begin{aligned} u^{k+\frac{i}{M}} &= (\mathrm{I} - \mathrm{P}_{i,h}) u^{k+\frac{i-1}{M}} + \mathrm{P}_{i,h} u \\ &= u^{k+\frac{i-1}{M}} + \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i (f - \mathrm{A}u^{k+\frac{i-1}{M}}), i = 1, \cdots, M \end{aligned} \tag{3.57}$$

and the additive Schwarz method reads

$$\begin{aligned} u^{k+1} &= (\mathrm{I} - \sum_{i=1}^{M} \mathrm{P}_{i,h}) u^k + \sum_{i=1}^{M} \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i f \\ &= u^k + (\sum_{i=1}^{M} \mathrm{Q}_i)(f - \mathrm{A}u^k) \quad . \end{aligned} \tag{3.58}$$

The error equation for the multiplicative case takes the form

$$e^{k+1} = u - u^{k+1} = (\mathrm{I} - \mathrm{P}_{M,h}) \cdots (\mathrm{I} - \mathrm{P}_{1,h})(u - u^k) \tag{3.59}$$

and for the additive case:

$$e^{k+1} = u - u^{k+1} = (\mathrm{I} - \sum_{i=1}^{M} \mathrm{P}_{i,h})(u - u^k) \quad . \tag{3.60}$$

For what concerns the additive Schwarz method with M sub-domains let's introduce the matrix

$$\mathrm{P}_{as} = (\sum_{i=1}^{M} \mathrm{Q}_i)^{-1} \tag{3.61}$$

and nothing that

$$\mathrm{P}_{as}^{-1} \mathrm{A} = \sum_{i=1}^{M} \mathrm{P}_{i,h} \tag{3.62}$$

the one level additive Schwarz method can be regarded as a fixed-point problem for the following support

$$\mathrm{P}_{as} := (\sum_{i=1}^{M} \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i)^{-1} = (\sum_{i=1}^{M} \mathrm{Q}_i)^{-1} \tag{3.63}$$

so that the preconditioned matrix of the system (3.47) becomes

$$\mathrm{P}_{as}^{-1} \mathrm{A} = \sum_{i=1}^{M} \mathrm{P}_{i,h} = \mathrm{Q}_a \quad . \tag{3.64}$$

As suggested by (3.58), the additive Schwarz method is simply a Richardson method for (3.47) with preconditioner P_{as} for A. Since P_{as} is symmetric and positive definite, the convergence of the preconditioned system can be more effectively accelerated by the conjugate gradient (CG) method [120], which converges faster than the Richardson method does. Therefore, by using the CG method, the convergence rate is : (see for example [58])

$$\|u^k - u\|_{\mathrm{A}} \leq 2(\frac{\sqrt{\mathcal{K}(\mathrm{P}_{as}^{-1}\mathrm{A})} - 1}{\sqrt{\mathcal{K}(\mathrm{P}_{as}^{-1}\mathrm{A})} + 1}) \|u^0 - u\|_{\mathrm{A}} \tag{3.65}$$

where $\|v\|_{\mathrm{A}} = \sqrt{(v,v)_{\mathrm{A}}}$ is the norm associates with the scalar product $(v,w)_{\mathrm{A}}$ introduced in (3.53). Concerning the condition number $\mathcal{K}(\mathrm{P}_{as}^{-1}\mathrm{A})$, denoting as usual by H the maximum diameter of the sub-domains Ω_i, $i = 1, \cdots, M$ and by βH the linear measure of the overlapping region between two adjacent sub-domains ($0 < \beta \leq 1$), the following estimate holds :

$$\mathcal{K}(\mathrm{P}_{as}^{-1}\mathrm{A}) \leq C \frac{1}{\beta H} \tag{3.66}$$

where C is a constant possibly dependent on the coefficients of the Laplace operator, see [42, 43]. By relation (3.66), the condition number depends on the number of sub-domains for a given

amount of overlap, thus when the number of sub-domains is getting greater, the preconditioned system is not scalable (that is the number of iterations required to obtain the convergence is dependent on the number of sub-domains). The reason is the lack of communications among sub-domains "distant", since the only communication between sub-domains is through overlap region. We will see in the next section how the two level Schwarz solves this drawback.

3.6.3 The colouring technique

Contrary to the additive Schwarz method, the multiplicative method has very little potential for parallel implementation. Since, many sub-domains do not have to share any grid point, a strategy of sub-domain colouring can adopted to allow a simultaneous and independently update of subsets of equations by (3.57)(see Figure 3.9 for an example). The colouring is such

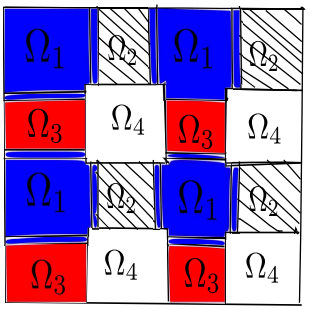

Figure 3.9: sub-domain colouring with four colours.

that no two overlapping sub-domains have the same colour. The number of parallel processors that may be used is given by the number of sub-domains in a given colour. A thorough discussion of this technique can be found in [124].

3.6.4 Two level Schwarz methods

In the previous section, we have seen that the convergence rate of the preconditioned iterative Schwarz methods deteriorates when the number of sub-domains becomes large. This is due to the lack of informations among distant sub-domains. The two level method overcomes this drawback by introducing a coarse global mesh over the whole domain in order to guarantee a mechanism of global communication among all sub-domains. This involves the smoothing of the original (fine) problem and the definition of an auxiliary (coarse) problem on an embedded mesh that is coarser than the original one. In this case, the preconditioner for the original problem is obtained as a superposition of the solution of an auxiliary problem on a coarser grid and a smoother on the original grid. The abstract formulation of the problem is the following. Consider two finite dimensional problems, the original one

$$\text{find} \quad u_h \in \mathrm{V}_h : \quad a_h(u_h, v_h) = (f_h, v_h), \quad \forall\, v_h \in \mathrm{V}_h \tag{3.67}$$

and the auxiliary one

$$\text{find} \quad u_H \in \mathrm{V}_H : \quad a_H(u_H, v_H) = (f_H, v_H), \quad \forall\, v_H \in \mathrm{V}_H \tag{3.68}$$

where $a_h(\cdot,\cdot)$ and $a_H(\cdot,\cdot)$ are bilinear forms on V_h and V_H respectively.
Given a basis φ_j, $j = 1, \cdots, N_h$ of V_h and a basis ψ_l, $l = 1, \cdots, N_H$ of V_H, we construct the matrices

$$(A_h)_{sj} := a_h(\varphi_j, \varphi_s)$$
$$(A_H)_{ml} := a_H(\psi_l, \psi_m)$$

The two spaces V_h and V_H are related by an operator $\mathrm{I}_h : \mathrm{V}_H \to \mathrm{V}_h$. Basically, I_h is the linear interpolation from the coarse grid to the fine grid and its representation is R^T the prolongation

operator (see section 3.5.2); h and H are the maximum diameter of the fine and coarse grid respectively. Two different adjoint operators can be associated with I_h :

$$\begin{cases} \mathrm{I}_h^{\mathrm{T}} : \mathrm{V}_h \to \mathrm{V}_H \\ (\mathrm{I}_h^{\mathrm{T}} w_h, v_H)_{\mathrm{V}_H} = (w_h, \mathrm{I}_h v_H)_{\mathrm{V}_h} \quad \forall\, w_h \in \mathrm{V}_h, v_H \in \mathrm{V}_H \end{cases} \tag{3.69}$$

and

$$\begin{cases} \mathrm{J}_h^{\mathrm{T}} : \mathrm{V}_h \to V_H \\ a_H(\mathrm{J}_h^{\mathrm{T}} w_h, v_H) = a_h(w_h, J_h v_H), \forall w_h \in \mathrm{V}_h, v_H \in \mathrm{V}_H \end{cases} \tag{3.70}$$

In the matrix form, the last definition gives $\mathrm{A}_H \mathrm{J}_h^{\mathrm{T}} = \mathrm{I}_h^{\mathrm{T}} \mathrm{A}_h$, thus $\mathrm{J}_h^{\mathrm{T}} = \mathrm{A}_H^{-1} \mathrm{I}_h^{\mathrm{T}} \mathrm{A}_h$. A preconditioner P_h for A_h can be constructed in the following way :

$$\mathrm{P}_h^{-1} := \mathrm{Q}_H^{-1} + \mathrm{Q}_h^{-1}, \quad \mathrm{Q}_h^{-1} = \mathrm{I}_h \mathrm{A}_H^{-1} \mathrm{I}_h^{\mathrm{T}} \tag{3.71}$$

where Q_H is any convenient symmetric positive definite matrix simpler than A_h itself. The preconditioned matrix becomes $\mathrm{P}_h^{-1} \mathrm{A}_h = \mathrm{I}_h \mathrm{J}_h^{\mathrm{T}} + \mathrm{Q}_H^{-1} \mathrm{A}_h$.
Now how to generate a coarse grid for Schwarz? As usual, the technique for constructing an overlapping decomposition of Ω into M sub-domains $\Omega_1, \cdots, \Omega_M$ consist to assume that a non overlapping partition $w_1, \cdots, w_M$ of Ω is available. One possibility is to choose each subregion w_i as an element from the coarse finite element triangulation of Ω of size H. Next, each w_i is extended to a larger domain Ω_i, consisting of all points in Ω at a distance not larger than βH from w_i with $0 < \beta \leq 1$. The restriction and extension matrices, R_i and $\mathrm{R}_i^{\mathrm{T}}$ as well as the local matrices are defined accordingly. Assuming the fine grid of diameter size h is a refinement of the coarse grid of size H, we denote by $\mathrm{R}_H^{\mathrm{T}}$, the interpolation map of coarse grid functions to the fine grid functions. By using piecewise-linear elements, $\mathrm{R}_H^{\mathrm{T}}$ interpolates the nodal values from the coarse grid (of w_i) to all the vertices of the fine grid. Its transpose R_H is a weighted restriction map. The corresponding stiffness matrix A_H on the coarse mesh is given by $\mathrm{A}_H = \mathrm{R}_H \mathrm{A}_h \mathrm{R}_H^{\mathrm{T}}$.
In conclusion, we have, see (3.63),

$$\mathrm{Q}_h^{-1} = \sum_{i=1}^{M} \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i, \qquad \mathrm{Q}_H^{-1} = \mathrm{R}_H^{\mathrm{T}} \mathrm{A}_H^{-1} \mathrm{R}_H$$

where the index i refers to the ith sub-domain, $i = 1, \cdots, M$ and the index H refers to the coarse grid (where the subdomains play the role of elements). Thus, the resulting preconditioner is :

$$\mathrm{P}_{cas}^{-1} := (\sum_{i=0}^{M} \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i)^{-1} \tag{3.72}$$

where we have set for notational convenience $\mathrm{R}_0 := \mathrm{R}_H$ and $\mathrm{A}_0 := \mathrm{A}_H$.

3.6.5 Convergence estimates

We have seen that the alternating Schwarz method presented in the last section furnishes a pair of sequences u^k and $u^{k+\frac{1}{2}}$ that converge to the solution u of the Dirichlet boundary problem. The original proof from Lions [83] is based on a weak formulation and not on the maximum principle, like the classical one. We recall, see (3.40), that the sequences u^k and $u^{k+\frac{1}{2}}$ generated by the multiplicative Schwarz method satisfies

$$u - u^{k+\frac{1}{2}} = (\mathrm{I} - \mathrm{P}_1)(u - u^k)$$
$$u - u^{k+1} = (\mathrm{I} - \mathrm{P}_2)(u - u^{k+\frac{1}{2}})$$

where, for $i = 1, 2$, $\mathrm{P}_i = \mathrm{S}_i\mathrm{P}_i^0$, S_i is the immersion of $\mathrm{V}_i^0 := \{v \in \mathrm{V} \mid v = 0 \text{ in } \Omega \setminus \overline{\Omega}_i \}$ into $\mathrm{V} := \mathrm{H}_0^1(\Omega)$ and P_i^0 is the orthogonal projection of V onto V_i^0 with respect to the bilinear forms (3.26). Let us denote by $\|\cdot\|$ the norm associated with the form $a(\cdot,\cdot)$ which is equivalent to the usual norm of $\mathrm{H}^1(\Omega)$. Moreover, for any pair of vectors spaces W_1 and W_2, denote by $\mathrm{W}_1 \oplus \mathrm{W}_2$ their direct sum; namely the vector space of all elements w that can be written in a unique way as the sum of an element in W_1 and W_2.

Theorem 3.14. *Assume that Ω is a bounded domain in $\mathbb{R}^2$, with a Lipschitz boundary $\partial\Omega$. If $\mathrm{V}_1^0 \oplus \mathrm{V}_2^0 = \mathrm{V}$, then both u^k and $u^{k+\frac{1}{2}}$ converge to u in $\mathrm{H}_0^1(\Omega)$ as $k \to \infty$.*

It is worthwhile to note that the assumption of theorem 3.14 is always satisfies, provided $\Omega = \Omega_1 \cup \Omega_2$, (see [83]). Under the slightly more restrictive assumption $\mathrm{V}_1^0 \oplus \mathrm{V}_2^0 = \mathrm{V}$, which is satisfied for a large class of subdomains Ω_1 and Ω_2, then the following theorem is an estimate of the rate of convergence. We need

Lemma 3.15. *If $\mathrm{V}_1^0 \oplus \mathrm{V}_2^0 = \mathrm{V}$, then there exists a constant $C_0 \geq 1$ such that*

$$\|v\| \leq C_0(\|\mathrm{P}_1 v\|^2 + \|\mathrm{P}_2 v\|^2)^{\frac{1}{2}} \quad \forall v \in \mathrm{V} \quad . \tag{3.73}$$

Then the convergence result is the following:

Theorem 3.16. *If $\mathrm{V}_1^0 \oplus \mathrm{V}_2^0 = \mathrm{V}$, then the iteration operators $(\mathrm{I}-\mathrm{P}_1)(\mathrm{I}-\mathrm{P}_2)$ and $(\mathrm{I}-\mathrm{P}_2)(\mathrm{I}-\mathrm{P}_1)$ are contractions in $\mathrm{H}_0^1(\Omega)$ with respect to the norm induced by the bilinear form $a(\cdot,\cdot)$ of (3.26).*

As a consequence of this theorem, the convergence of the alternating Schwarz process takes place with a geometric rate, more precisely, setting $K_0 := (1 - C_0^{-2})^{\frac{1}{2}}$, we have

$$\begin{aligned} \|e^{k+1}\| &= \|(\mathrm{I} - \mathrm{P}_2)(\mathrm{I} - \mathrm{P}_1)e^k\| \\ &\leq K_0\|e^k\| \leq K_0^{k+1}\|e^0\| \end{aligned}$$

and analogously,

$$\begin{aligned} \|e^{k+\frac{1}{2}}\| &= \|(\mathrm{I} - \mathrm{P}_1)(\mathrm{I} - \mathrm{P}_2)e^{k+\frac{1}{2}}\| \\ &\leq K_0^{k+1}\|e^{\frac{1}{2}}\| = K_0^{k+1}\|(\mathrm{I} - \mathrm{P}_1)e^0\| \\ &\leq K_0^{k+1}\|e^0\| \end{aligned}$$

Convergence of the two level method

We assume that the fine grid $\mathcal{I}_h$ and the coarse grid $\mathcal{I}_H$ underlying problems (3.67) and (3.68) respectively are comparable, i.e., for any $K_h \in \mathcal{I}_h$ and $K_H \in \mathcal{I}_H$ with $K_h \cap K_H \neq \emptyset$, there exist two positive constant C_0 and C_1 such that

$$C_0 diam\, K_H \leq diam\, K_h \leq C_1 diam\, K_H \quad . \tag{3.74}$$

- If then the two grids are quasi uniform and of comparable size (i.e $C_0\,, C_1 \simeq 1$), by using either block-Jacobi or a symmetric block Gauss-Seidel iteration as a smoother, the preconditioner P_{cas} is uniformly optimal for the operator A_h (the spectrum of $\mathrm{P}_{cas}^{-1}\mathrm{A}_h$ is uniformly bounded with respect to h)

- If instead $\mathcal{I}_H$ is genuinely coarser than $\mathcal{I}_h$, which means that there exist triangles K_h and K_H, $K_h \cap K_H \neq \emptyset$ for which (3.74) holds only for extremely small C_0, then the above smoothers are no longer sufficient to guarantee that P_{cas} is uniformly optimal for A_h

- However, if the smoothing is based on the overlapping Schwarz method Q_h given by the symmetric multiplicative preconditioner, see (3.55), or Q_h is given by the additive preconditioner, see (3.63), then P_{cas}^{-1} is uniformly optimal provided the coarse grid $\mathcal{I}_H$ is comparable with the sub-domain partition. This means that there exist positive constants $\widehat{C_0}$ and $\widehat{C_1}$ such that for each $K_H \in \mathcal{I}_H$ such that $K_H \cap \Omega_i \neq \emptyset$, it holds

$$\widehat{C_0} diam\, K_H \leq diam\, \Omega_i \leq \widehat{C_1} diam\, K_H \quad .$$

The proof of these two results can be found in [20] and the reference therein. In particular, it is underlined that the spectrum of $P_{cas}^{-1} A_h$ is bounded independently of h and H if the linear measure β of the overlapping region is kept proportional to H, say, given by βH. In addition, the convergence of the iterative procedures is poor for very small values of β, but improves rapidly when the overlap increases. In [114], it is reported that the relation (3.66) is satisfied with P_{cas} equal to (3.72) and that the preconditioner P_{cas} is scalable.

3.6.6 Why use a Schwarz preconditioner?

It is well known that in the elliptic problems the preconditioners based on non overlapping techniques do not always converge. As the final aim of this work is the formulation of a new method for the numerical solution of a PDE system (i.e. 2D Navier-Stokes equations), method in which the computational kernels reduced to elliptic problems, it seems very convenient to use overlapping additive Schwarz preconditioner for solving the algebraic systems. In the next section of this chapter, we deal with some applications of multi-domain methods to problems where the iterations are based on transmission conditions at interfaces and finally, we provide some numerical experiments.

3.7 Application of multi-domain methods to other problems

3.7.1 The generalized Stokes problem

We consider the generalized Stokes problem [112] :

$$\begin{cases} \alpha\sigma + div\, \underline{u} = g & \text{in } \Omega \\ \alpha\underline{u} - a\mu\Delta\underline{u} + \beta\nabla\sigma = f & \text{in } \Omega, \end{cases} \tag{3.75}$$

where α, β, a and μ are positive constants, f and g are given functions and Ω is an open domain of $\mathbb{R}^m$ ($m = 2$ or 3). The meaning of the various constant and functions appearing in (3.75) depends on what physical problem the model system (3.75) stands for. Here, it represents an intermediate step of the linearization process for the full Navier-Stokes equations for compressible flows. The vector variable $\underline{u}$ denotes the velocity field, the scalar function σ is the logarithm of the density, f, g, a and β are known at previous time-level, α is the inverse of the time-step, and μ is an average of the dynamic viscosity. Several sets of boundary condition render (3.75) a mathematically well posed boundary value problem. Among these, we consider those supplementing (3.75) in the simulation of external flows around an airfoil. In this case, the computational domain Ω is the complement of the airfoil profile, truncated in the far field by the boundary Γ_∞. Γ_B is the boundary of the solid body and by $\Gamma_\infty^- \equiv \{x \in \mathbb{R}^m |\ x \in \partial\Omega \setminus \Gamma_B,\ \underline{u}_\infty \cdot n < 0\}$, $\Gamma_\infty^+ \equiv \partial\Omega \setminus (\Gamma_B \cup \Gamma_\infty^-)$ is the inflow and the out flow external boundaries respectively. Clearly $\underline{u}_\infty$ and n denote the free stream velocity and the unit outward normal vector to $\partial\Omega$ respectively.
The boundary conditions we are considering take the following form

$$\begin{aligned} \underline{u} &= 0 \quad \text{on } \Gamma_B \\ \underline{u} &= \underline{u}_\infty \quad \text{on } \Gamma_\infty^- \\ S(\underline{u}, \sigma) \equiv a\mu \frac{\partial \underline{u}}{\partial n} - \beta\sigma n &= 0 \quad \text{on } \Gamma_\infty^+, \end{aligned} \tag{3.76}$$

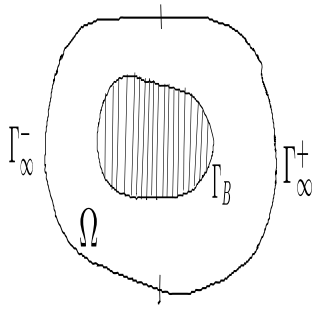

Figure 3.10: The computational domain for the flow around and airfoil.

where $(\frac{\partial u}{\partial n}) \equiv \sum_j (\partial_j u_i) n_j, j = 1, \cdots, m$. We emphasize that a Dirichlet condition can be imposed on σ on Γ_∞^-, while m conditions are enforced on Γ_∞^+. The last condition in (3.76) expresses the vanishing of both normal and shear stresses given respectively by $a\mu \frac{\partial u}{\partial n} \cdot n - \beta\sigma$ and $a\mu \frac{\partial u}{\partial n} \cdot \tau^s$, $(s = 1, \cdots, m-1)$.
In order to show that problem (3.75) - (3.76) is well posed, it requires its variational formulation. We define $V_0 \equiv \{v \in \mathrm{H}^1(\Omega) \mid v = 0 \text{ on } \Gamma_\infty^- \cup \Gamma_B\}$, $\Sigma_0 \equiv \mathrm{L}^2(\Omega)$. We recall that $\mathrm{H}^k(\Omega)$ denotes the Sobolev space of measurable functions whose distributional derivatives of order less or equal than k belong to the space $\mathrm{L}^2(\Omega)$ (see, e.g., Lions-Magenes [87]). The norm in this space will be denoted by $\|\cdot\|_k$. The weak formulation of the problem (3.75) with boundary condition (3.76) is the following: we choose for simplicity $\underline{u}_\infty = 0$, then find $\underline{u} \in V_0,\ \sigma \in \Sigma_0$ such that

$$\begin{cases} \displaystyle\int_\Omega (\alpha\sigma + div\, \underline{u})\varphi \, d\Omega = \int_\Omega g\varphi \, d\Omega \quad \forall\ \varphi \in \Sigma_0 \\ \displaystyle\int_\Omega (\alpha \underline{u}\, \underline{v} + a\mu \nabla \underline{u} \cdot \nabla \underline{v} - \beta\sigma div\, \underline{v})\, d\Omega = \int_\Omega f \underline{v}\, d\Omega \quad \forall\ \underline{v} \in V_0 \quad . \end{cases} \tag{3.77}$$

If it is not otherwise specified, the integrals are extended to the whole domain Ω. Consider the bilinear form on $\mathrm{V}_0 \times \Sigma_0$

$$A[(\underline{u}, \sigma), (\underline{v}, \varphi)] \equiv \int_\Omega (\alpha\beta\sigma\varphi + \beta div\, \underline{u}\varphi + \alpha \underline{u}\, \underline{v} + a\mu \nabla \underline{u} \cdot \nabla \underline{v} - \beta\sigma div\, \underline{v})\, d\Omega \tag{3.78}$$

which is associated to the weak problem (3.77). It is easily seen that A is continuous and coercive in $\mathrm{V}_0 \times \Sigma_0$, since

$$A[(\underline{u}, \sigma), (\underline{u}, \sigma)] = \alpha\beta \|\sigma\|_0^2 + \alpha \|\underline{u}\|_0^2 + a\mu \|\nabla \underline{u}\|_0^2 \quad . \tag{3.79}$$

Therefore, by Lax-Milgram lemma, for each $f \in \mathrm{L}^2(\Omega)$, $g \in \mathrm{L}^2(\Omega)$, the weak problem (3.77) admits a unique solution and we have the following estimate

$$\alpha \|\underline{u}\|_0^2 + 2a\mu \|\nabla \underline{u}\|_0^2 + \alpha\beta \|\sigma\|_0^2 \le \alpha^{-1} (\|f\|_0^2 + \beta \|g\|_0^2) \quad . \tag{3.80}$$

Our aim is to give a correct domain decomposition procedure for the solution of problem (3.75) and (3.76). We assume for simplicity that $\underline{u}_\infty = 0$, and the computational domain is subdivided into two subdomains as indicated in Figure 3.11 . We define

$$\begin{aligned} V_1 &\equiv \{\, \underline{v} \in \mathrm{H}^1(\Omega_1) \mid \underline{v} = 0 \quad \text{on} \quad \Gamma_\infty^- \} \\ V_{0,1} &\equiv \{\, \underline{v} \in V_1 \mid \underline{v} = 0 \quad \text{on} \quad \Gamma \} \\ V_2 &\equiv \{\, \underline{v} \in \mathrm{H}_0^1(\Omega_2) \mid \underline{v} = 0 \quad \text{on} \quad \Gamma_B \} \end{aligned}$$

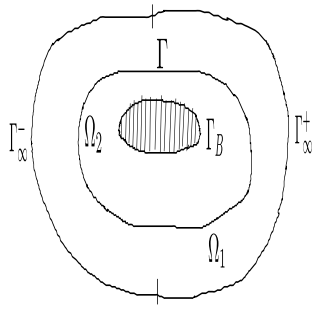

Figure 3.11: The sub-domain decomposition of the computational domain of Figure 3.10 .

and consider the multi-domain problem:
find $\underline{u}^1 \in V_1$, $\sigma^1 \in \mathrm{L}^2(\Omega_1)$, $\underline{u}^2 \in V_2$, $\sigma^2 \in \mathrm{L}^2(\Omega_2)$ such that

$$\begin{aligned}
\int_{\Omega_1} (\alpha\sigma^1 + div\,\underline{u}^1)\varphi\, d\Omega &= \int_{\Omega_1} g\varphi\, d\Omega \quad \forall\ \varphi \in \mathrm{L}^2(\Omega_1)\\
\int_{\Omega_1} (\alpha\underline{u}^1\underline{v} + a\mu\nabla\underline{u}^1\cdot\nabla\underline{v} - \beta\sigma^1 div\,\underline{v})\, d\Omega &= \int_{\Omega_1} f\underline{v}\, d\Omega, \quad \forall\ \underline{v} \in V_{0,1}\\
\underline{u}^1 &= \underline{u}^2 \ \text{ on } \Gamma\\
\int_{\Omega_2} (\alpha\sigma^2 + div\,\underline{u}^2)\psi\, d\Omega &= \int_{\Omega_2} g\psi\, d\Omega \quad \forall\ \psi \in \mathrm{L}^2(\Omega_2)\\
\int_{\Omega_2} (\alpha\underline{u}^2\underline{w} + a\mu\nabla\underline{u}^2\cdot\nabla\underline{w} - \beta\sigma^2 div\,\underline{w})\, d\Omega &= \int_{\Omega_2} f\underline{w}\, d\Omega, \quad \forall\ \underline{w} \in \mathrm{H}_0^1(\Omega_2)\\
\int_{\Omega_2} (\alpha\underline{u}^2 R_2\chi + a\mu\nabla\underline{u}^2\cdot\nabla R_2\chi - \beta\sigma^2 div\, R_2\chi)\, d\Omega &= \int_{\Omega_2} f R_2\chi\, d\Omega + \int_{\Omega_1} f R_1\chi\, d\Omega -\\
\int_{\Omega_1} (\alpha\underline{u}^1 R_1\chi + a\mu\nabla\underline{u}^1\cdot\nabla R_1\chi - \beta\sigma^1 div\, R_1\chi)\, d\Omega &\quad \forall\ \chi \in \mathrm{H}^{\frac{1}{2}}(\Gamma),
\end{aligned} \tag{3.81}$$

where R_1 and R_2 are any possible pair of (continuous) operators from $\mathrm{H}^{\frac{1}{2}}(\Gamma)$ to V_1 and V_2 which satisfy $R_k\chi = \chi$ on Γ, $k = 1, 2$. From $(3.81)_6$, it follows that $S(\underline{u}^1, \sigma^1) = S(\underline{u}^2, \sigma^2)$ on Γ.

Theorem 3.17. *The weak problem (3.77) and its multi-domain formulation (3.81) are equivalent.*

Proof. Let $(\underline{u}, \sigma)$ be a solution to the weak problem (3.77); then we set $\underline{u}^k \equiv u|_{\Omega_k}$, $\sigma^k \equiv \sigma|_{\Omega_k}$, $k = 1, 2$. Then obviously $\underline{u}^k$ and σ^k is a solution to the multi-domain problem. Conversely, let $\underline{u}^k$ and σ^k $(k = 1, 2)$ be a solution to the multi-domain formulation. Since $\underline{u}^1 = \underline{u}^2$ on Γ, we define

$$\underline{u} \equiv \begin{cases} \underline{u}^1 & \text{in } \Omega_1 \\ \underline{u}^2 & \text{in } \Omega_2 \end{cases}, \qquad \sigma \equiv \begin{cases} \sigma^1 & \text{in } \Omega_1 \\ \sigma^2 & \text{in } \Omega_2 \end{cases} \tag{3.82}$$

and we have $\underline{u} \in V_0$, $\sigma \in \Sigma_0$, then the equation $(3.77)_1$ of the weak formulation is satisfied since $\int_\Omega (\alpha\sigma + div\,\underline{u})\varphi\, d\Omega = \sum_k \int_{\Omega_k} (\alpha\sigma^k + div\,\underline{u}^k)\varphi\, d\Omega_k = \int_\Omega g\varphi\, d\Omega$, $\forall\ \varphi \in \mathrm{L}^2(\Omega)$. Now, we take $\underline{v} \in V_0$ and set $\chi \equiv v|_\Gamma \in \mathrm{H}^{\frac{1}{2}}(\Gamma)$. Let us define $R\chi \in V_0$ such that $R\chi = R_k\chi$ in Ω_k, $\underline{v}^k \in V_0$ $(k = 1, 2)$ as $\underline{v}^k|_{\Omega_k} = v|_{\Omega_k} - R_k\chi$ and $\underline{v}^k|_{\Omega/\Omega_k} = 0$, that is $\underline{v}^1 = 0$ in Ω_2, $\underline{v}^2 = 0$ in Ω_1. We note

that $\underline{v}^1|_{\Omega_1} \in V_{0,1}$ and $\underline{v}^2|_{\Omega_2} \in \mathrm{H}_0^1(\Omega_2)$ and that $\underline{v} = \underline{v}^1 + \underline{v}^2 + R\chi$. Thus we get

$$\int_\Omega (\alpha \underline{u}\underline{v} + a\mu \nabla \underline{u} \cdot \nabla \underline{v} - \beta\sigma div\, \underline{v})\, d\Omega = \sum_k \int_{\Omega_k} (\alpha \underline{u}^k \underline{v}^k + a\mu \nabla \underline{u}^k \cdot \nabla \underline{v}^k - \beta\sigma^k div\, \underline{v}^k)\, d\Omega_k +$$

$$\sum_k \int_{\Omega_k} (\alpha \underline{u}^k R_k\chi + a\mu \nabla \underline{u}^k \cdot \nabla R_k\chi - \beta\sigma^k div\, R_k\chi)\, d\Omega_k$$

$$= \sum_k \int_{\Omega_k} f(\underline{v}^k + R_k\chi)\, d\Omega_k$$

$$= \int_\Omega f\underline{v}\, d\Omega$$

where we have used $(3.81)_2$, $(3.81)_5$ and $(3.81)_6$. □

In order to make the above domain decomposition method appealing in view of numerical computation, an iterative procedure can allow the resolution of the two subproblems, one within Ω_1 and the other in Ω_2. It is the following: given an initial guess for $\underline{u}$ at the interface Γ, then go from step $m-1$ to the next step m by solving the two following problems :

- Within Ω_1, solve a problem like (3.75) with the boundary conditions

$$\begin{cases} \underline{u}_m^1 = \underline{u}_\infty \quad \text{on } \Gamma_\infty^- \\ S(\underline{u}_m^1, \sigma_m^1) = 0 \quad \text{on } \Gamma_\infty^+ \\ \underline{u}_m^1 = \theta_m \underline{u}_{m-1}^2 + (1-\theta_m)\underline{u}_{m-1}^1\theta \quad \text{on } \Gamma \quad , \end{cases} \tag{3.83}$$

  θ_m being a positive acceleration parameter that is determined in order to ensure (and possibly, to accelerate) the convergence of the iterative scheme

- Then in Ω_2, solve a problem like (3.75) with the boundary conditions

$$\begin{cases} S(\underline{u}_m^2, \sigma_m^2) = S(\underline{u}_m^1, \sigma_m^1) \quad \text{on } \Gamma \\ \underline{u}_m^2 = 0 \quad \text{on } \Gamma_B \quad . \end{cases} \tag{3.84}$$

We give the weak formulation of the iterative procedure which is very important in the study of the convergence. Let's define the bilinear forms

$$a_1[(\underline{u}^1, \sigma^1), (\underline{v}, \varphi)] \equiv \int_{\Omega_1} (\beta\alpha\sigma^1\varphi + \beta div\, \underline{u}^1\varphi + \alpha \underline{u}^1\underline{v} + a\mu \nabla \underline{u}^1 \cdot \nabla \underline{v} - \beta\sigma^1 div\, \underline{v})\, d\Omega \tag{3.85}$$

$$a_2[(\underline{u}^2, \sigma^2), (\underline{w}, \psi)] \equiv \int_{\Omega_2} (\beta\alpha\sigma^2\psi + \beta div\, \underline{u}^2\psi + \alpha \underline{u}^2\underline{w} + a\mu \nabla \underline{u}^2 \cdot \nabla \underline{w} - \beta\sigma^2 div\, \underline{w})\, d\Omega \ , \tag{3.86}$$

which are continuous and coercive on $\mathrm{H}^1(\Omega_k) \times \mathrm{L}^2(\Omega_k)$, $k = 1, 2$. We remark however that these forms are not symmetric, hence they don't define a scalar product in $\mathrm{H}^1(\Omega_k) \times \mathrm{L}^2(\Omega_k)$. Then define the extension operator $E_k \ : \mathrm{H}^{\frac{1}{2}}(\Gamma) \to \mathrm{V}_k \times \mathrm{L}^2(\Omega_k)$ in the following way :

$$\begin{cases} E_1\chi \in \mathrm{V}_1 \times \mathrm{L}^2(\Omega_1) \\ a_1[E_1\chi, (\underline{v}, \varphi)] = 0 \quad \forall\, \underline{v} \in \mathrm{V}_{0,1},\ \varphi \in \mathrm{L}^2(\Omega_1) \\ (E_1\chi)_{1|\Gamma} = \chi \end{cases} \tag{3.87}$$

$$\begin{cases} E_2\chi \in \mathrm{V}_2 \times \mathrm{L}^2(\Omega_2) \\ a_2[E_2\chi, (\underline{w}, \psi)] = 0 \quad \forall\, \underline{w} \in \mathrm{H}_0^1(\Omega_2),\ \psi \in \mathrm{L}^2(\Omega_2) \\ (E_2\chi)_{1|\Gamma} = \chi \quad . \end{cases} \tag{3.88}$$

Then the iterative scheme introduced in (3.83), (3.84) can be reformulated in a variational form in the following way : solve for $m \geq 1$, $\underline{u}_m^1 \in \mathrm{V}_1$, $\sigma_m^1 \in \mathrm{L}^2(\Omega_1)$,

$$\begin{cases} a_1[(\underline{u}_m^1, \sigma_m^1), (\underline{v}, \varphi)] = (f, \underline{v}) + \beta(g, \varphi) \quad \forall\, \underline{v} \in \mathrm{V}_{0,1},\ \varphi \in \mathrm{L}^2(\Omega_1) \\ \underline{u}_m^1|_\Gamma = \theta_m \underline{u}_{m-1}^2|_\Gamma + (1-\theta_m)\underline{u}_{m-1}^1|_\Gamma \equiv g_{m-1} \quad \text{on } \Gamma \end{cases} \tag{3.89}$$

and $\underline{u}_m^2 \in \mathrm{V}_2$, $\sigma_m^2 \in \mathrm{L}^2(\Omega_2)$,

$$\begin{cases} a_2[(\underline{u}_m^2, \sigma_m^2), (\underline{w}, \psi)] = (f, \underline{w}) + \beta(g, \psi)\, \forall\, \underline{v} \in \mathrm{H}_0^1(\Omega_2),\ \psi \in \mathrm{L}^2(\Omega_2) \\ a_2[(\underline{u}_m^2, \sigma_m^2), E_2\chi] = (f, (E_2\chi)_1) + \beta(g, (E_2\chi)_2) - a_1[(\underline{u}_m^1, \sigma_m^1), E_1\chi] + (f, (E_1\chi)_1) + \\ \beta(g, (E_1\chi)_2) \quad \forall\ \chi \in \mathrm{H}^{\frac{1}{2}}(\Gamma) \quad , \end{cases} \tag{3.90}$$

where $(\cdot\,,\cdot)$ denotes the scalar product in $\mathrm{L}^2(\Omega_1)$ or $\mathrm{L}^2(\Omega_2)$ and the initial guess g_0 can be arbitrarily chosen in $\mathrm{H}^{\frac{1}{2}}(\Gamma)$. Here θ_m is a positive relaxation parameter which will be determined in the sequel in such a way that the scheme converges. We can remark that (3.90) is equivalent to the following problem $\underline{u}_m^2 \in \mathrm{V}_2\,,\ \ \sigma_m^2 \in \mathrm{L}^2(\Omega_2)$,

$$\begin{cases} a_2[(\underline{u}_m^2, \sigma_m^2), (\underline{w}, \psi)] = (f, \underline{w}) + \beta(g, \psi) - \\ a_1[(\underline{u}_m^1, \sigma_m^1), E_1(\underline{w}|_\Gamma)] + (f, E_1(\underline{w}|_\Gamma)_1) + \beta(g, E_1(\underline{w}|_\Gamma)_2)\ \forall\, \underline{w} \in \mathrm{V}_2,\ \psi \in \mathrm{L}^2(\Omega_2) \quad . \end{cases}$$

Obviously, the right hand side is a continuous linear functional on $\mathrm{V}_2 \times \mathrm{L}^2(\Omega_2)$.

3.7.2 The inviscid generalized Stokes problem

We consider now the inviscid counterpart of the generalized Stokes problem (3.75) which is

$$\begin{cases} \alpha\sigma + div\, \underline{u} = g \\ \alpha \underline{u} + \beta \nabla\sigma = f \quad . \end{cases} \tag{3.91}$$

It is obtained by setting $\mu = 0$ in (3.75). The boundary conditions associated to problem (3.91) are namely

$$\begin{aligned} &\underline{u} \cdot n = \underline{u}_\infty \cdot n \quad \text{on } \Gamma_\infty^- \\ &\underline{u} \cdot n = 0 \quad \text{on } \Gamma_B \\ &\quad \sigma = 0 \quad \text{on } \Gamma_\infty^+ \qquad . \end{aligned} \tag{3.92}$$

The weak formulation in this case consists to find $(\underline{u}, \sigma)$ such that $(\underline{u} - \underline{w}_\infty) \in W_0$ and $\sigma \in S_0$, where $\underline{w}_\infty \in \mathrm{L}^2(\Omega)$ with $div\, \underline{w}_\infty \in \mathrm{L}^2(\Omega)$ is such that $\underline{w}_\infty \cdot n = \underline{u}_\infty \cdot n$ on Γ_∞^- and $\underline{w}_\infty \cdot n = 0$ on Γ_B,

$$\begin{aligned} W_0 &\equiv \{\, \underline{v} \in \mathrm{L}^2(\Omega) \mid div\, \underline{v} \in \mathrm{L}^2(\Omega),\ \underline{v} \cdot n = 0 \text{ on } \Gamma_\infty^- \cup \Gamma_B \} \\ S_0 &\equiv \{ s \in \mathrm{H}^1(\Omega) \mid s = 0 \text{ on } \Gamma_\infty^+ \} \end{aligned} \tag{3.93}$$

and $(\underline{u}, \sigma)$ satisfies

$$\begin{cases} \displaystyle\int_\Omega (\alpha\sigma + div\, \underline{u})\varphi\, d\Omega = \int_\Omega g\varphi\ d\Omega\ ,\ \forall\, \varphi \in \mathrm{L}^2(\Omega) \\ \displaystyle\int_\Omega (\alpha \underline{u}\, \underline{v} - \beta\sigma div\, \underline{v})\, d\Omega = \int_\Omega f \underline{v}\ d\Omega\ ,\ \forall\, \underline{v} \in W_0 \end{cases} \tag{3.94}$$

Now, the multi-domain formulation in the case $\underline{u}_\infty = 0$ of the inviscid generalized Stokes problem is the following : define

$$
\begin{aligned}
W_1 &\equiv \{\, \underline{v} \in \mathrm{L}^2(\Omega_1) \mid div\, \underline{v} \in \mathrm{L}^2(\Omega_1),\ \underline{v}\cdot n = 0 \text{ on } \Gamma_\infty^- \,\} \\
W_{0,1} &\equiv \{\, \underline{v} \in W_1 \mid\ \underline{v}\cdot n = 0 \text{ on } \Gamma \,\} \\
W_2 &\equiv \{\, \underline{v} \in \mathrm{L}^2(\Omega_2) \mid\ div\, \underline{v} \in \mathrm{L}^2(\Omega_2),\ \underline{v}\cdot n = 0 \text{ on } \Gamma_B \,\} \\
W_{0,2} &\equiv \{\, \underline{v} \in W_2 \mid\ \underline{v}\cdot n = 0 \text{ on } \Gamma \,\} \\
S_1 &\equiv \{\sigma \in \mathrm{H}^1(\Omega_1) \mid\ \sigma = 0 \text{ on } \Gamma_\infty^+ \,\} \\
S_2 &\equiv \mathrm{H}^1(\Omega_2) \quad ,
\end{aligned}
$$

then, find $\underline{u}^1 \in W_1$, $\sigma^1 \in S_1$, $\underline{u}^2 \in W_2$, $\sigma^2 \in S_2$ such that

$$
\begin{aligned}
\int_{\Omega_1} (\alpha\sigma^1 + div\, \underline{u}^1)\varphi\, d\Omega &= \int_{\Omega_1} g\varphi\, d\Omega \quad , \forall\ \varphi \in \mathrm{L}^2(\Omega_1) \\
\int (\alpha \underline{u}^1 \underline{v} - \beta\sigma^1 div\, \underline{v})\, d\Omega &= \int_{\Omega_1} f\underline{v}\, d\Omega \quad , \forall\ v \in W_{0,1} \\
\underline{u}^1 \cdot n &= \underline{u}^2 \cdot n \ \text{ on } \ \Gamma \\
\int_{\Omega_2} (\alpha\sigma^2 + div\, \underline{u}^2)\psi\, d\Omega &= \int_{\Omega_2} g\psi\, d\Omega \quad , \forall\ \psi \in \mathrm{L}^2(\Omega_2) \\
\int_{\Omega_2} (\alpha \underline{u}^2 \underline{w} - \beta\sigma^2 div\, \underline{w})\, d\Omega &= \int_{\Omega_2} f\underline{w}\, d\Omega, \quad \forall\ \underline{w} \in W_{0,2}) \\
\int_{\Omega_2} (\alpha \underline{u}^2 R_2\chi - \beta\sigma^2 div\, R_2\chi)\, d\Omega &= \int_{\Omega_2} f\, R_2\chi d\Omega + \int_{\Omega_1} f\, R_1\chi\, d\Omega - \\
\int_{\Omega_1} (\alpha \underline{u}^1 R_1\chi - \beta\sigma^1 div\ R_1\chi)\, d\Omega &\quad , \forall\ \chi \in \mathrm{H}^{\frac{1}{2}}(\Gamma) \quad ,
\end{aligned}
\tag{3.95}
$$

where R_1 and R_2 are any possible pair of (continuous) operators from $\mathrm{H}^{-\frac{1}{2}}(\Gamma)$ to W_1 and W_2 respectively which satisfy $R_k\chi = \chi$ on Γ, $k = 1, 2$, (n is the unit normal vector to Γ which is directed from Ω_1 to Ω_2).
In the same way as the procedure used in the viscous case, we show that (3.91) and (3.92) are equivalent to (3.95). The only point that needs to be specified here is about the matching conditions on Γ. Since $\underline{u}^1 \cdot n = \underline{u}^2 \cdot n$ on Γ (see $(3.95)_3$), then it follows that the function $\underline{u}$ built up as in (3.82) belongs to the space W_0 defined in (3.93). On the other hand, the last condition in (3.93) states that $\sigma_1 = \sigma_2$ on Γ, hence if we construct σ as in (3.82), such a function belongs to the space S_0 defined in (3.93).
Also as in the viscous case, let us introduce now a suitable iterative procedure for the solution of the multidomain problem. At each step, we proceed as follows :

- In Ω_1, solve (3.91) with the boundary conditions

$$
\begin{cases}
\underline{u}_m^1 \cdot n = \underline{u}_\infty \cdot n \quad \text{on } \Gamma_\infty^- \\
\quad \sigma_m^1 = 0 \quad \text{on } \Gamma_\infty^+ \\
\underline{u}_m^1 \cdot n = \theta_m \underline{u}_{m-1}^2 \cdot n + (1-\theta_m)\underline{u}_{m-1}^1 \cdot n \quad \text{on } \Gamma
\end{cases}
\tag{3.96}
$$

- Then in Ω_2, solve the problem (3.91) with the following boundary conditions :

$$
\begin{cases}
\quad \sigma_m^2 = \sigma_m^1 \quad \text{on } \Gamma \\
\underline{u}_m^2 \cdot n = 0 \quad \text{on } \Gamma_B \quad .
\end{cases}
\tag{3.97}
$$

We continue by just underlying the differences between this case and the viscous case. We define at first the following bilinear forms

$$b_1[(\underline{u}^1,\sigma^1),(\underline{v},\varphi)] \equiv \int_{\Omega_1} (\beta\alpha\sigma^1\varphi + \beta div\,\underline{u}^1\varphi + \alpha\underline{u}^1\underline{v} - \beta\sigma^1 div\,\underline{v})\,d\Omega \tag{3.98}$$

$$b_2[(\underline{u}^2,\sigma^2),(\underline{w},\psi)] \equiv \int_{\Omega_2} (\beta\alpha\sigma^2\psi + \beta div\,\underline{u}^2\psi + \alpha\underline{u}^2\underline{w} - \beta\sigma^2 div\,\underline{w})\,d\Omega \tag{3.99}$$

which correspond to (3.85) and (3.86) in the degenerate case $\mu = 0$. It can be shown that these forms are continuous in $\mathrm{H}(div\,;\Omega_k) \equiv \{\,\underline{u} \in \mathrm{L}^2(\Omega_k) \mid div\,\underline{u} \in \mathrm{L}^2(\Omega_k)\,\} \times \mathrm{L}^2(\Omega_k)$, but they are not coercive. Then define the (continuous) operators $F_k : \mathrm{H}^{-\frac{1}{2}}(\Gamma) \to \mathrm{W}_k \times S_k$ in the following way : $F_1\chi \in \mathrm{W}_1 \times S_1$,

$$\begin{cases} b_1[F_1\chi,(\underline{v},\varphi)] = 0\ \forall\,\underline{v} \in \mathrm{W}_{0,1}\,,\ \varphi \in \mathrm{L}^2(\Omega_1), \\ \quad (F_1\chi)_{1|\Gamma}\cdot n = \chi, \end{cases} \tag{3.100}$$

and $F_2\chi \in \mathrm{W}_2 \times S_2$,

$$\begin{cases} b_2[F_2\chi,(\underline{w},\psi)] = 0\ \forall\,\underline{w} \in \mathrm{W}_{0,2}\,,\ \psi \in \mathrm{L}^2(\Omega_2), \\ \quad (F_2\chi)_{1|\Gamma}\cdot n = \chi. \end{cases} \tag{3.101}$$

3.7.3 Convergence estimates

For the viscous and the inviscid cases, we have the following theorems, proved in [112].

Theorem 3.18. *There exists a positive constant $\theta^* \in\,]0,1]$ such that, if $\sup\theta_m < \theta^*$ and $\inf\theta_m > 0$, then the sequences $\underline{u}^1_m$, $\underline{u}^2_m$, σ^1_m, σ^2_m converge to the solution $\underline{u}^1$, $\underline{u}^2$, σ^1, σ^2 of problem (3.81). Convergence is in the $\mathrm{H}^1(\Omega)$-norm for velocity and $\mathrm{L}^2(\Omega)$-norm for density.*

Theorem 3.19. *There exists a positive constant $\theta^* \in]0,1]$ such that, if $\sup\theta_m < \theta^*$ and $\inf\theta_m > 0$, then the sequences $\underline{u}^1_m$, $\underline{u}^2_m$,σ^1_m, σ^2_m converge to the solution $\underline{u}^1$, $\underline{u}^2$, σ^1, σ^2 of problem (3.95). Namely, we have that $\underline{u}^k_m$ and $div\,\underline{u}^k_m$ convergence in $\mathrm{L}^2(\Omega_k)$ to $\underline{u}^k$ and $div\,\underline{u}^k$, respectively, and that σ^k_m converges in $\mathrm{H}^1(\Omega_k)$ to σ^k.*

3.8 Numerical experiences

We provide some numerical experiments to illustrate the performance of the overlapping additive Schwarz preconditioner.

Description of the preconditioner

Assume at first that our computational domain Ω is subdivided into a finite number of non-overlapping sub-domains Ω_i see Figure 3.12. Then for all the nodes belonging to the internal boundary of each sub-domains Ω_i, we find all the elements which have these nodes in common and add theses elements to each non-overlapping sub-domains Ω_i considered. After adding these elements, for each non-overlapping sub-domain Ω_i we obtain a new sub-domain which is an overlapping sub-domain called $\tilde{\Omega}_i$ see Figure 3.13. In fact, in the Figure 3.13, the union of the two coloured regions is the overlapping sub-domain $\tilde{\Omega}_i$ with respect to the initial non-overlapping sub-domain Ω_i. The darker region represents the strip of overlapping. Whenever the overlapping sub-domain $\tilde{\Omega}_i$ is obtained, we have to apply the following algorithm in order to have the desired preconditioner. We have:

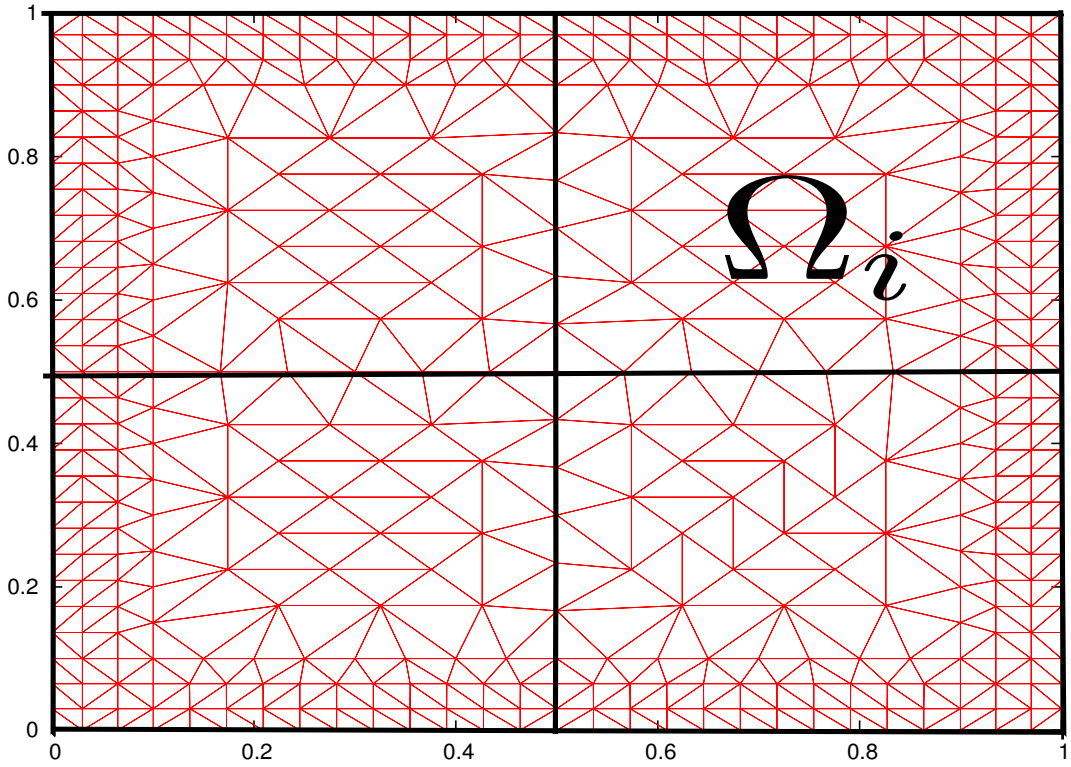

Figure 3.12: Nonoverlapping subdomain Ω_i .

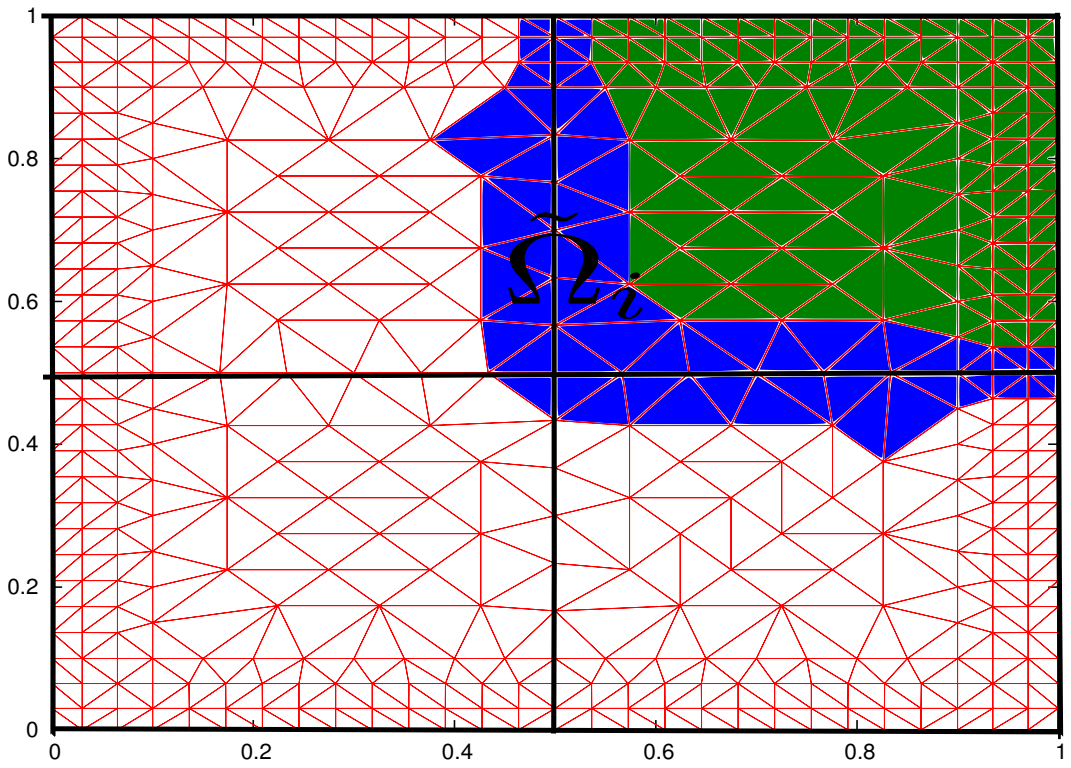

Figure 3.13: Overlapping sub-domain $\tilde{\Omega}_i$.

- To construct the vector I_i of the position of all the node of $\tilde{\Omega}_i$ in A, where A represents the final matrix of the algebraic system (the general stiffness matrix). As our aim is to define a new approach for the solution of the Navier-Stokes equations, the matrix A will represent the matrix of the algebraic system when solving

- □ for the intermediate velocity field
- □ for the updated pressure field
- □ for the correction velocity field
- □ for the temperature field

- To construct R_i and $\mathrm{R}_i^{\mathrm{T}}$, respectively restriction and extension operator
- To compute $\mathrm{A}_i = \mathrm{R}_i \mathrm{A} \mathrm{R}_i^{\mathrm{T}}$ the local sub-matrix with respect to each overlapping sub-domain $\tilde{\Omega}_i$ (the dimension of A_i is given by the node's number of $\tilde{\Omega}_i$)
- To make the Incomplete Lower-Upper (ILU) factorization for each local sub-matrix A_i and use this factorization to compute the approximated inverse matrix A_i^{-1} of the local sub-matrix A_i
- To compute $\mathrm{T}_i = \mathrm{R}_i^{\mathrm{T}} \mathrm{A}_i^{-1} \mathrm{R}_i$ the preconditioner matrix with respect to each local sub-matrix A_i (the dimension of T_i is the same as A)
- To compute $\mathrm{P}_i = \mathrm{T}_i \mathrm{A}$, the preconditioned matrix with respect to each local sub-matrix A_i .
- To compute $\mathcal{P} = \sum_{i=0}^{M} \mathrm{P}_i$ (respectively $\mathcal{Q} = \sum_{i=0}^{M} \mathrm{T}_i$) the final preconditioned matrix (respectively the final preconditioner) desired, where M being the number of all the overlapping sub-domains, plus the coarse mesh
- To compute the final right hand term by the product $\mathcal{Q}\mathrm{b} = \mathrm{c}$; at this point we have transformed the initial algebraic system $\mathrm{A}u_h = \mathrm{b}$ in the equivalent well-conditioned system $\mathcal{P}u_h = \mathrm{c}$.

Remark 3.20. *The coarse mesh is obtained using the sub-domains Ω_i, in particular knowing the number n_i of vertices of the polygonal boundary $\partial\Omega_i$ of Ω_i we subdivide each Ω_i in the union of $n_i - 2$ "big" triangles. Theses big triangles are easily obtained fixing the attention on a vertex V_e of the n_i vertices of $\partial\Omega_i$ and drawing the diagonals of the polygon $\partial\Omega_i$ starting from the vertex V_e (see Figure 3.14). Finally, the coarse mesh is obtained by the union of all the big triangles of all the sub-domains Ω_i.*

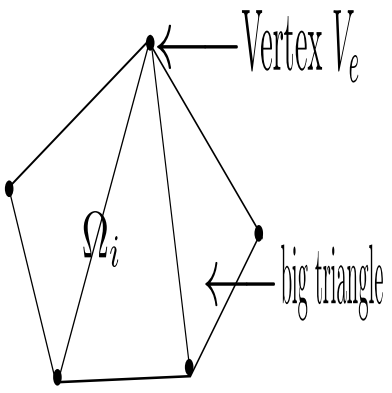

Figure 3.14: Sub-domain Ω_i and its big triangles.

Implementation and results

We have implemented the preconditioner and have tested its efficiency solving the same problems of the tests of chapter 2 section 2.4 using the Bi-CGSTAB and the preconditioned additive Schwarz Bi-CGSTAB iterative methods. We note that the inversion of all the local sub-matrices

by means of ILU factorization have been performed taking the fill-in parameter equal to the maximum number of nodes belonging to the support of each basis functions of $\tilde{\Omega}_i$. With the fixed tolerance $(1.E-13)$, we obtained the same accuracy reported in Tables 2.3 and 2.4. We denote by N_{SC} the total number of sub-domains plus the coarse mesh.
In the following Tables we denote by:

- $iter(\mathrm{A}_h)$ the number of iterations without using the Schwarz additive preconditioner
- $iter(\mathrm{P}_{cas}\mathrm{A}_h)$ the number of iterations using the Schwarz additive preconditioner
- $\mathcal{K}(\mathrm{A}_h)$ the condition number of the matrix not preconditioned
- $\mathcal{K}(\mathrm{P}_{cas}\mathrm{A}_h)$ the condition number of the matrix preconditioned with the Schwarz additive preconditioner.

Test 3.8.1

We have considered the problem of the Test 2.4.3.1 and have reported in the Table 3.1 the number of iterations of the algebraic system when using the non-symmetric matrix A_h with and without the additive Schwarz preconditioner obtained in the new conservative FE case with 10 sub-domains(coarse mesh included). Also, we have reported in the table 3.2 (respectively Table 3.3) the number of iterations and the condition number of the symmetric matrix A_h of the traditional FE case, when it is not preconditioned and when it is preconditioned using 10 (respectively 5) sub-domains (coarse mesh included).

Table 3.1: Iteration number with $N_{SC} = 10$.

	Conservative FE	
N_h	$\neq iter(\mathrm{A}_h)$	$\neq iter(\mathrm{P}_{cas}\mathrm{A}_h)$
49	26	16
81	52	22
137	54	25
169	36	22

Table 3.2: Iteration and condition numbers with $N_{SC} = 10$.

	Traditional FE			
N_h	$\neq iter(\mathrm{A}_h)$	$\mathcal{K}(\mathrm{A}_h)$	$\neq iter(\mathrm{P}_{cas}\mathrm{A}_h)$	$\mathcal{K}(\mathrm{P}_{cas}\mathrm{A}_h)$
49	16	$1.842E+1$	13	$3.266E+0$
81	32	$1.702E+1$	16	$6.586E+0$
137	35	$1.534E+1$	15	$1.628E+1$
169	24	$6.336E+1$	17	$1.304E+1$

Table 3.3: Iteration and condition numbers with $N_{SC} = 5$.

	Traditional FE			
N_h	$\neq iter(\mathrm{A}_h)$	$\mathcal{K}(\mathrm{A}_h)$	$\neq iter(\mathrm{P}_{cas}\mathrm{A}_h)$	$\mathcal{K}(\mathrm{P}_{cas}\mathrm{A}_h)$
81	32	$1.702E+1$	18	$8.330E+0$
137	35	$5.134E+1$	17	$2.025E+1$
169	24	$6.336E+1$	17	$1.512E+1$

Test 3.8.2

We have considered the problem of the Test 2.4.3.2 and have reported in Table 3.4 the number of iterations when using the non-symmetric matrix A_h with and without the additive Schwarz preconditioner obtained in the new conservative FE with 10 sub-domains (coarse mesh included). Also, we have reported in Table 3.5 (respectively Table 3.6) the iteration and the condition numbers of the symmetric matrix A_h in the traditional FE case when it is not preconditioned and when it is preconditioner using 10 (respectively 5) sub-domains (coarse mesh included) .

Table 3.4: Iteration number with $N_{SC} = 10$.

	Conservative FE	
N_h	$\neq iter(\mathrm{A}_h)$	$\neq iter(\mathrm{P}_{cas}\mathrm{A}_h)$
49	42	23
81	76	39
137	99	44
169	70	49

Table 3.5: Iteration and condition numbers with $N_{SC} = 10$.

	Traditional FE			
N_h	$\neq iter(\mathrm{A}_h)$	$\mathcal{K}(\mathrm{A}_h)$	$\neq iter(\mathrm{P}_{cas}\mathrm{A}_h)$	$\mathcal{K}(\mathrm{P}_{cas}\mathrm{A}_h)$
49	34	$4.530E+1$	15	$5.854E+0$
81	49	$4.732E+1$	17	$1.025E+1$
137	61	$1.254E+2$	19	$3.649E+1$
169	47	$1.580E+2$	19	$5.405E+1$

Table 3.6: Iteration and condition numbers with $N_{SC} = 5$.

	Traditional FE			
N_h	$\neq iter(\mathrm{A}_h)$	$\mathcal{K}(\mathrm{A}_h)$	$\neq iter(\mathrm{P}_{cas}\mathrm{A}_h)$	$\mathcal{K}(\mathrm{P}_{cas}\mathrm{A}_h)$
81	49	$4.732E+1$	20	$1.732E+1$
137	61	$1.254E+2$	18	$4.233E+1$
169	47	$1.580E+2$	18	$3.277E+1$

Remark 3.21. *Checking Tables (3.2), (3.3), (3.5), (3.6), in the case of the traditional FE, we conclude that the theoretically foreseen scalability property is respected.*

Remark 3.22. *It has been demonstrated [1] that in presence of a non symmetric matrix* A_h *of the system (2.56), the scalability property of the Schwarz preconditioner is respected by the symmetric part of* A_h *under suitable conditions.*

3.9 Conclusions of the chapter

- Using the Bi-CGSTAB solver with stopping criteria of the residual value $\leq 10E-13$, the numerical solution coincides with the one obtained by the Gauss method up to eight decimal digit
- The error norms of solutions of chapters 2 and 3 are equals up to eight decimal digit

- The fluxes conservation is guaranteed independently of the number of sub-domains
- The accuracy of the solution is independent of the sub-domains decompositions (scalability property)
- The Schwarz preconditioner reduces substantially (in particular for symmetric matrices) the number of iterations necessary to obtain the desired accuracy.

Chapter 4

Parabolic and convection diffusion problems

This chapter is devoted to the approximation of the second order parabolic and convection-diffusion equations. We will present a short review of the theory concerning the existence and uniqueness to these initial boundary value problems; then we will conclude by presenting our advancing time scheme for the solution of convection diffusion problems since it is one of the equations involved in the solution of the 2D Navier-Stokes equations that we are interested in.

4.1 Initial boundary value problems and weak formulation

Let us assume that Ω is a bounded domain in $\mathbb{R}^2$ with Lipschitz boundary. We consider a second order differential operator L given by:

$$Lu := -\sum_{i,j=1}^{2} D_i(a_{i,j}D_j u) + \sum_{i=1}^{2}[D_i(b_i u) + c_i D_i u] + a_0 u \tag{4.1}$$

We give these two definitions.

Definition 4.1. *The operator L is elliptic if there exists a constant $\alpha_0 > 0$ such that*

$$\sum_{i,j=1}^{2} a_{i,j}(x)\xi_i\xi_j \geq \alpha_0|\xi|^2 \tag{4.2}$$

for each $\xi \in \mathbb{R}^2$ and for almost every $x \in \Omega$.

Definition 4.2. *The operator*

$$\frac{\partial}{\partial t} + L \tag{4.3}$$

is said to be parabolic if L is elliptic.

We note that the coefficients of the operator L (namely $a_{i,j}$, b_i, c_i and a_0) do not depend on t. An example is provided by the heat equation $\frac{\partial u}{\partial t} - \Delta u = f$ in which $L = -\Delta$.
We indicate by $Bu = g$ any of the boundary conditions considered in Chapter 1. In order to simplify the discussion, we will consider only the case of Dirichlet homogeneous boundary condition, that is $u = 0$ on $\Sigma_T := (0,T) \times \partial\Omega$.
Since the problem contains the time derivative operator, an initial condition has to be introduced in order to determine the solution u. Thus, the following initial boundary value problem holds:

$$\begin{aligned} \frac{\partial u}{\partial t} + Lu = f \quad &\text{in} \quad \mathcal{Q}_T := (0,T) \times \Omega \\ u = 0 \quad &\text{on} \quad \Sigma_T := (0,T) \times \partial\Omega \\ u|_{t=0} = u_0 \quad &\text{on} \quad \Omega\,, \end{aligned} \tag{4.4}$$

where $f = f(t,x)$, and $u_0 = u_0(x)$ are given data.
We give a weak formulation of the problem. We denote $\mathsf{V} := \mathrm{H}_0^1(\Omega)$. We also denote by $\mathrm{L}^2(0,T;\mathsf{V})$ the space

$$\mathrm{L}^2(0,T;\mathsf{V}) := \{ \quad u : (0,T) \to \mathsf{V} : u \text{ is measurable and } \int_0^T \|u(t)\|_1^2 \, dt < \infty \quad \}$$

and similarly we define $\mathcal{C}^0([0,T];\mathrm{L}^2(\Omega))$.
The weak formulation of (4.4) is the following: given $f \in \mathrm{L}^2(\mathcal{Q}_T)$ and $u_0 \in \mathrm{L}^2(\Omega)$, find $u \in \mathrm{L}^2(0,T;\mathsf{V}) \cap \mathcal{C}^0([0,T];\mathrm{L}^2(\Omega))$ such that

$$\begin{aligned} \frac{d}{dt}(u(t),v) + a(u(t),v) &= (f(t),v) \quad \forall\, v \in \mathsf{V} \\ u(0) &= u_0 \end{aligned} \tag{4.5}$$

where $(\cdot\,,\cdot)$ denotes the scalar product in $\mathrm{L}^2(\Omega)$, the bilinear form $a(\cdot\,,\cdot)$ is

$$a(w,v) := \int_\Omega [\sum_{i,j=1}^2 a_{i,j} D_j w D_i v - \sum_{i=1}^2 (b_i w D_i v + c_i v D_i w) + a_0 w \cdot v].$$

In the case $L = -\Delta$, we have $a(w,v) = \int_\Omega \nabla w \cdot \nabla v \; dx$.

4.1.1 Mathematical analysis

Several methods can be used to prove the existence and uniqueness of a solution to (4.5). We are presenting the one based on the Faedo-Galerkin approach with suitable energy estimates. Let us assume that the bilinear form $a(\cdot\,,\cdot)$ is continuous and satisfies the Gårding inequality (sometimes also named weakly coerciveness) that is:

$$\text{there exists two constants } \alpha > 0 \text{ and } \lambda \geq 0 : \quad a(v,v) + \lambda \|v\|_0^2 \geq \alpha \|v\|_1^2 \quad \forall\, v \in \mathsf{V} \,. \tag{4.6}$$

More often it is satisfied taking $\lambda = 0$, that means the bilinear form $a(\cdot\,,\cdot)$ is strongly coercive, this is for example the case of the heat equation aforementioned. Let us notice that all norms refer to the space variables, i.e., $\|\cdot\|_k$ is the norm in the Sobolev space $\mathrm{H}^k(\Omega)$ for $k \geq 0$. In general, the inequality (4.6) is satisfied for all the boundary conditions mentioned in Chapter 1 provided that for each $i,j = 1,,2$, the coefficients $a_{i,j}$, b_i, c_i and a_0 of the operator L belong to $\mathrm{L}^\infty(\Omega)$. In fact using (4.2), we have

$$\begin{aligned} a(v,v) &= \int_\Omega [\sum_{i,j} a_{i,j} D_i v \cdot D_j v - \sum_i (b_i - c_i) v D_i v + a_0 v^2] \; d\Omega \\ &\geq \beta_0 \|Dv\|_0^2 - \|b - c\|_{\mathrm{L}^\infty(\Omega)} \|Dv\|_0 \|v\|_0 + \|a_0\|_{\mathrm{L}^\infty(\Omega)} \|v\|_0^2 \end{aligned}$$

but using this arithmetic-geometric mean inequality for each $\epsilon > 0$

$$\|Dv\|_0 \|v\|_0 \leq \epsilon \|Dv\|_0^2 + \frac{1}{4\epsilon} \|v\|_0^2$$

It follows that (4.6) holds choosing for instance

$$\lambda > C \frac{1}{\alpha_0} (\|b - c\|_{\mathrm{L}^\infty(\Omega)} - \|a_0\|_{\mathrm{L}^\infty(\Omega)})$$

where $C = C(2,\Omega)$ is a suitable constant.
We have this existence theorem.

Theorem 4.3. *Assume that the bilinear form $a(\cdot\,,\cdot)$ is continuous in $\mathsf{V}\times\mathsf{V}$ and that (4.6) is satisfied with $\lambda = 0$. Then given $f \in \mathrm{L}^2(\mathcal{Q}_T)$ and $u_0 \in \mathrm{L}^2(\Omega)$, there exists a unique solution $u \in \mathrm{L}^2(0,T;\mathsf{V})\cap\mathcal{C}^0([0,T];\mathrm{L}^2(\Omega))$ to (4.5). Moreover, $\frac{\partial u}{\partial t} \in \mathrm{L}^2(0,T;\mathsf{V}')$ and the energy estimate*

$$\|u(t)\|_0^2 + \alpha\int_0^t \|u(\tau)\|_1^2 d\tau \le \|u_0\|_0^2 + \frac{1}{\alpha}\int_0^t \|f(\tau)\|_0^2 d\tau \tag{4.7}$$

holds for each $t \in [0,T]$.

Proof. We are going to construct an approximate sequence solving suitable finite dimensional problems. Since V is a closed subspace of the Hilbert space $\mathrm{H}^1(\Omega)$, it is a separable Hilbert space. Let $\{\phi_j\}_{j\ge1}$ be a complete orthonormal basis in V and define $\mathsf{V}^N := span\{\phi_1,\cdots,\phi_N\}$. We consider the approximate problem: for each $t\in[0,T]$, find $u^N(t)\in\mathsf{V}^N$ such that

$$\begin{aligned} &\frac{d}{dt}(u^N(t),\phi_j) + a(u^N(t),\phi_j) = (f(t),\phi_j), \quad \forall\ j=1,\cdots,N, \quad t\in(0,T)\\ &u^N(0) = u_0^N := P_N(u_0) = \sum_{s=1}^N \rho_s\phi_s \end{aligned} \tag{4.8}$$

where P_N is the orthogonal projection in $\mathrm{L}^2(\Omega)$ on V^N, hence the vector $\underline{\rho}$ is the solution of the linear system $\mathrm{M}\underline{\rho} = (u_0,\phi_j)$, M being the mass matrix $\mathrm{M}_{js} := (\phi_j,\phi_s)$. Since $\{\phi_j\}, j=1,\cdots,N$ is a basis for V^N, the equation in (4.8) is therefore satisfied for each $v^N\in\mathsf{V}^N$.
Writing $u^N(t) = \sum_{s=1}^N C_s^N(t)\phi_s$, (4.8) is equivalent to solving

$$\begin{aligned} \mathrm{M}\frac{d}{dt}C^N(t) + \mathrm{A}C^N(t) &= F(t)\\ \mathrm{M}C^N(0) &= C_0 \end{aligned} \tag{4.9}$$

where for $i,j=1,\cdots,N$, $\mathrm{A}_{i,j} := a(\phi_j,\phi_i)$, $F_i(t) := (f(t),\phi_i)$, $C_{0,i} := (u_0,\phi_i)$. Since M is positive definite, we find a unique solution C^N to (4.9). Since $F\in\mathrm{L}^2(0,T)$, it follows that $C^N\in\mathrm{H}^1(0,T)$ i.e $u^N\in\mathrm{H}^1(0,T;\mathsf{V})$. Choosing $u^N(t)$ in (4.8) as a test function, we have

$$(\frac{d}{dt}u^N(t),u^N(t)) + a(u^N(t),u^N(t)) = (f(t),u^N(t))$$

and therefore, using (4.6) it comes

$$\frac{1}{2}\frac{d}{dt}\|u^N(t)\|_0^2 + \alpha\|u^N(t)\|_1^2 \le \|f(t)\|_0\|u^N(t)\|_0.$$

Applying the arithmetic-mean inequality for each $\epsilon>0$ it comes

$$\frac{1}{2}\frac{d}{dt}\|u^N(t)\|_0^2 + \alpha\|u^N(t)\|_1^2 \le \epsilon\|u^N(t)\|_0^2 + \frac{1}{\epsilon}\|f(t)\|_0^2.$$

Integrating over $(0,\tau)$, $\tau\in(0,T]$, it comes by taking $\epsilon=\alpha$

$$\|u^N(\tau)\|_0^2 + 2\alpha\int_0^\tau \|u^N(t)\|_1^2 dt \le \|u_0\|_0^2 + 2\alpha\int_0^\tau \|u^N(t)\|_0^2 dt + \frac{2}{\alpha}\int_0^\tau \|f(t)\|_0^2 dt$$

i.e

$$\|u^N(\tau)\|_0^2 + 2\alpha[\int_0^\tau (\|u^N(t)\|_1^2 - \|u^N(t)\|_0^2)dt] \le \|u_0\|_0^2 + \frac{2}{\alpha}\int_0^\tau \|f(t)\|_0^2 dt$$

i.e

$$\|u^N(\tau)\|_0^2 + 2\alpha \int_0^\tau \|Du^N(t)\|_0^2 dt \leq \|u_0\|_0^2 + \frac{2}{\alpha}\int_0^\tau \|f(t)\|_0^2 dt$$

and we obtain

$$\|u^N(\tau)\|_0^2 + C_1\alpha \int_0^\tau \|u^N(t)\|_1^2 dt \leq \|u_0\|_0^2 + C_2\frac{1}{\alpha}\int_0^\tau \|f(t)\|_0^2 dt \tag{4.10}$$

Thus the sequence u^N is bounded in $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega)) \cap \mathrm{L}^2(0,T;\mathsf{V})$. Hence by the compact embedding theorem, there exists a subsequence (still denoted by u^N) which converges to the weak star topology of $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$ and weakly in $\mathrm{L}^2(0,T;\mathsf{V})$ (see [136]). This means that there exists $u \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega)) \cap \mathrm{L}^2(0,T;\mathsf{V})$ such that

$$\int_0^T (u^N(t),\varphi(t))dt \longrightarrow \int_0^T (u(t),\varphi(t))dt \quad \text{as} \quad N \to \infty$$

for each $\varphi \in \mathrm{L}^1(0,T;\mathrm{L}^2(\Omega))$ and

$$\int_0^T (\nabla u^N(t),\phi(t))dt \longrightarrow \int_0^T (\nabla u(t),\phi(t))dt \quad \text{as} \quad N \to \infty$$

for each $\phi \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$. In order to pass to the limit in (4.8), we take $\Psi \in \mathcal{C}^1([0,T])$ with $\Psi(T) = 0$. By multiplying (4.8) by Ψ and integrating by parts over $(0,T)$, the first term at the left hand side gives (since $\Psi(T) = 0$)

$$\int_0^T (\frac{du}{dt}(t),\phi_j)\Psi dt = -\int_0^T (u^N(t),\phi_j)\frac{d\Psi}{dt}(t)dt - (u_0^N,\phi_j)\Psi(0)$$

Since u_0^N converges in $\mathrm{L}^2(\Omega)$ to u_0, by choosing arbitrarily N and passing to the limit in (4.8), we finally obtain

$$-\int_0^T (u(t),\phi_j)\frac{d\Psi}{dt}(t)dt - (u_0,\phi_j)\Psi(0) + \int_0^T a(u(t),\phi_j)\Psi(t)dt = \int_0^T (f(t),\phi_j)\Psi(t)dt, \quad \forall\ j = 1,\cdots,N\,. \tag{4.11}$$

Since the linear combinations of ϕ_j are dense in V, we can also write (4.11) for each $v \in \mathsf{V}$. Moreover, taking $\Psi \in \mathcal{D}(0,T)$, (4.11) is nothing else that (4.5). Thus we have constructed a solution $u \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega)) \cap \mathrm{L}^2(0,T;\mathsf{V})$ of problem (4.5).
It remains to show that $u(0) = u_0$ and $\frac{\partial u}{\partial t} \in \mathrm{L}^2(0,T;\mathsf{V}')$. Let us recall in fact that from the latter result it follows that $u \in \mathrm{L}^2(0,T;\mathsf{V}) \cap \mathrm{H}^1(0,T;\mathsf{V}')$ thus $u \in \mathcal{C}^0([0,T];\mathrm{L}^2(\Omega))$ see [86].
Now we prove that $u(0) = u_0$. From (4.5), we have that $(u(t),v) \in \mathrm{H}^1(0,T)$, hence it is a continuous function on $[0,T]$. Multiplying (4.5) by $\Psi \in \mathcal{C}^1([0,T])$ with $\Psi(T) = 0$ and integrating by parts it comes

$$-\int_0^T (u(t),v)\frac{d\Psi}{dt}(t)dt - (u(0),v)\Psi(0) + \int_0^T a(u(t),v)\Psi(t)dt = \int_0^T (f(t),v)\Psi(t)dt \quad \forall\ v \in \mathsf{V}$$

thus, taking $\Psi(0) = 1$, we get $(u(0) - u_0, v) = 0 \quad \forall\, v \in \mathsf{V}$ this implies that $u(0) = u_0$.
Finally from (4.5) it follows that

$$\frac{\partial u}{\partial t} + Lu = f \tag{4.12}$$

in the sense of distributions of $\mathcal{Q}_T$. Since $Lu \in \mathrm{L}^2(0,T;\mathsf{V}')$ we find that $\frac{\partial u}{\partial t} \in \mathrm{L}^2(0,T;\mathsf{V}')$. The uniqueness of the solution is a consequence of (4.7) which is indeed an a-priori estimate for any solution $u \in \mathrm{L}^2(0,T;\mathsf{V}) \cap \mathrm{H}^1(0,T;\mathsf{V}')$ to (4.5). Rewriting (4.5) as

$$\langle \frac{\partial u}{\partial t}(t), v\rangle + a(u(t),v) = (f(t),v) \quad \forall\, v \in \mathsf{V}$$

where $\langle \cdot\,,\cdot \rangle$ denotes the duality pairing between V' and V. Taking $v = u(t)$, it follows from (see for instance [86] or [131]) that

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|_0^2 + a(u(t),u(t)) = (f(t),u(t))$$

and thus (4.7) is obtained by proceedings as in the proof of (4.10). □

Remark 4.4. *To prove the existence of a solution an approach based on the semi-group theory can be used, in which the solution u to (4.4) is formally given by*

$$u(t) = \exp(-t\mathrm{A})u_0 + \int_0^t \exp(-(t-s)\mathrm{A})f(s)\; ds \tag{4.13}$$

where A *represents a suitable "realization"of L with the associated boundary condition $u = 0$ on $\partial\Omega$. We refer to [102] for details on this approach.*

We have this a-priori estimate:

Proposition 4.5. *Assume (4.6) is satisfied with $\lambda = 0$ and that $f \in \mathrm{L}^2(\mathcal{Q}_T)$, $u_0 \in \mathsf{V}$, $a_{i,j}, b_i \in \mathcal{C}^1(\overline{\Omega})$ and $c_i, a_0 \in \mathrm{L}^\infty(\Omega)$. Then the solution u to (4.5) belongs to $\mathrm{L}^\infty(0,T;\mathsf{V}) \cap \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))$ and satisfies the energy estimate*

$$\sup_{t\in(0,T)} \|u(t)\|_1^2 + \int_0^T \|\frac{\partial u}{\partial t}(t)\|_0^2 dt \le C_\alpha(\|u_0\|_1^2 + \int_0^T \|f(t)\|_0^2 dt) \tag{4.14}$$

where $C_\alpha > 0$ is a constant independent of T.

The proof of this result can be found in [113].

Corollary 4.6. *Assume that the solution u obtained in Proposition 4.5 satisfies*

$$\|u(t)\|_2^2 \le C(\|Lu(t)\|_0^2 + \|u(t)\|_1^2) \quad \textit{almost everywhere in } \ [0,T] \tag{4.15}$$

then $u \in \mathrm{L}^2(0,T;\mathrm{H}^2(\Omega)) \cap \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega)) \cap \mathcal{C}^0([0,T];\mathsf{V})$ and satisfies the estimate

$$\max_{t\in[0,T]} \|u(t)\|_1^2 + \int_0^T (\|\frac{\partial u}{\partial t}(t)\|_0^2 + \|u(t)\|_2^2) dt \le C_\alpha(\|u_0\|_1^2 + \int_0^T \|f(t)\|_0^2 dt) \tag{4.16}$$

Proof. Estimate (4.16) follows from (4.15), (4.14) and (4.7), since $L\underline{u} = f - \frac{\partial u}{\partial t}$. Moreover, by using an interpolation (see [86]) one has that $\mathrm{L}^2(0,T;\mathrm{H}^2(\Omega)) \cap \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega)) \subset \mathcal{C}^0([0,T];\mathrm{H}^1(\Omega))$. □

Remark 4.7. *The assumption (4.15) is satisfied for homogeneous Dirichlet problem if $\partial\Omega \in \mathcal{C}^2$ see [86] or if Ω is a plane convex polygonal domain see [52].*

We are giving some methods of approximation to the solution to (4.5).

4.1.2 Semi-discrete approximation by FE

A step towards the approximation of the solution to (4.5) entails the discretization of the space variable only. It leads to a system of ordinary differential equations whose solution $u_h(t)$ is an approximation of the exact solution for each $t \in [0,T]$. We will focus only on the finite element case. For other spatial discretizations such as the spectral collocation or FD methods, we suggest the books [113] and [122] respectively. The problem by variational formulation (4.5) leads to a semi-discrete problem by approximating the space V by a finite dimensional space V_h. The semi-discrete problem is the following: given $f \in \mathrm{L}^2(\mathcal{Q}_T)$ and $u_{0,h} \in \mathsf{V}_h$, a suitable approximation of the initial datum $u_0 \in \mathrm{L}^2(\Omega)$, for each $t \in [0,T]$, find $u_h(t) \in \mathsf{V}_h$ such that

$$\begin{aligned} \frac{d}{dt}(u_h(t), v_h) + a(u_h(t), v_h) = (f(t), v_h) \quad \forall\ \underline{v_h} \in \mathsf{V}_h,\ t \in (0,T) \\ u_0(0) = u_{0,h}\ . \end{aligned} \tag{4.17}$$

We have assumed that Ω is a polygonal domain with Lipschitz boundary and the boundary condition is of homogeneous Dirichlet kind and so V_h is a finite dimensional subspace of $\mathrm{H}^1_0(\Omega)$. Writing $u_h(t) = \sum_j \varepsilon_j \varphi_j$, where $\{\varphi_j\}_{j=1}^{N_h}$ is a basis of V_h and $u_{0,h} = \sum_j \varepsilon_{0,h}\varphi_j$, problem (4.17) can be written as

$$\begin{aligned} \mathrm{M}\frac{d}{dt}\varepsilon(t) + \mathrm{A}\varepsilon(t) = F(t) \\ \varepsilon(0) = \varepsilon_0 \end{aligned} \tag{4.18}$$

where $\mathrm{M}_{i,j} = (\mathrm{M}_{f_e})_{i,j} := (\varphi_i, \varphi_j)$, $\mathrm{A}_{i,j} = (\mathrm{A}_{f_e})_{i,j} := a(\varphi_j, \varphi_i)$, $F_i(t) := (f(t), \varphi_i)$, $i,j = 1, \cdots, N_h$.
Since M_{f_e} is positive definite, there exists a unique solution $\varepsilon(t)$ to (4.18). By repeating the proof of Theorem (4.3), we see that the solution u_h satisfies an energy estimate like (4.7) provided $u_{0,h}$ converges to u_0 in $\mathrm{L}^2(\Omega)$. This proves the stability of the method.
The convergence of u_h to u and an estimate of the order of convergence is given by this proposition.

Proposition 4.8. *Let $\mathcal{I}_h$ be a regular family of triangulations and assume that piecewise-linear or bilinear elements are used. Assume moreover that (4.6) (with $\lambda = 0$) and (4.15) are satisfied and that $f \in \mathrm{L}^2(\mathcal{Q}_T)$, $u_0 \in \mathsf{V}$, $a_{i,j}$, $b_i \in \mathcal{C}^1(\overline{\Omega})$, c_i, $a_0 \in \mathrm{L}^\infty(\Omega)$. Then the solutions u and u_h to (4.5) and (4.17) respectively satisfy*

$$\|u(t) - u_h(t)\|_0^2 + \alpha \int_0^t \|u(\tau) - u_h(\tau)\|_1^2 d\tau \le C_{\alpha,\gamma} h^{2k} N(u) e^t \quad \textit{for each} \quad t \in [0,T]\,, \tag{4.19}$$

where α is the coerciveness constant in (4.6), $N(u)$ is a suitable function depending on u and on $\frac{\partial u}{\partial t}$, γ is the continuity constant of the bilinear form $a(\cdot,\cdot)$ and $C_{\alpha,\gamma} > 0$ is a suitable constant independent of h (the parameter characteristic of the triangulation).

Proof. From the coercivity, we can write

$$\begin{aligned} \alpha\|u - u_h\|^2_{\mathrm{H}^1(\Omega)} &\le a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \quad \forall\ v_h \in \mathsf{V}_h. \end{aligned} \tag{4.20}$$

By subtraction between (4.5) and (4.17), it comes

$$\frac{d}{dt}(u - u_h, w_h) + a(u - u_h, w_h) = (\frac{\partial(u - u_h)}{\partial t}, w_h) + a((u - u_h), w_h) = 0$$

with $w_h = v_h - u_h$ i.e $a(u - u_h, w_h) = -(\frac{\partial(u - u_h)}{\partial t}, w_h)$. Therefore (4.20) becomes

$$\alpha\|u - u_h\|_1^2 \le a(u - u_h, u - v_h) - (\frac{\partial(u - u_h)}{\partial t}, w_h)\ . \tag{4.21}$$

We treat separately the terms in the right hand side of (4.21).

- Using the continuity of the bilinear form and the Young inequality, it comes

$$\begin{aligned} a(u-u_h, u-v_h) &\leq \gamma \|u-u_h\|_1 \|u-v_h\|_1 \\ &\leq \frac{\alpha}{2}\|u-u_h\|_1^2 + \frac{\gamma^2}{2\alpha}\|u-v_h\|_1^2 \ . \end{aligned}$$

- Writing w_h in the form $w_h = (v_h - u) + (u - u_h)$, it comes

$$-(\frac{\partial(u-u_h)}{\partial t}, w_h) = (\frac{\partial(u-u_h)}{\partial t}, u-v_h) - \frac{1}{2}\frac{d}{dt}\|u-u_h\|_0^2 \ .$$

Replacing these two results in (4.21), it comes

$$\begin{aligned} \alpha\|u-u_h\|_1^2 \leq \frac{\alpha}{2}\|u-u_h\|_1^2 + \frac{\gamma^2}{2\alpha}\|u-v_h\|_1^2 + \\ (\frac{\partial(u-u_h)}{\partial t}, u-v_h) - \frac{1}{2}\frac{d}{dt}\|u-u_h\|_0^2 \end{aligned}$$

i.e

$$\frac{1}{2}\frac{d}{dt}\|u-u_h\|_0^2 + \frac{\alpha}{2}\|u-u_h\|_1^2 \leq \frac{\gamma^2}{2\alpha}\|u-v_h\|_1^2 + (\frac{\partial(u-u_h)}{\partial t}, u-v_h) \ .$$

Multiplying by two the two members and integrating in $(0,t)$, it comes

$$\begin{aligned} \|(u-u_h)(t)\|_0^2 + \alpha\int_0^t \|(u-u_h)(s)\|_1^2\, ds \leq \|(u-u_h)(0)\|_0^2 + \frac{\gamma^2}{\alpha}\int_0^t \|u(s)-v_h\|_1^2\ ds + \\ 2\int_0^t (\frac{\partial}{\partial t}(u-u_h)(s), u(s)-v_h)\, ds \ . \end{aligned} \tag{4.22}$$

Integrating by parts the last integral with respect to the temporal variable and using the Young inequality, it comes

$$\begin{aligned} \int_0^t (\frac{\partial}{\partial t}(u-u_h)(s), u(s)-v_h)\ ds = \\ -\int_0^t ((u-u_h)(s), \frac{\partial}{\partial t}(u(s)-v_h))\, ds + \\ ((u-u_h)(t), (u-v_h)(t)) - ((u-u_h)(0), (u-v_h)(0)) \\ \leq \\ \frac{1}{4}\int_0^t \|(u-u_h)(s)\|_0^2\ ds\ + \int_0^t \|\frac{\partial}{\partial t}(u(s)-v_h)\|_0^2\ ds + \\ \frac{1}{4}\|(u-u_h)(t)\|_0^2 + \|(u-v_h)(t)\|_0^2 + \\ \|(u-u_h)(0)\|_0 \cdot \|(u-v_h)(0)\|_0 \ . \end{aligned}$$

Therefore, (4.22) becomes

$$\frac{1}{2}\|(u-u_h)(t)\|_0^2+\alpha\int_0^t\|(u-u_h)(s)\|_1^2\,ds\le$$
$$\frac{\gamma^2}{\alpha}\int_0^t\|u(s)-v_h\|_1^2\,ds+2\int_0^t\|\frac{\partial(u(s)-v_h)}{\partial t}\|_0^2\,ds+$$
$$2\|(u-v_h)(t)\|_0^2+\|(u-u_h)(0)\|_0^2\ +2\|(u-u_h)(0)\|_0\|(u-v_h)(0)\|_0+$$
$$\frac{1}{2}\int_0^t\|(u-u_h)(s)\|_0^2\,ds\ .$$

Choosing for almost $t\in[0,T]$, $v_h=I_h^k(u(t))$, where I_h^k is the finite element interpolation operator (2.15), we find (see [108])

$$h\|u-I_h^k(u)\|_1+\|u-I_h^k(u)\|_0\le Ch^{k+1}|u|_{k+1}\ .$$

Now we have:

$$F_1=\frac{\gamma^2}{\alpha}\int_0^t\|u(s)-v_h\|_1^2\,ds\le C_1h^{2k}\int_0^t|u(s)|_{k+1}^2\,ds$$
$$F_2=2\int_0^t\|\frac{\partial(u-v_h)}{\partial t}(s)\|_0^2\,ds\le C_2h^{2k+2}\int_0^t|\frac{\partial u}{\partial t}(s)|_{k+1}^2\,ds$$
$$F_3=2\|(u-v_h)(t)\|_0^2\le C_3h^{2k+2}|u|_{k+1}^2$$
$$F_4=\|(u-u_h)(0)\|_0^2+2\|(u-u_h)(0)\|_0\|(u-v_h)(0)\|_0\le C_4h^{2k+2}|u(0)|_{k+1}^2\ .$$

Thus $F_1+F_2+F_3+F_4\le Ch^{2k}N(u)$, where $N(u)$ is a suitable function depending on u and on $\frac{\partial u}{\partial t}$. Therefore, we have

$$\frac{1}{2}\|(u-u_h)(t)\|_0^2+\alpha\int_0^t\|(u-u_h)(s)\|_1^2\,ds\le C_{\alpha,\gamma}h^{2k}N(u)+\frac{1}{2}\int_0^t\|(u-u_h)(s)\|_0^2\,ds\ .$$

Applying the Gronwall lemma (see [108]), we obtain the a-priori error estimate :

$$\forall\ t>0\ ,\|(u-u_h)(t)\|_0^2+2\alpha\int_0^t\|u-u_h\|_1^2\le C_{\gamma,\alpha}h^{2k}N(u)e^t \tag{4.23}$$

□

We note that the ingredients in the proof above leads to an a priori error estimate with respect to the norm space $\mathrm{L}^2(\mathbb{R}^+;\mathsf{V})\cap\mathcal{C}^0(\mathbb{R}^+;\mathrm{L}^2(\Omega))$ and for piecewise-linear polynomial only. If we assume more regularity on the data and furthermore that suitable compatibility conditions between the initial datum and the boundary data are satisfied at $t=0$ on $\partial\Omega$, the solution u to (4.5) is indeed more regular. This implies in principle that the convergence of u_h to u is of higher order. An example of error estimate of higher order is given by the following proposition.

Proposition 4.9. *Let $\mathcal{I}_h$ be a regular family of triangulations. Assume that (4.6) holds with $\lambda=0$ and the solution $\varphi(r)$ of the adjoint problem*

$$\varphi(r)\in\mathsf{V}:\quad a(v,\varphi(r))=(r,v)\quad\forall v\in\mathsf{V}$$

satisfies $\varphi(r) \in \mathrm{H}^2(\Omega)$ *when* $r \in \mathrm{L}^2(\Omega)$. *Assume moreover that* $u_0 \in \mathrm{H}^{k+1}(\Omega)$, $k \geq 1$, *and the solution* u *to (4.5) is such that* $\frac{\partial u}{\partial t} \in \mathrm{L}^1(0,T;\mathrm{H}^{k+1}(\Omega))$. *Then, using piecewise-polynomials of degree less that or equal to* k *in the definition of the finite element space* V_h, *for each* $t \in [0,T]$, *the solution* u_h *to (4.17) satisfies*

$$\|u(t) - u_h(t)\|_0 \leq \|u_0 - u_{0,h}\|_0 + Ch^{k+1}(\|u_0\|_{k+1} + \int_0^t \|\frac{\partial u}{\partial t}(\tau)\|_{k+1} d\tau) \tag{4.24}$$

where $C > 0$ *is a suitable constant independent of* h .

We refer to [113] for the proof of the proposition.

4.1.3 Time advancing by FD

In order to have a full discretization of (4.5), we consider a uniform mesh for the time variable t and define

$$t_n := n\Delta t, \quad n = 0, 1, \cdots, \mathcal{N} \tag{4.25}$$

$\Delta t > 0$ being the time-step and $\mathcal{N} := [T/\Delta T]$, the integer part of $T/\Delta T$. Then we replace the time derivative by means of suitable quotients and thus construct a sequence $u_h^n(x)$ that approximates the exact solutions $u(t_n, x)$.
Let us first describe the procedure on a general system of ordinary equations

$$\begin{aligned} \frac{dy(t)}{dt} &= \Psi(t, y(t)), \quad t \in (0,T) \\ y(0) &= y_0 \end{aligned} \tag{4.26}$$

The well known $\theta-$ method consists in replacing (4.26) by the following scheme: find y^n such that

$$\begin{aligned} \frac{1}{\Delta t}(y^{n+1} - y^n) &= \theta\Psi(t_{n+1}, y^{n+1}) + (1-\theta)\Psi(t_n, y^n), \quad n = 0, 1, \cdots, \mathcal{N}-1 \\ y^0 &= y_0 \end{aligned} \tag{4.27}$$

where $\theta \in [0,1]$ is a parameter. When $\theta = 0$ or $\theta = 1$, the method is called forward Euler, or backward Euler respectively while for $\theta = 1/2$ it is named Crank-Nicolson .
Since we are considering only the finite element case, by applying the $\theta-$ scheme to the semi-discrete approximation (4.17), we obtain the problem: find $u_h^n \in \mathsf{V}_h$ such that

$$\begin{aligned} \frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v_h) + a(\theta u_h^{n+1} + (1-\theta)u_h^n, v_h) &= (\theta f(t_{n+1}) + (1-\theta)f(t_n), v_h) \quad \forall\, v_h \in \mathsf{V}_h \\ u_h^0 &= u_{0,h} \end{aligned} \tag{4.28}$$

for each $n = 0, 1, \cdots, \mathcal{N}-1$, having assumed that $f \in \mathrm{L}^2(Q_T)$.
At each time step, one has to solve the linear system

$$(\mathrm{M} + \theta\Delta t\mathrm{A})\underline{\xi}^{n+1} = \underline{\eta}^n \tag{4.29}$$

where $\underline{\eta}^n$ is known from the previous time instant. The matrices M and A are defined in (4.18) and

$$\underline{u}_h^{n+1} = \sum_{j=1}^{N_h} \xi_j^{n+1}\varphi_j$$

φ_j being the base functions of V_h .

Remark 4.10. *If we assume that (4.6) holds with $\lambda = 0$, the matrix $(\mathrm{M} + \theta\Delta t\mathrm{A})$ is positive definite. Thus (4.29) has a unique solution. Moreover, $(\mathrm{M}+\theta\Delta t\mathrm{A})$ is symmetric if the bilinear form is symmetric, i.e., $a(z,v) = a(v,z)$ for each $z, v \in \mathsf{V}$.*

In order to give the stability property of u_h^n, we need to introduce this notation. For any function $\phi \in \mathrm{L}^2(\Omega)$, define

$$\|\phi\|_{-1,h} := \sup_{v_h \in \mathsf{V}_h,\, v_h \neq 0} \frac{(\phi, v_h)}{\|v_h\|_1} \tag{4.30}$$

which is a norm on V_h since $\|\phi\|_{-1,h} \leq \|\phi\|_0$ for each $\phi \in \mathrm{L}^2(\Omega)$. This stability result holds.

Theorem 4.11. *Assume that (4.6) is satisfied with $\lambda = 0$ and that the map $t \to \|f(t)\|_0$ is bounded in $[0,T]$. When $0 \leq \theta \leq 1/2$, assume moreover that $\mathcal{I}_h$ is quasi-uniform family of triangulations and that the following restriction on the time step is satisfied*

$$\Delta t(1 + C_3 h^{-2}) < \frac{2\alpha}{(1-2\theta)\gamma^2} \ , \tag{4.31}$$

where C_3 is the constant appearing in an inverse type inequality (2.49), while α and γ are the coerciveness and continuity constants respectively. Then u_h^n defined in (4.28) satisfies

$$\|u_h^n\|_0 \leq C_\theta(\|u_{0,h}\|_0 + \sup_{t\in[0,T]} \|f(t)\|_0), \ \ n = 0, 1, \cdots, \mathcal{N} \ , \tag{4.32}$$

where $C_\theta > 0$ is a non-decreasing function of α^{-1}, γ and T, and is independent of $\mathcal{N}$, Δt and h.

Proof. Take $v_h = \theta u_h^{n+1} + (1-\theta)u_h^n$ in (4.28). It comes that

$$\frac{1}{2}\|u_h^{n+1}\|_0^2 - \frac{1}{2}\|u_h^n\|_0^2 + (\theta - \frac{1}{2})\|u_h^{n+1} - u_h^n\|_0^2 + \\ \Delta t a(\theta u_h^{n+1} + (1-\theta)u_h^n, \theta u_h^{n+1} + (1-\theta)u_h^n) = \Delta t(\theta f(t_{n+1}) + (1-\theta)f(t_n), \theta u_h^{n+1} + (1-\theta)u_h^n) \, .$$

Using the coerciveness assumption (4.6) for each $0 < \varepsilon \leq 1$ from Young inequality, it comes that

$$\|u_h^{n+1}\|_0^2 - \|u_h^n\|_0^2 + (2\theta - 1)\|u_h^{n+1} - u_h^n\|_0^2 + \\ 2(1-\varepsilon)\alpha\Delta t\|\theta u_h^{n+1} + (1-\theta)u_h^n\|_1^2 \leq \frac{\Delta t}{2\varepsilon\alpha}\|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{-1,h}^2 \ . \tag{4.33}$$

When $1/2 \leq \theta \leq 1$, the left hand side is larger than $\|u_h^{n+1}\|_0^2 - \|u_h^n\|_0^2$ and in particular, we can set $\varepsilon = 1$. When $0 \leq \theta < 1/2$, we proceed as follows: choosing $v_h = u_h^{n+1} - u_h^n$ in (4.28), we find

$$\|u_h^{n+1} - u_h^n\|_0^2 = -\Delta t \, a(\theta u_h^{n+1} + (1-\theta)u_h^n, u_h^{n+1} - u_h^n) + \\ \Delta t \, (\theta f(t_{n+1}) + (1-\theta)f(t_n), u_h^{n+1} - u_h^n) \leq \gamma\Delta t \, \|\theta u_h^{n+1} + (1-\theta)u_h^n\|_1 \|u_h^{n+1} - u_h^n\|_1 + \\ \Delta t \, \|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{-1,h}\|u_h^{n+1} - u_h^n\|_1 \ . \tag{4.34}$$

Then, by means of an inverse type inequality (2.49), we have

$$\|u_h^{n+1} - u_h^n\|_0 \leq \Delta t \, (1 + C_3 h^{-2})^{1/2}[\, \gamma\|\theta u_h^{n+1} + (1-\theta)u_h^n\|_1 + \|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{-1,h} \,] \, . \tag{4.35}$$

Setting for each $\eta > 0$

$$\mathcal{K}_\eta := [\, 2(1-\varepsilon)\alpha - (1-2\theta)\gamma(\gamma+\eta)\Delta t \, (1 + C_3 h^{-2}) \,],$$

it follows

$$\|u^{n+1}\|_0^2 - \|u_h^n\|_0^2 + \Delta t\, \mathcal{K}_\eta \|\theta u_h^{n+1} + (1-\theta)u_h^{n+1}\|_1^2 \le C_{\varepsilon,\eta}\Delta t\,(1 + C_3\, h^{-2})\|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{-1,h}^2 \ .$$

Choosing ε and η small enough due to (4.31), we have $\mathcal{K}_\eta > 0$ and moreover $1 + C_3\, h^{-2} \le C_*$ for a suitable $C_* > 0$, therefore

$$\|u_h^{n+1}\|_0^2 - \|u_h^n\|_0^2 \le C_{\epsilon,\eta}\Delta t \|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{-1,h}^2 \ . \tag{4.36}$$

Now let m to be a fixed index, $1 \le m \le \mathcal{N}$, by summing up from $n = 0$ to $n = m-1$, we find

$$\|u_h^n\|_0^2 \le \|u_{0,h}\|_0^2 + C\Delta t \sum_{n=0}^{m-1} \|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{-1,h}^2$$

and the result follows. □

The next theorem is the error estimate between the semi-discrete solution $u_h(t_n)$ and the fully-discrete one u_h^n for any fixed h.

Theorem 4.12. *Assume that (4.6) is satisfied with $\lambda = 0$ and $\frac{\partial u_h}{\partial t}(0) \in \mathrm{L}^2(\Omega)$, $f \in \mathrm{L}^2(\mathcal{Q}_T)$ with $\frac{\partial f}{\partial t} \in \mathrm{L}^2(\mathcal{Q}_T)$. When $0 \le \theta < 1/2$, assume moreover that $\mathcal{I}_h$ is a quasi-uniform family of triangulations and that the time-step restriction (4.31) is satisfied. Then the functions u_h^n and $u_h(t)$ defined in (4.28) and in (4.17) respectively satisfy*

$$\|u_h^n - u_h(t_n)\|_0 \le C_\theta \Delta t (\|\frac{\partial u_h}{\partial t}(0)\|_0^2 + \int_0^T \|\frac{\partial f}{\partial t}(s)\|_0^2\ )^{1/2} \tag{4.37}$$

for each $n = 0, 1, \cdots, \mathcal{N}$. When $\theta = 1/2$, under the additional assumptions $\frac{\partial^2 f}{\partial t^2} \in \mathrm{L}(\mathcal{Q}_T)$ and $\frac{\partial^2 u_h}{\partial t^2}(0) \in \mathrm{L}^2(\Omega)$, the following estimate also holds

$$\|u_h^n - u_h(t_n)\|_0 \le C(\Delta t)^2 (\|\frac{\partial^2 u_h}{\partial t^2}(0)\|_0^2 + \int_0^T \|\frac{\partial f}{\partial t}(s)\|_0^2)^{1/2} \tag{4.38}$$

for each $n = 0, 1, \cdots, \mathcal{N}$. $C_\theta > 0$ and $C > 0$ are non-decreasing functions of α^{-1}, γ and T, and are independent of $\mathcal{N}$, Δt and h .

We refer to [113] for the proof of this theorem.

4.2 Approximation of convection-dominated problems

In practical applications, convection-diffusion equations are generally employed to describe the transport processes involving fluid motion. With the progress in computer power, the differential convection-diffusion equations can be studied by pursuing the numerical solutions of their discretized counterparts. Therefore, accurate and stable numerical solutions of the convection-diffusion equations are of vital importance. By inspecting the convection-diffusion equations, one can find that it contains two distinct differential operators derived from their respective physical processes: the convection and the diffusion operators. The convection operator consists of variations (first-order spatial derivatives) of the transported variables, which arise from fluid flow. On the other hand, the diffusion operator (represented by second-order spatial derivatives) is due to transport at the molecular level. These two terms can be treated separately and then combined to form the resulting discretized expression. The main problem encountered in general on the treatment of the convective operator. We are going to review standard methods in the approximation of the convection dominated problems.

4.2.1 The streamline diffusion method

The streamline diffusion is the prevalent method in the numerical treatment of stationary convection-dominated problems. The basic idea is due to Brooks and Hughes [16], who called the method the streamline upwind Petrov-Galerkin (SUPG) method.
We describe the idea of the method for the special case of boundary value problem (4.39) under consideration. Let $\Omega \subset \mathbb{R}^2$ denotes a bounded domain with Lipschitz continuous boundary. Given a function $f : \Omega \to \mathbb{R}$, a function $u : \Omega \to \mathbb{R}$ is to be determinated such that

$$\begin{aligned} Lu &= f \quad \text{in} \quad \Omega \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{4.39}$$

where $Lu := -\nabla\cdot(\mathrm{K}\,\nabla u) + \underline{c}\cdot\nabla u + ru$ with sufficiently smooth coefficients

$$\mathrm{K} : \Omega \to \mathbb{R}^{2,2}, \quad \underline{c} : \Omega \to \mathbb{R}^2, \quad r : \Omega \to \mathbb{R}$$

We consider $\mathrm{K}(x) \equiv \varepsilon\mathrm{I}$ with a constant coefficient $\varepsilon > 0$, $\underline{c} \in \mathcal{C}^1(\overline{\Omega}, \mathbb{R}^2)$, $r \in \mathcal{C}(\overline{\Omega})$, $f \in \mathrm{L}^2(\Omega)$. We also assume that the inequality

$$r - \frac{1}{2}\nabla\cdot\underline{c} \geq r_0 \tag{4.40}$$

is valid in Ω, where $r_0 > 0$ is a constant. Then the variational formulation of (4.39) reads as follows:

$$\text{find } u \in \mathsf{V} := \mathrm{H}_0^1(\Omega) \text{ such that } \quad a(u,v) = \int_\Omega fv\,dx \quad \text{for all } v \in \mathsf{V} \tag{4.41}$$

where the bilinear form $a(\cdot,\cdot)$ is :

$$a(u,v) = \int_\Omega [\,\varepsilon\nabla u\cdot\nabla v + \underline{c}\cdot\nabla uv + ruv\,]\,dx \quad u, v \in \mathsf{V}\ . \tag{4.42}$$

Given a regular family of triangulations $\mathcal{I}_h$, let $\mathsf{V}_h \subset \mathsf{V}$ denote the set of continuous functions that are piecewise polynomial of degree $k \in \mathbb{N}$ and satisfy the boundary conditions, i.e.,

$$\mathsf{V}_h := \{\ v_h \in \mathsf{V} \mid v_h|_K \in \mathcal{P}_k(K) \quad \text{for all} \quad K \in \mathcal{I}_h\ \}\ . \tag{4.43}$$

If in addition the solution $u \in \mathsf{V}$ of (4.41) belongs to the space $\mathrm{H}^{k+1}(\Omega)$, we have by (2.16), the following estimate for the interpolant $I_h(u)$:

$$\|u - I_h(u)\|_{l,K} \leq C_{int} h_K^{k+1-l} |u|_{k+1,K} \tag{4.44}$$

for all $0 \leq l \leq k+1$ and all $K \in \mathcal{I}_h$. Since the spaces V_h are of finite dimension, a so-called inverse inequality similar to (2.49) also holds:

$$\|\Delta v_h\|_{0,K} \leq \frac{C_{inv}}{h_K}|v_h|_{1,K} \tag{4.45}$$

for all $v_h \in \mathsf{V}_h$ and all $K \in \mathcal{I}_h$. We note that it is important that the constants $C_{int},\ C_{inv} > 0$ from (4.44) and (4.45) respectively, do not depend on u or v_h and on the particular element $K \in \mathcal{I}_h$.
The basic idea of the streamline-diffusion method consists in the addition of suitably weighted residuals to the variational formulation (4.42). Because of the assumption $u \in \mathrm{H}^{k+1}(\Omega)$, $k \in \mathbb{N}$, the differential equation can be interpreted as an equation in $\mathrm{L}^2(\Omega)$. In particular, it is valid on any element $K \in \mathcal{I}_h$ in the sense of distribution strong form i.e.,

$$-\varepsilon\Delta u + \underline{c}\cdot\nabla u + ru = f \quad \text{almost everywhere in } K \text{ and for all } K \in \mathcal{I}_h\ .$$

Taking an element wise defined mapping $\tau : \mathsf{V}_h \to \mathrm{L}^2(\Omega)$ and multiplying the local differential equation in $\mathrm{L}^2(K)$ by the restriction over $\tau(v_h)$ to K, then scaling by a suitable parameter $\delta_K \in \mathbb{R}$ and summing the results over all elements $K \in \mathcal{I}_h$, we obtain

$$\sum_{K\in\mathcal{I}_h} \delta_K \langle -\varepsilon\Delta u + \underline{c}\cdot\nabla u + ru, \tau(v_h)\rangle_{0,K} = \sum_{K\in\mathcal{I}_h} \delta_K \langle f, \tau(v_h)\rangle_{0,K}$$

Adding this relation to the equation (4.41) restricted to V_h, we see that the weak formulation $u \in \mathsf{V}_h \cap \mathrm{H}^{k+1}(\Omega)$, satisfies the following variational equation:

$$a_h(u, v_h) = \langle f, v_h\rangle_h \quad \text{for all} \quad v_h \in \mathsf{V}_h$$

where

$$a_h(u, v_h) := a(u, v_h) + \sum_{K\in\mathcal{I}_h} \delta_K \langle -\varepsilon\Delta u + \underline{c}\cdot\nabla u + ru, \tau(v_h)\rangle_{0,K} \ ,$$
$$\langle f, v_h\rangle_h = \langle f, v\rangle_0 + \sum_{K\in\mathcal{I}_h} \delta_K \langle f, \tau(v_h)\rangle_{0,K} \ .$$

Thus the discrete problem is the following :

$$\text{find} \quad u_h \in \mathsf{V}_h \quad \text{such that} \quad a_h(u_h, v_h) = \langle f, v_h\rangle_h \quad \text{for all} \quad v_h \in \mathsf{V}_h \tag{4.46}$$

Corollary 4.13. *Suppose the problems (4.41) and (4.46) have a solution $u \in \mathsf{V} \cap \mathrm{H}^{k+1}(\Omega)$ and $u_h \in \mathsf{V}_h$ respectively. The following error equation holds:*

$$a_h(u - u_h, v_h) = 0 \quad \textit{for all} \quad v_h \in \mathsf{V}_h \ . \tag{4.47}$$

Remark 4.14.

- *In the streamline-diffusion method (SDFEM), the mapping τ used in (4.46) is chosen as $\tau(v_h) := \underline{c}\cdot\nabla v_h$*
- *Another choice of the mapping τ is given by $\tau(v_h) := -\varepsilon\Delta v_h + \underline{c}\cdot\nabla v_h + rv_h$. This is the so-called Galerkin/ Least Squares-FEM (GLS FEM) method. A detail of this approach can be found in [62].*

4.2.2 Interpretation of the additional term in the case of linear elements

If the finite element space V_h is formed by piecewise linear functions (i.e., in the above definition (4.43) of V_h we have $k = 1$), we get $\Delta v_h|_{\mathsf{K}} = 0$ for all $K \in \mathcal{I}_h$. If in addition there is no reactive term (i.e., $r = 0$), the discrete bilinear form is

$$a_h(u_h, v_h) = \int_\Omega \varepsilon\nabla u_h \cdot \nabla v_h \ dx + \langle \underline{c}\cdot\nabla u_h, v_h\rangle_0 + \sum_{K\in\mathcal{I}_h} \delta_K \langle \underline{c}\cdot\nabla u_h, \underline{c}\cdot\nabla v_h\rangle_{0,K} \ .$$

Since the scalar product appearing in the sum can be rewritten as:
$\langle \underline{c}\cdot\nabla u_h, \underline{c}\cdot\nabla v_h\rangle_{0,K} = \int_K (\underline{c}\underline{c}^T\nabla u_h)\cdot\nabla v_h \ dx$, we obtain the following equivalent representation:

$$a_h(u_h, v_h) = \sum_{K\in\mathcal{I}_h} \int_K ((\varepsilon\mathrm{I} + \delta_K \underline{c}\underline{c}^T)\nabla u_h)\cdot\nabla v_h \ dx + \langle \underline{c}\cdot\nabla u_h, v_h\rangle_0 \, .$$

This shows that the additional term introduces an element dependent extra diffusion in the direction of the convective field $\underline{c}$ which motivates the name of the method.

4.2.3 Analysis of the streamline diffusion method

To start the analysis of stability and convergence properties of the streamline diffusion method, we consider the term $a_h(v_h, v_h)$ for arbitrary $v_h \in \mathsf{V}_h$ and the structure of the discrete bilinear form $a_h(\cdot\,,\cdot)$ that allows us to derive the estimate

$$a_h(v_h, v_h) \geq \varepsilon |v_h|_1^2 + r_0 \|v_h\|_0^2 + \sum_{K \in \mathcal{I}_h} \delta_K \langle -\varepsilon \Delta v_h + \underline{c} \cdot \nabla v_h + r v_h, \underline{c} \cdot \nabla v_h \rangle_{0,K} \ . \tag{4.48}$$

Theorem 4.15. *We have the following stability property:*

$$a_h(v_h, v_h) \geq \frac{1}{2} \|v_h\|_{sd}^2 \quad \forall \ v_h \in \mathsf{V}_h \ ,$$

where $\|.\|_{sd}$ is a suitable streamline diffusion norm.

Proof. To get an ellipticity estimate of $a(\cdot,\cdot)$ in (4.42), we take $u = v$, use the relation $2v(\underline{c} \cdot \nabla v) = \underline{c} \cdot \nabla v^2$, and performing an integration by parts of the middle term, we obtain

$$\begin{aligned} a(v,v) &= \varepsilon |v|_1^2 + \langle \underline{c} \cdot \nabla v, v \rangle_0 + \langle r v, v \rangle_0 \\ &= \varepsilon |v|_1^2 - \langle \frac{1}{2} \nabla \cdot \underline{c}, v^2 \rangle_0 + \langle r v, v \rangle_0 \\ &= \varepsilon |v|_1^2 + \langle r - \frac{1}{2} \nabla \cdot \underline{c}, v^2 \rangle_0 \end{aligned}$$

and using the inequality (4.40), we obtain the estimate

$$a(v,v) \geq \varepsilon |v|_1^2 + r_0 \|v\|_0^2 \tag{4.49}$$

and thus the estimate (4.48).
Neglecting for a moment the term $\underline{c} \cdot \nabla v_h$ in the sum (4.48) and using the elementary inequality $ab \leq a^2 + b^2/4$ for arbitrary $a, b \in \mathbb{R}$, we get

$$\begin{aligned} |\sum_{K \in \mathcal{I}_h} \delta_K \langle -\varepsilon \Delta v_h + r v_h, \underline{c} \cdot \nabla v_h \rangle_{0,K}| &\leq \sum_{K \in \mathcal{I}_h} \{ |\langle -\varepsilon \sqrt{|\delta_K|} \Delta v_h, \sqrt{|\delta_K|} \underline{c} \cdot \nabla v_h \rangle_{0,K}| + \\ |\langle \sqrt{|\delta_K|} r v_h, \sqrt{|\delta_K|} \underline{c} \cdot \nabla v_h \rangle_{0,K}| \} & \\ &\leq \sum_{K \in \mathcal{I}_h} \{ \varepsilon^2 |\delta_K| \|\Delta v_h\|_{0,K}^2 + |\delta_K| \|r\|_{\infty,K}^2 \|v_h\|_{0,K}^2 + \frac{|\delta_K|}{2} \|\underline{c} \cdot \nabla v_h\|_{0,K}^2 \}, \end{aligned}$$

where $\|r\|_{\infty,K} = \sup_K r$.
By means of the inverse inequality (4.45), it follows that

$$\begin{aligned} |\sum_{K \in \mathcal{I}_h} \delta_K \langle -\varepsilon \Delta v_h + r v_h, \underline{c} \cdot \nabla v_h \rangle_{0,K}| &\leq \sum_{K \in \mathcal{I}_h} \{ \varepsilon^2 |\delta_K| \frac{C_{inv}^2}{h_K^2} |v_h|_{1,K}^2 \ + \\ |\delta_K| \|r\|_{\infty,K}^2 \|v_h\|_{0,K}^2 + \frac{|\delta_K|}{2} \|\underline{c} \cdot \nabla v_h\|_{0,K}^2 \} \ . & \end{aligned}$$

Putting things together, we obtain

$$\begin{aligned} a_h(v_h, v_h) \geq \sum_{K \in \mathcal{I}_h} \{ (\varepsilon - \varepsilon^2 |\delta_K| \frac{C_{inv}^2}{h_K^2}) |v_h|_{1,K}^2 \ + \\ (r_0 - |\delta_K| \|r\|_{\infty,K}^2) \|v_h\|_{0,K}^2 + (\delta_K - \frac{|\delta_K|}{2}) \|\underline{c} \cdot \nabla v_h\|_{0,K}^2 \} & \end{aligned}$$

Choosing

$$0 < \delta_K \leq \frac{1}{2} \min \{ \frac{h_K^2}{\varepsilon C_{inv}^2}, \frac{r_0}{\|r\|_{\infty,K}^2} \} \ , \tag{4.50}$$

we get

$$a_h(v_h, v_h) \geq \frac{\varepsilon}{2}|v_h|_1^2 + \frac{r_0}{2}\|v_h\|_0^2 + \frac{1}{2}\sum_{K\in\mathcal{I}_h}\delta_K\|\underline{c}\cdot\nabla v_h\|_{0,K}^2$$

If we define the so-called streamline-diffusion norm

$$\|v\|_{sd} := \{\varepsilon|v|_1^2 + r_0\|v\|_0^2 + \sum_{K\in\mathcal{I}_h}\delta_K\|\underline{c}\cdot\nabla v\|_{0,K}^2\}^{\frac{1}{2}},\ v\in\mathsf{V},$$

we obtain from the choice (4.50) that

$$\frac{1}{2}\|v_h\|_{sd}^2 \leq a_h(v_h, v_h) \quad \text{for all} \quad v_h \in \mathsf{V}_h \tag{4.51}$$

i.e., the stability property is demonstrated. □

In the following, we develop some work useful for obtaining an error estimate and convergence demonstration.
Since estimate (4.51) holds only on the finite element spaces V_h, let us consider the norm of $I_h(u) - u_h \in \mathsf{V}_h$ and make use of the error equation (4.47):

$$\frac{1}{2}\|I_h(u) - u_h\|_{sd}^2 \leq a_h(I_h(u) - u_h, I_h(u) - u_h) = a_h(I_h(u) - u, I_h(u) - u_h)\ .$$

In particular, under the assumption $u \in \mathsf{V}\cap \mathrm{H}^{k+1}(\Omega)$, the following three estimates holds:

- $$\varepsilon\int_\Omega \nabla(I_h(u)-u)\cdot\nabla(I_h(u)-u_h)\,dx \leq \sqrt{\varepsilon}|I_h(u)-u|_1\|I_h(u)-u_h\|_{sd}$$
$$\leq C_{int}\sqrt{\varepsilon}h^k|u|_{k+1}\|I_h(u)-u_h\|_{sd}\ ,$$

- $$\int_\Omega [\underline{c}\cdot\nabla(I_h(u)-u) + r(I_h(u)-u)](I_h(u)-u_h)\,dx = \int_\Omega (r-\nabla\cdot\underline{c})(I_h(u)-u)(I_h(u)-u_h)\,dx\ -$$
$$\int_\Omega (I_h(u)-u)\underline{c}\cdot\nabla(I_h(u)-u_h)\,dx$$
$$\leq \|r-\nabla\cdot\underline{c}\|_\infty\|I_h(u)-u\|_0\|I_h(u)-u_h\|_0\ +$$
$$\|I_h(u)-u\|_0\|\underline{c}\cdot\nabla(I_h(u)-u_h)\|_0$$
$$\leq C\{\{\sum_{K\in\mathcal{I}_h}\|I_h(u)-u\|_{0,K}^2\}^{\frac{1}{2}}\ +$$
$$\{\sum_{K\in\mathcal{I}_h}\delta_K^{-1}\|I_h(u)-u\|_{0,K}^2\}^{\frac{1}{2}}\}\|I_h(u)-u_h\|_{sd}$$
$$\leq Ch^k\{\sum_{K\in\mathcal{I}_h}(1+\delta_K^{-1})h_K^2|u|_{k+1,K}^2\}^{\frac{1}{2}}\|I_h(u)-u_h\|_{sd}\ ,$$

  and

- $$|\sum_{K\in\mathcal{I}_h}\delta_K\langle -\varepsilon\Delta(I_h(u)-u) + \underline{c}\cdot\nabla(I_h(u)-u) + r(I_h(u)-u), \underline{c}\cdot\nabla(I_h(u)-u_h)\rangle_{0,K}| \leq$$
$$\sum_{K\in\mathcal{I}_h} C_{int}\sqrt{\delta_K}\ [\ \varepsilon h_K^{k-1} + \|\underline{c}\|_{\infty,K}h_K^k + \|r\|_{\infty,K}h_K^{k+1}\ ]\times$$
$$|u|_{k+1,K}\sqrt{\delta_K}\|\underline{c}\cdot\nabla(I_h(u)-u_h)\|_{0,K}$$
$$\leq C\{\sum_{K\in\mathcal{I}_h}\delta_K\ [\ \varepsilon h_K^{k-1} + h_K^k + h_K^{k+1}\ ]^2|u|_{k+1,K}^2\}^{\frac{1}{2}}\|I_h(u)-u_h\|_{sd}\ .$$

Condition (4.50), which was already required for estimate (4.51), implies that

$$\varepsilon\delta_K \leq \frac{h_K^2}{C_{inv}^2}$$

and so the application to the first term of the last bound leads to

$$|\sum_{K\in\mathcal{I}_h} \delta_K\langle -\varepsilon\Delta(I_h(u)-u) + \underline{c}\cdot\nabla(I_h(u)-u) + r(I_h(u)-u), \underline{c}\cdot\nabla(I_h(u)-u_h)\rangle_{0,K}| \leq$$
$$Ch^k\{\sum_{K\in\mathcal{I}_h}[\ \varepsilon+\delta_K\ ]|u|_{k+1,K}^2\}^{\frac{1}{2}}\|I_h(u)-u_h\|_{sd}$$

Collecting the estimates and dividing by $\|I_h(u)-u_h\|_{sd}$, we obtain the relation

$$\|I_h(u)-u_h\|_{sd} \leq Ch^k\{\sum_{K\in\mathcal{I}_h}[\ \varepsilon+\frac{h_K^2}{\delta_K}+h_K^2+\delta_K\ ]|u|_{k+1,K}^2\}^{\frac{1}{2}}$$

Finally, the terms in the square brackets will be equilibrated with the help of condition (4.50). We rewrite the $\varepsilon-$dependent term condition as

$$\frac{h_K^2}{\varepsilon C_{inv}^2} = \frac{2}{C_{inv}^2\|\underline{c}\|_{\infty,K}}P_{e_K}h_K$$

with

$$P_{e_K} := \frac{\|\underline{c}\|_{\infty,K}\ h_K}{2\varepsilon}$$

the local Péclet number, a refinement of the global Péclet number given by

$$P_e := \frac{\|\underline{c}\|_\infty\ diam(\Omega)}{\varepsilon} \tag{4.52}$$

The following distinctions concerning P_{e_K} are convenient:

$$P_{e_K} \leq 1 \quad \text{and} \quad P_{e_K} > 1 \quad .$$

In the first case, we choose

$$\delta_K = \delta_0 P_{e_K}h_K = \delta_1\frac{h_K^2}{\varepsilon}, \quad \delta_0 = \frac{2}{\|\underline{c}\|_{\infty,K}}\delta_1\,,$$

with appropriate constants $\delta_0 > 0$ and $\delta_1 > 0$, respectively which are independent of K and ε. Then we have

$$\varepsilon+\frac{h_K^2}{\delta_K}+h_K^2+\delta_K = (1+\frac{1}{\delta_1})\varepsilon+h_K^2+\delta_1\frac{2P_{e_K}}{\|\underline{c}\|_{\infty,K}}h_K \leq C(\varepsilon+h_K)\ ,$$

where $C>0$ is independent of K and ε.
In the second case, it is sufficient to choose $\delta_K = \delta_2 h_K$ with an appropriate constant $\delta_2 > 0$ that is independent of K and ε. Then

$$\delta_K = \frac{\delta_2}{P_{e_K}}P_{e_K}h_K = \frac{\delta_2\|\underline{c}\|_{\infty,K}}{2P_{e_K}}\frac{h_K^2}{\varepsilon}$$

and

$$\varepsilon+\frac{h_K^2}{\delta_K}+h_K^2+\delta_K = \varepsilon+(\frac{1}{\delta_2}+\delta_2)h_K+h_K^2 \leq C(\varepsilon+h_K)$$

with $C>0$ independent of K and ε.
We note that in both cases the constants can be chosen sufficiently small, independent of P_{e_K},

that is the condition (4.50) is satisfied.
Finally, we obtain

$$\begin{aligned}\|I_h(u)-u_h\|_{sd} &\le Ch^k\{\sum_{K\in\mathcal{I}_h}(\varepsilon+h_K)|u|^2_{h+1,K}\}^{\frac{1}{2}}\\ &\le Ch^k(\varepsilon+h)^{\frac{1}{2}}|u|_{k+1}\ .\end{aligned}$$

Now, we are ready to formulate a theorem relevant to the convergence property and error estimate.

Theorem 4.16. *Let the parameters δ_K be given by*

$$\delta_K=\begin{cases}\delta_1\dfrac{h_K^2}{\varepsilon}, & P_{e_K}\le 1,\\ \delta_2 h_K, & P_{e_K}>1,\end{cases}$$

where $\delta_1,\delta_2>0$ do not depend on K and ε and are chosen such that condition (4.50) is satisfied. If the weak solution u of (4.41) belongs to $\mathrm{H}^{k+1}(\Omega)$, then the solution u_h of (4.46) converges and satisfies the relation

$$\|u-u_h\|_{sd}\le C(\sqrt{\varepsilon}+\sqrt{h})h^k|u|_{k+1}\ ,$$

where $C>0$ is a constant independent of ε, h and u.

Proof. By the triangle inequality, we get

$$\|u-u_h\|_{sd}\le\|u-I_h(u)\|_{sd}+\|I_h(u)-u_h\|_{sd}\ .$$

We have already an estimate of the second addend. To deal with the first term, the estimates of the interpolation error (4.44) are used directly

$$\begin{aligned}\|u-I_h(u)\|^2_{sd} &\le \varepsilon|u-I_h(u)|^2_1+r_0\|u-I_h(u)\|^2_0+\sum_{K\in\mathcal{I}_h}\delta_K\|\underline{c}\cdot\nabla(u-I_h(u))\|^2_{0,\mathsf{K}}\\ &\le C^2_{int}\sum_{K\in\mathcal{I}_h}[\ \varepsilon h_K^{2k}+r_0h_K^{2(k+1)}+\delta_K\|\underline{c}\|^2_{\infty,K}h_K^{2k}\ ]|u|^2_{k+1,K}\\ &\le Ch_K^{2k}\sum_{K\in\mathcal{I}_h}[\ \varepsilon+h_K^2+\delta_K\ ]|u|^2_{k+1,K}\\ &\le C(\varepsilon+h)h_K^{2k}|u|^2_{k+1}\ .\end{aligned}$$

and $\|u-I_h(u)\|_{sd}\le C(\varepsilon+h)^{\frac{1}{2}}h^k|u|_{k+1}$. Summing the terms relevant to the two addends, we obtain the result. □

Remark 4.17. *In the case of Péclet numbers $P_{e_K}>1$, we have $\varepsilon<\frac{1}{2}\|\underline{c}\|_{\infty,K}h_K$ and thus*

$$\|u-u_h\|_0+\{\delta_2\sum_{K\in\mathcal{I}_h}h_K\|\underline{c}\cdot\nabla(u-u_h)\|^2_{0,K}\}^{\frac{1}{2}}\le Ch^{k+1/2}|u|_{k+1}\ .$$

So the L^2- error norm of the solution is not optimal in comparison with the estimate of the interpolation error

$$\|u-I_h(u)\|_0\le C\,h^{k+1}|u|_{k+1}\ ,$$

while the L^2- error of the directional derivative of u in the direction of the velocity $\underline{c}$ is optimal.

4.2.4 The characteristic Galerkin method

We assume that Ω is bounded in $\mathbb{R}^2$, with Lipschitz boundary $\partial\Omega$, and consider the parabolic initial-boundary value problem: for each $t \in [0,T]$ find $u(t)$ such that

$$\begin{aligned} \frac{\partial u}{\partial t} + Lu &= f \quad \text{in} \quad \mathcal{Q}_T := (0,T)\times\Omega \\ u &= 0 \quad \text{on} \quad \Sigma_T := (0,T)\times\partial\Omega \\ u &= u_0 \quad \text{in} \quad \Omega \text{ for } t = 0 \ , \end{aligned} \tag{4.53}$$

where L is the second-order elliptic operator

$$Lw := -\varepsilon\Delta w + \sum_{i=1}^{2} D_i(c_i w) + a_0 w \tag{4.54}$$

With no loss of generality, we consider the case in which $\varepsilon \lll \|\underline{c}\|_{\mathrm{L}^\infty(\Omega)}$. We also assume that there exists two positive constant μ_0 and μ_1 such that

$$0 < \mu_0 \le \mu(x) := \frac{1}{2} div\, \underline{c}(x) + a_0(x) \le \mu_1 \tag{4.55}$$

for almost every $x \in \Omega$.
The method of characteristic stems from considering the non-stationary advection-diffusion equation (4.53) from a Lagrangian instead of Eulerian point of view, and can be traced back to Pironneau [103], Douglas and Russell [40], Ewing, Russell and Wheeler [45]. We first define the characteristic lines associated to a vector field $\underline{c} = \underline{c}(t,x)$. Being given $x \in \overline{\Omega}$ and $s \in [0,T]$, they are vector functions $\mathrm{X} = \mathrm{X}(t;s,x)$ such that

$$\begin{cases} \dfrac{d\mathrm{X}}{dt}(t;s,x) = \underline{c}(t,\mathrm{X}(t;s,x)) \ , \quad t \in (0,T) \\ \mathrm{X}(s;s,x) = x \ . \end{cases} \tag{4.56}$$

The existence and uniqueness of the characteristic lines for each choice of s and x hold under mild assumptions on $\underline{c}$, for instance $\underline{c}$ continuous in $[0,T]\times\overline{\Omega}$ and Lipschitz continuous in $\overline{\Omega}$, uniformly with respect to $t \in [0,T]$, see [59].
From a geometric point of view, $\mathrm{X}(t;s,x)$ provides the position at time t of a particle which has been driven by the field $\underline{c}$ and that occupied the position x at the time s. The uniqueness result gives in particular that

$$\mathrm{X}(t;s,\mathrm{X}(s;\tau,x)) = \mathrm{X}(t;\tau,x) \tag{4.57}$$

for each $t,s,\tau \in [0,T]$ and $x \in \overline{\Omega}$. Hence $\mathrm{X}(t;s,\mathrm{X}(s;t,x)) = \mathrm{X}(t;t,x) = x$, i.e., for fixed t and s, the inverse function of $x \to \mathrm{X}(s;t,x)$ is given by $y \to \mathrm{X}(t;s,y)$.
Therefore defining

$$\overline{u}(t,y) := u(t,\mathrm{X}(t;0,y)) \tag{4.58}$$

or equivalently, $u(t,x) = \overline{u}(t,\mathrm{X}(0;t,x))$. From (4.56) it follows that

$$\begin{aligned} \frac{\partial \overline{u}}{\partial t}(t,y) &= \frac{\partial u}{\partial t}(t,\mathrm{X}(t;0,y)) + \sum_{i=1}^{2} D_i u(t,\mathrm{X}(t;0,y)) \frac{d\mathrm{X}_i}{dt}(t;0,y) \\ &= (\frac{\partial u}{\partial t} + \underline{c}\cdot\nabla u)(t,\mathrm{X}(t;0,y)) \ . \end{aligned} \tag{4.59}$$

According to the notation introduced in (4.58), we can rewrite the non-stationary advection-diffusion equation as

$$\frac{\partial \overline{u}}{\partial t} - \varepsilon\overline{\Delta u} + (\overline{div\, \underline{c}} + \overline{a_0})\overline{u} = \overline{f} \quad \text{in} \quad \mathcal{Q}_T \ . \tag{4.60}$$

We are thus ready to discretize (4.60). The time derivative could be approximated by the backward scheme, i.e.,

$$\frac{\partial \overline{u}}{\partial t}(t_{n+1}, y) \cong \frac{\overline{u}(t_{n+1}, y) - \overline{u}(t_n, y)}{\Delta t} \ . \tag{4.61}$$

If we set $y = \mathrm{X}(0; t_{n+1}, x)$, from (4.58) we obtained

$$\frac{\partial \overline{u}}{\partial t}(t_{n+1}, \mathrm{X}(0; t_{n+1}, x)) \cong \frac{u(t_{n+1}, x) - u(t_n, \mathrm{X}(t_n; t_{n+1}, x))}{\Delta t} \ .$$

Denoting by $\mathrm{X}^n(x)$ a suitable approximation of $\mathrm{X}(t_n; t_{n+1}, x), n = 0, 1, \cdots, \mathcal{N}-1$ and by $u^n \circ \mathrm{X}^n$ an approximation of u at the point X^n, we can finally obtain the following implicit discretization scheme for the problem (4.60): set $u^0 := u_0$, then for $n = 0, 1, \cdots, \mathcal{N} - 1$ solve

$$\frac{u^{n+1} - u^n \circ \mathrm{X}^n}{\Delta t} - \varepsilon \Delta u^{n+1} + [div\, \underline{c}(t_{n+1}) + a_0] u^{n+1} = f(t_{n+1}) \quad \text{in} \quad \Omega \ . \tag{4.62}$$

This formulation is completed by the boundary condition $u^{n+1} = 0$ on $\partial\Omega$.
One can typically choose a backward Euler scheme also for discretizing

$$\frac{d\mathrm{X}}{dt}(t; t_{n+1}, x) = \underline{c}(t, \mathrm{X}(t; t_{n+1}, x)) \ . \tag{4.63}$$

This produces the following approximation of $\mathrm{X}(t_n; t_{n+1}, x)$:

$$\mathrm{X}^n_{(1)}(x) := x - \underline{c}(t_{n+1}, x)\Delta t \ . \tag{4.64}$$

We notice that $\mathrm{X}^n_{(1)}$ is a second order approximation of $\mathrm{X}(t_n; t_{n+1}, x)$ since we are integrating (4.63) on the time interval (t_n, t_{n+1}) which has length Δt .
A more accurate scheme is provided by the second order Runge-Kutta scheme

$$\mathrm{X}^n_{(2)}(x) := x - \underline{c}(t_{n+\frac{1}{2}}, x - \underline{c}(t_{n+1}, x)\frac{\Delta t}{2})\Delta t \ , \tag{4.65}$$

which gives a third order approximation of $\mathrm{X}(t_n; t_{n+1}, x)$.

Remark 4.18. *It is important to verify that $\mathrm{X}^n_{(i)}(x) \in \Omega$ for each $x \in \overline{\Omega}, i = 1, 2$, so we can compute $u^n \circ \mathrm{X}^n_{(i)}$.*

If we assume that $\underline{c}(t, x) \neq 0$ for each $t \in [0, T]$ and $x \in \partial\Omega$, therefore, $\mathrm{X}^n_{(i)}(x) = x$ for $x \in \partial\Omega, i = 1, 2$. If we denote by $x^* \in \partial\Omega$ the point having minimal distance from $x \in \Omega$, we have

$$\begin{aligned} |\mathrm{X}^n_{(1)}(x) - x| &= |\underline{c}(t_{n+1}, x)|\Delta t = |\underline{c}(t_{n+1}, x) - \underline{c}(t_{n+1}, x^*)|\Delta t \\ &\leq |\underline{c}(t_{n+1})|_{Lip(\overline{\Omega})} |x - x^*| \ , \end{aligned}$$

where

$$|g|_{Lip(\overline{\Omega})} := \sup_{x_1, x_2 \in \overline{\Omega}, x_1 \neq x_2} \frac{|g(x_1) - g(x_2)|}{|x_1 - x_2|} \ .$$

By assuming that

$$\max_{t \in [0,T]} |\underline{c}(t)|_{Lip(\overline{\Omega})} \Delta t < 1 \ , \tag{4.66}$$

it follows at once that $\mathrm{X}^n_{(1)} \in \Omega$ for each $x \in \Omega$ and for each $n = 0, 1, \cdots, \mathcal{N} - 1$.
A similar result can be obtained for $\mathrm{X}^n_{(2)}(x)$.
Let now consider the second order approximation (4.64) referring to Pironneau [104] for higher order scheme based on (4.65).

Theorem 4.19. *If we suppose that*

$$div\, \underline{c}(t, x) + a_0(x) > 0 \tag{4.67}$$

then solution of (4.53) depends on the problem data for each $t \in [0, T]$ and almost every $x \in \Omega$.

Proof. In fact, by multiplying (4.62) by u^{n+1} and integrating over Ω, one obtains

$$\|u^{n+1}\|_0^2 + \varepsilon\Delta t\|\nabla u^{n+1}\|_0^2 \leq (\|u^n \circ \mathrm{X}_{(1)}^n\|_0 + \Delta t\|f(t_{n+1})\|_0)\|u^{n+1}\|_0 \ . \tag{4.68}$$

From (4.66) it also follows that the map $\mathrm{X}_{(1)}^n$ is injective. Therefore, we can introduce the change of variable $y = \mathrm{X}_{(1)}^n(x)$, and setting $\mathrm{Y}_{(1)}^n(y) := (\mathrm{X}_{(1)}^n)^{-1}(y)$ we have

$$\|u^n \circ \mathrm{X}_{(1)}^n\|_0^2 = \int\limits_{\mathrm{X}_{(1)}^n(\Omega)} [u^n(y)]^2 \circ \mathrm{Y}_{(1)}^n(y)|det(Jac\ \mathrm{X}_{(1)}^n)^{-1}|\ dy \ . \tag{4.69}$$

On the other hand,

$$|det(Jac\,\mathrm{X}_{(1)}^n)(x)| \geq 1 - \Delta t\, C_1\, \|Jac\ \underline{c}(t_{n+1})\|_{\mathrm{L}^\infty(\Omega)} > 0$$

for almost every $x \in \Omega$, provided that

$$\mu_1^*\Delta t \leq C_2 \ , \tag{4.70}$$

where

$$\mu_1^* := \max_{t\in[0,T]} \|Jac\ \underline{c}(t)\|_{\mathrm{L}^\infty(\Omega)}$$

and $C_1 > 0$, $0 < C_2 < C_1^{-1}$ are suitable constants. Therefore, choosing the smallest constant C_2 in (4.70), from (4.69) one has

$$\|u^n \circ \mathrm{X}_{(1)}^n\|_0^2 \leq (1 + \Delta t C_3\mu_1^*)\|u^n\|_0^2 \ . \tag{4.71}$$

We note that condition (4.70) implies (4.66) if C_2 is small enough. From (4.68) and (4.71) we have

$$\begin{aligned}\|u^{n+1}\|_0^2 + \varepsilon\Delta t\|\nabla u^{n+1}\|_0^2 &\leq (\|u^n \circ \mathrm{X}_{(1)^n}\|_0 + \Delta t\|f(t_{n+1})\|_0)\|u^{n+1}\|_0 \\ &\leq \frac{1}{2}\{[\|u^n \circ \mathrm{X}_{(1)}^n\|_0 + \Delta t\|f(t_{n+1})\|_0]^2 + \|u^{n+1}\|_0^2\}\end{aligned}$$

by Young

$$2\|u^{n+1}\|_0^2 + 2\varepsilon\Delta t\|\nabla u^{n+1}\|_0^2 \leq [\|u^n \circ \mathrm{X}_{(1)}^n\|_0 + \Delta t\|f(t_{n+1})\|_0]^2 + \|u^{n+1}\|_0^2$$

and

$$\|u^{n+1}\|_0^2 + 2\varepsilon\Delta t\|\nabla u^{n+1}\|_0^2 \leq [\|u^n \circ \mathrm{X}_{(1)}^n\|_0 + \Delta t\|f(t_{n+1})\|_0]^2$$

and so

$$\{\|u^{n+1}\|_0^2 + 2\varepsilon\Delta t\|\nabla u^{n+1}\|_0^2\}^{\frac{1}{2}} \leq \|u^n \circ \mathrm{X}_{(1)}^n\|_0 + \Delta t\|f(t_{n+1})\|_0 \ ,$$

therefore we finally obtain for each $n = 0, 1, \cdots, \mathcal{N} - 1$

$$\begin{aligned}(\|u^{n+1}\|_0^2 + 2\varepsilon\Delta t\|\nabla u^{n+1}\|_0^2)^{\frac{1}{2}} &\leq (1 + C_3\mu_1^*\Delta t)^{\frac{1}{2}}\|u^n\|_0 + \Delta t\|f(t_{n+1})\|_0 \\ &\leq (1 + C_3\mu_1^*\Delta t)^{\frac{n+1}{2}}\|u_0\|_0 + \Delta t\sum_{k=1}^{n+1}(1 + C_3\mu_1^*\Delta t)^{\frac{n+1-k}{2}}\|f(t_k)\|_0 \\ &\leq (\|u_0\|_0 + t_{n+1}\max_{t\in[0,T]}\|f(t)\|_0)\exp(\frac{C_3}{2}\mu_1^* t_{n+1}) \ .\end{aligned} \tag{4.72}$$

□

The method we have described above also applies to the fully-discretized problem obtained from (4.62) by using for example, the finite element method. The resulting scheme would read: given $u_h^0 := u_{0,h} \in \mathsf{V}_h$, for each $n = 0, 1, \cdots, \mathcal{N}-1$ find $u_h^{n+1} \in \mathsf{V}_h$ such that

$$\begin{aligned} &\frac{1}{\Delta t}(u_h^{n+1} - u_h^n \circ \mathrm{X}^n, v_n) + \\ &\varepsilon(\nabla u_h^{n+1}, \nabla v_h) + ([div\, \underline{c}(t_{n+1}) + a_0]u_h^{n+1}, v_h) = (f(t_{n+1}), v_h) \quad \forall\ v_h \in \mathsf{V}_h\ , \end{aligned} \tag{4.73}$$

where $(\cdot,\cdot)$ denotes the $\mathrm{L}^2(\Omega)$ scalar product and V_h is a suitable finite element subspace of $\mathsf{V} = \mathrm{H}_0^1(\Omega)$. By assuming that (4.67) and (4.70) hold and that $\underline{c}(t)|_{\partial\Omega} = 0$ for each $t \in [0,T]$, the stability can be proven exactly as before. Pironneau [104] has shown that when V_h is taken as the space of continuous and piecewise-linear polynomials vanishing on $\partial\Omega$ we have that $\|\underline{u}(t_n) - \underline{u}_h^n\| = O(h + \Delta t + \frac{h^2}{\Delta t})$.
When implementing this method, some problems appear. In fact, one has to compute the integrals $(u_h^n \circ \mathrm{X}^n, v_h)$, and this usually accomplished by means of a quadrature formula. Hence, the effects of this procedure can give rise to instability phenomena see [97]. Moreover, numerical integration requires the knowledge of the value of $u_h^n \circ \mathrm{X}^n$ at some nodal points. This means that, for any fixed node x_k, it is necessary to know which triangle $K \in \mathcal{I}_h$ contains the point $\mathrm{X}^n(x_k)$. Other remarks on the implementation of the method can be found in Pironneau [104] and Priestley [106]. For some developments on the the methods of characteristics, we suggest to the reader these papers [17, 39, 98, 105, 118, 138].

4.3 New fractional step for convection diffusion equations

4.3.1 Weak and algebraic formulation

In order to fix the ideas let us consider the equation that expresses the conservation of energy written in term of a temperature T. It is given by

$$\frac{\partial \mathsf{T}}{\partial t} + \underline{c} \cdot \nabla \mathsf{T} - \Gamma\ \Delta \mathsf{T} = \mathsf{S} \quad \text{in} \quad \Omega \tag{4.74}$$

where $\underline{c} = (u, v)$ is a known velocity field, S is the source term, Γ is a known thermal diffusion parameter and Ω is a bounded subset of $\mathbb{R}^2$ with Lipschitz boundary $\partial\Omega$. We complete (4.74) by adding initial condition

$$\mathsf{T}|_{t=0} = \mathsf{T}^0 \tag{4.75}$$

and suitable boundary conditions (any kind specified in chapter 1 section 1.1.1). Here we consider Dirichlet homogeneous boundaries conditions

$$\mathsf{T} = 0 \quad \text{on} \quad \partial\Omega\ . \tag{4.76}$$

We apply the implicit discretization and the characteristics in the convective term of equation (4.74). It comes for $n = 0, \cdots, \mathcal{N}-1$,

$$\frac{\mathsf{T}^{n+1} - \mathsf{T}^n(\tilde{\underline{X}})}{\Delta t} - \Gamma\, \Delta \mathsf{T}^{n+1} = \mathsf{S}^{n+1} \tag{4.77}$$

where $\tilde{\underline{X}}$ is the foot of characteristics.
Let $\mathcal{I}_h$ be a suitable triangulation of Ω. The weak formulation of (4.77) with (4.76) is the following:

$$\int_\Omega \frac{\mathsf{T}^{n+1} - \mathsf{T}^n(\tilde{\underline{X}})}{\Delta t}\varphi\, d\Omega - \int_\Omega \Gamma\, \Delta \mathsf{T}^{n+1}\varphi\, d\Omega = \int_\Omega \mathsf{S}^{n+1}\varphi\, d\Omega \quad \forall\ \varphi \in \mathrm{H}_0^1(\Omega) \tag{4.78}$$

i.e.

$$\int_\Omega \mathtt{T}^{n+1}\,\varphi\,d\Omega \;-\int_\Omega \mathtt{T}^{n}(\underline{\tilde{X}})\varphi\,d\Omega \;+\Delta t\int_\Omega \Gamma\,\nabla\mathtt{T}^{n+1}\cdot\,\nabla\varphi\,d\Omega = \Delta t\int_\Omega \mathtt{S}^{n+1}\,\varphi\,d\Omega \quad \forall\,\varphi\in \mathrm{H}^1_0(\Omega)\ . \tag{4.79}$$

By taking basis functions traditional 2D polynomials of second degree with six nodes on the boundary, it comes

$$\sum_{K\in\mathcal{I}_h}\Big[\sum_i \mathtt{T}_i^{n+1}\int_K \varphi_i\,\varphi_j\,dK \;-\Delta t\Gamma\sum_i \mathtt{T}_i^{n+1}\int_K \nabla\varphi_i\cdot\,\nabla\varphi_j\,d\,K\Big] = \sum_{K\in\mathcal{I}_h}\Big[\sum_i\int_K \mathtt{T}_i^{n}(\underline{\tilde{X}})\varphi_i\varphi_j\,d\,K \;+\Delta t\sum_i \mathtt{S}_i^{n+1}\int_K \varphi_i\,\varphi_j\,dK\Big] \quad i,j=1,\cdots,N_h \tag{4.80}$$

i.e. in algebraic form:

$$\mathrm{M}\underline{\mathtt{T}}^{n+1}+\Gamma\,\Delta t\mathrm{A}\underline{\mathtt{T}}^{n+1}=\Delta t\mathrm{M}\underline{\mathtt{S}}^{n+1}+\int_\Omega \underline{\mathtt{T}}^{n}(\underline{\tilde{X}})\varphi\,d\Omega$$

thus

$$(\,\mathrm{M}+\Gamma\,\Delta t\mathrm{A}\,)\underline{\mathtt{T}}^{n+1}=\Delta t\mathrm{M}\underline{\mathtt{S}}^{n+1}+\int_\Omega \underline{\mathtt{T}}^{n}(\underline{\tilde{X}})\,\varphi\,d\Omega \tag{4.81}$$

where

$$\mathrm{M}:=\int_\Omega \varphi_i\,\varphi_j\,d\Omega \quad \text{is the mass matrix}$$

and

$$\mathrm{A}:=\int_\Omega \nabla\varphi_i\cdot\,\nabla\varphi_j\,d\Omega \quad \text{is the stiffness matrix}\ .$$

The application of the formulation (4.81) requires the respect of the following iterative procedure [74]:

Iterative procedure

step 1) The nodal values of $\underline{\mathtt{T}}$ are given at time instant $t=t_n$

step 2) Given the velocity $\underline{c}$ at time instant $t=t_{n+1}$

step 3) Compute $\int_\Omega \underline{\mathtt{T}}^{n}(\underline{\tilde{X}})\varphi\,d\Omega$ (details of its computation in the next section)

step 4) Compute the nodal values of $\underline{\mathtt{T}}$ at time instant $t=t_{n+1}$

step 5) Go to step 1 and repeat the above procedure until the desired solution is obtained.

4.3.2 Computation of the step 3 of the iterative procedure

We have $\int_\Omega \underline{\mathtt{T}}^{n}(\underline{\tilde{X}})\varphi\,d\Omega=\sum_K\int_K(\sum_i \mathtt{T}_i^{n}(\underline{\tilde{X}})\varphi_i)\varphi_j\,dK$.
To compute $\underline{\mathtt{T}}^{n}(\underline{\tilde{X}})$ and $\int_\Omega \underline{\mathtt{T}}^{n}(\underline{\tilde{X}})\varphi\,d\Omega$, we proceed as follows:

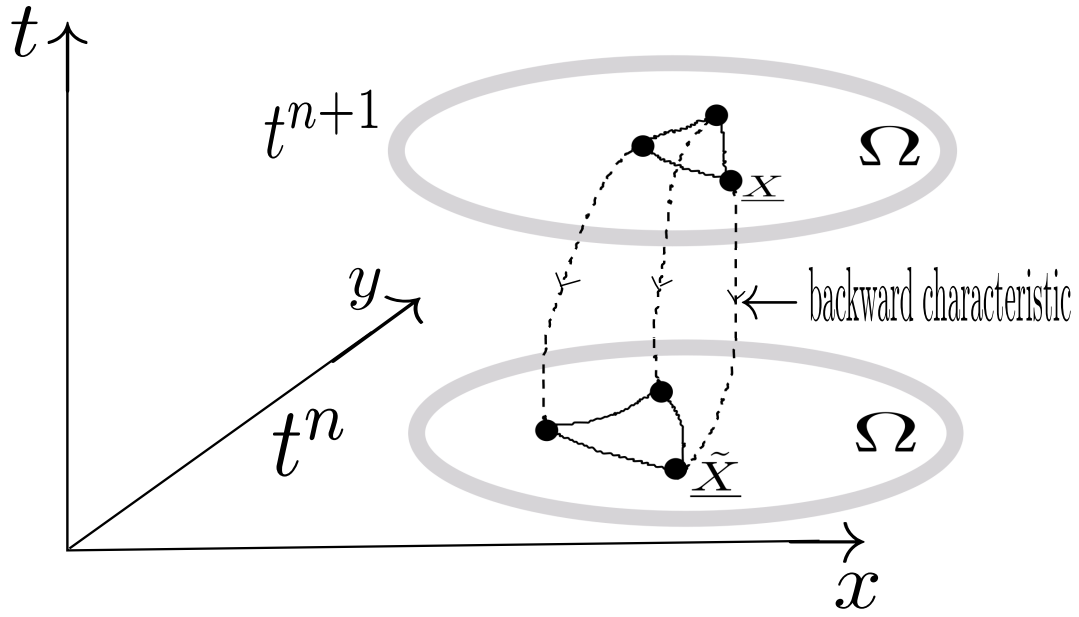

Figure 4.1: Backward characteristics, foot $\underline{\tilde{X}}$ of trajectory.

- Compute the foot of characteristic $\underline{\tilde{X}}$ by means of one of these expressions

$$\begin{aligned}
&\underline{\tilde{X}} = \underline{X} - \Delta t\, \underline{c}^{n+1}(\underline{X}) \\
&\underline{\tilde{X}} = \underline{X} - \Delta t[\ 2\underline{c}^{n+1}(\underline{X}) - \underline{c}^{n}(\underline{X})\ ] \\
&\underline{\tilde{X}} = \underline{X} - \Delta t\, \underline{c}^{n+1}(\underline{X} - \frac{\Delta t}{2}\, \underline{c}^{n+1}(\underline{X})) \\
&\underline{\tilde{X}} = \underline{X} - \Delta t[\ 2\,\underline{c}^{n+1}(\underline{X} - \frac{\Delta t}{2}\, \underline{c}^{n+1}(\underline{X})) - \underline{c}^{n}(\underline{X} - \frac{\Delta t}{2}\, \underline{c}^{n+1}(\underline{X}))\ ]\ ,
\end{aligned} \tag{4.82}$$

  where $\underline{X}$ is the physical coordinate of the node of the element in the triangulation $\mathcal{I}$ at time instant $t = t_{n+1}$, see Figure 4.1 (the coordinate of the barycenter of the element included)

- Find the element in the triangulation $\mathcal{I}_h$ that contains the foot of characteristic $\underline{\tilde{X}}$ and interpolate the value of $\underline{\mathtt{T}}$ obtained at time instant $t = t_n$ that gives the nodal values $\underline{\mathtt{T}}^n(\underline{\tilde{X}})$

- Use a Gauss quadrature formula i.e.

$$\int_K (\sum_i \mathtt{T}_i^n(\underline{\tilde{X}})\varphi_i)\varphi_j \cong \sum_{l=1}^{7}(\sum_{i=1}^{6} \mathtt{T}_i^n(\underline{\tilde{X}_l})\varphi_i(\xi_l,\eta_l))\,\varphi_j(\xi_l,\eta_l)\,\mathtt{J}(\xi_l,\eta_l)\,w_l\,, \quad i,j = 1,\cdots,6$$

where w_l are the weights of the Gauss points, $\mathtt{J}(\xi_l,\eta_l)$ is the modulus of the Jacobian of the transformation involving the element K and the reference element, all evaluated at the seven integrations points (vertice, mid-edge and centroid points).

Remark 4.20. *The computation of the foot of characteristic should be consistent with the boundary condition considered. For instance, the foot of characteristic has always to lie inside the physical domain* Ω.

4.3.3 Conservation property

In chapter 2, it has been presented a new FE scheme by means of which, using 2D second degree polynomials for base and test functions, is possible to guarantee the conservation of the

mathematical fluxes in elliptic problems. As a matter of the fact the new fractional step for convection diffusion problems can be interpreted like a classical parabolic ones (see equation (4.77)), we decided to carry out some numerical experiences in order to heuristically verify the efficiency of the conservative FE in parabolic problems. In the section 4.3.4, the Tables 4.8 - 4.10 show the results of the numerical tests.

Remark 4.21. *We should remind that our conservative approach does not respect the equilibrium between source term and mathematical (numerical) fluxes. In order to obtain a method genuinely conservative for example like those developed in the FVE framework, a generalization of the techniques presented in [109] could be afforded.*

4.3.4 Numerical tests

In order to check the accuracy of the formulation (4.81), several problems with known analytical solutions have been considered. The computations were performed in $\Omega = [0,1]^2$. In Tables 4.1 - 4.7, we reported the L^∞ and L^2 error norms generated by the numerical solutions. We have considered in all cases Dirichlet boundary conditions. We denote by $\underline{\tilde{X}}_1$, $\underline{\tilde{X}}_2$, $\underline{\tilde{X}}_3$, $\underline{\tilde{X}}_4$, the foot of characteristic obtained using the expression $(4.82)_1$, $(4.82)_2$, $(4.82)_3$ and $(4.82)_4$ respectively. Let us denote by N_h the total number of nodes with respect to each domain considered, by Δt the time increment, by Nts the number of time steps, by $\underline{c} = (u, v)$ the velocity and by $\mathtt{T}$ the temperature. We define the following two expressions for the analytical velocity and temperature:

$$\underline{c}^I = (u^I, v^I) \text{ such that } \quad u^I = -\cos(\pi x)\sin(\pi y)\exp(-2t)$$
$$v^I = \sin(\pi x)\cos(\pi y)\exp(-2t)$$

$$\underline{c}^{II} = (u^{II}, v^{II}) \text{ such that } \quad u^{II} = -\cos(\pi x)\sin(\pi y)$$
$$v^{II} = \sin(\pi x)\cos(\pi y)$$

$$\mathtt{T}^I = xy(1-x)(1-y)\exp(-2t)$$

$$\mathtt{T}^{II} = xy(1-x)(1-y)\ .$$

Test 4.3.4.1

Table 4.1: $N_h = 889$, $\Delta t = 0.001$, $Nts = 10$, $\underline{c} = \underline{c}^I$ and $\mathtt{T} = \mathtt{T}^I$.

$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$\underline{\tilde{X}}_1$	$3.281E-4$	$9.416E-4$
$\underline{\tilde{X}}_2$	$3.084E-4$	$9.168E-4$
$\underline{\tilde{X}}_3$	$3.346E-4$	$9.792E-4$
$\underline{\tilde{X}}_4$	$3.142E-4$	$9.460E-4$

Test 4.3.4.2

Table 4.2: $N_h = 109$, $\Delta t = 0.001$, $Nts = 10$, $\underline{c} = \underline{c}^I$ and $\mathtt{T} = \mathtt{T}^I$.

$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$\underline{\tilde{X}}_1$	$2.077E-4$	$4.418E-4$
$\underline{\tilde{X}}_2$	$2.072E-4$	$4.414E-4$
$\underline{\tilde{X}}_3$	$2.078E-4$	$4.420E-4$
$\underline{\tilde{X}}_4$	$2.072E-4$	$4.4152E-4$

Test 4.3.4.3

Table 4.3: $N_h = 241$, $\Delta t = 0.1$, $Nts = 10$, $\underline{c} = \underline{c}^I$ and $\mathtt{T} = \mathtt{T}^I$.

$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$\underline{\tilde{X}}_1$	$4.523E-4$	$1.140E-3$
$\underline{\tilde{X}}_2$	$4.430E-4$	$1.133E-3$
$\underline{\tilde{X}}_3$	$4.543E-4$	$1.148E-3$
$\underline{\tilde{X}}_4$	$4.488E-4$	$1.139E-3$

Test 4.3.4.4

Table 4.4: $N_h = 241$, $\Delta t = 0.1$, $Nts = 10$, $\underline{c} = \underline{c}^I$ and $\mathtt{T} = \mathtt{T}^{II}$.

$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$\underline{\tilde{X}}_1$	$6.486E-3$	$1.746E-2$
$\underline{\tilde{X}}_2$	$6.455E-3$	$1.744E-2$
$\underline{\tilde{X}}_3$	$6.499E-3$	$1.749E-2$
$\underline{\tilde{X}}_4$	$6.164E-3$	$1.466E-2$

Test 4.3.4.5

Table 4.5: $N_h = 733$, $\Delta t = 0.0725$, $Nts = 10$, $\underline{c} = \underline{c}^I$ and $\mathtt{T} = \mathtt{T}^I$.

$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$\underline{\tilde{X}}_1$	$2.760E-4$	$8.901E-4$
$\underline{\tilde{X}}_2$	$2.626E-4$	$8.691E-4$
$\underline{\tilde{X}}_3$	$2.792E-4$	$9.134E-4$
$\underline{\tilde{X}}_4$	$2.653E-4$	$8.876E-4$

Test 4.3.4.6

Table 4.6: $N_h = 457$, $\Delta t = 0.01$, $Nts = 1$, $Nts = 25$ and $Nts = 50$, $\underline{c} = \underline{c}^{II}$ and $\mathtt{T} = \mathtt{T}^I$.

	$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$Nts = 1$	$\underline{\tilde{X}}_1$	$4.873E-4$	$1.395E-3$
$Nts = 25$	$\underline{\tilde{X}}_1$	$9.872E-3$	$3.279E-2$
$Nts = 50$	$\underline{\tilde{X}}_1$	$9.948E-3$	$2.884E-2$

Test 4.3.4.7

Table 4.7: $N_h = 457$, $\Delta t = 0.001$, $Nts = 1$ and $Nts = 500$, $\underline{c} = \underline{c}^{II}$ and $\mathtt{T} = \mathtt{T}^I$.

	$\underline{\tilde{X}}$	$L^\infty-$ norm	L^2-norm
$Nts = 1$	$\underline{\tilde{X}}_1$	$2.571E-4$	$6.010E-4$
$Nts = 500$	$\underline{\tilde{X}}_1$	$8.310E-3$	$2.400E-2$

In order to verify the accuracy of the conservative FE scheme, we solved three problems like (4.77) with the following known solutions u:

$$u_I = (x+y)\exp(-t)$$
$$u_{II} = xy(1-x)(1-y)\exp(-2t)$$
$$u_{III} = -\cos(\pi\, x)\sin(\pi\, y)\exp(-2\, t)\ .$$

In all the computations, we assigned Dirichlet boundary conditions and chose the parameters: 72 elements with $N_h = 169$ and $\Delta t = 0.005$. In Tables 4.8 - 4.10 are reported the L^∞ error norm values calculated by traditional (FE) and conservative finite elements (FEC) for $Nts = 2,\ 20,\ 100$ and 200. $Max\ u$ is the maximum value of the known solution after Nts time steps.

Test 4.3.4.8

Table 4.8: $u = u_I$ and $L^\infty-$ norm after Nts time steps.

	$Nts = 2$	$Nts = 20$	$Nts = 100$	$Nts = 200$
$L^\infty-$norm by FE	$4.691E-5$	$7.278E-5$	$5.092E-5$	$3.084E-5$
$L^\infty-$norm by FEC	$8.894E-5$	$1.748E-4$	$1.250E-4$	$7.588E-5$
$Max\ u$	$1.980E+0$	$1.800E+0$	$1.210E+0$	$7.357E-1$

Test 4.3.4.9

Table 4.9: $u = u_{II}$ and $L^\infty-$ norm after Nts time steps.

	$Nts = 2$	$Nts = 20$	$Nts = 100$	$Nts = 200$
$L^\infty-$norm by FE	$5.169E-5$	$5.637E-5$	$2.661E-5$	$9.793E-6$
$L^\infty-$norm by FEC	$1.438E-3$	$4.604E-3$	$2.223E-3$	$8.215E-4$
$Max\ u$	$6.120E-2$	$5.117E-2$	$2.200E-2$	$8.450E-3$

Test 4.3.4.10

Table 4.10: $u = u_{III}$ and $L^\infty-$ norm after Nts time steps.

	$Nts = 2$	$Nts = 20$	$Nts = 100$	$Nts = 200$
$L^\infty-$norm by FE	$2.633E-3$	$2.420E-3$	$1.080E-3$	$4.007E-4$
$L^\infty-$norm by FEC	$1.907E-2$	$2.770E-2$	$1.251E-2$	$4.600E-3$
$Max\ u$	$9.800E-1$	$8.100E-1$	$3.600E-1$	$1.353E-1$

4.4 Conclusions of the chapter

Checking Tables 4.8-4.10 and making the comparison with the error norms, we conclude that the conservative FE method applied to to parabolic problems keeps the conservation of numerical fluxes (property heuristically verified), but has not an optimal order like in elliptic problems (see Conclusions of chapter 2).

Chapter 5

Solution of Navier-Stokes equations

In this chapter, we aim to address a new numerical approach for the solution of 2D incompressible Navier-Stokes problems presented in chapter 1. In particular we will present an equal-order finite element (with second degree polynomials) fractional step method, then due to the L.B.B. condition we will formalize at algebraic level a stabilization technique. Finally some numerical results, with two test problems having known exact analytic solutions and a real problem that is the natural convection in a square cavity, will be given.

5.1 Equal-order finite element and characteristic-fractional step approach

The problem we addressed in Chapter 1 was to solve:

$$\frac{\partial \underline{u}}{\partial t} + \underline{u} \cdot \nabla \underline{u} - \mu \Delta \underline{u} + \nabla p = \underline{f} \quad \text{in} \quad \Omega \times (0,T) \tag{5.1}$$

$$\nabla \cdot \underline{u} = 0 \quad \text{in} \quad \Omega \times (0,T) \tag{5.2}$$

$$\underline{u} = 0 \quad \text{on} \quad \partial\Omega \times (0,T) \ , \qquad \underline{u}|_{t=0} = \underline{u}_0 \quad \text{in} \quad \Omega \times (0,T) \tag{5.3}$$

$$\frac{\partial \mathtt{T}}{\partial t} + \underline{u} \cdot \nabla \mathtt{T} - \lambda \Delta \mathtt{T} = \mathtt{S} \quad \text{in} \quad \Omega \times (0,T) \tag{5.4}$$

$$\mathtt{T} = 0 \quad \text{on} \quad \partial\Omega \times (0,T) \ , \qquad \mathtt{T}|_{t=0} = \mathtt{T}_0 \quad \text{in} \quad \Omega \times (0,T) \tag{5.5}$$

where Ω is a bounded subset of $\mathbb{R}^2$ with Lipschitz boundary $\partial\Omega$.

5.1.1 Choice of functional spaces

In order to obtain a weak formulation of problem (5.1) - (5.3), we formally multiply (5.1) by a suitable functional $\underline{\psi}$ and by integration in Ω we obtain:

$$\int_\Omega \frac{\partial \underline{u}}{\partial t} \cdot \underline{\psi} \, d\Omega + \int_\Omega [(\underline{u} \cdot \nabla)\underline{u}] \cdot \underline{\psi} \, d\Omega - \int_\Omega \mu \Delta \underline{u} \cdot \underline{\psi} \, d\Omega + \int_\Omega \nabla p \cdot \underline{\psi} \, d\Omega = \int_\Omega \underline{f} \cdot \underline{\psi} \, d\Omega \, . \tag{5.6}$$

Using the Green formula we have:

$$-\int_\Omega \mu \Delta \underline{u} \cdot \underline{\psi} \, d\Omega = \int_\Omega \mu \nabla \underline{u} \cdot \nabla \underline{\psi} \, d\Omega - \int_{\partial\Omega} \mu \frac{\partial \underline{u}}{\partial n} \cdot \underline{\psi} \, d\Gamma \tag{5.7}$$

and

$$\int_\Omega \nabla p \cdot \underline{\psi} \, d\Omega = -\int_\Omega p \nabla \cdot \underline{\psi} \, d\Omega + \int_{\partial\Omega} p \, \underline{\psi} \cdot n \, d\Gamma \, . \tag{5.8}$$

Keeping into account the homogeneous boundary conditions and by substitution of the last two relations in the momentum equation, we obtain

$$\int_\Omega \frac{\partial \underline{u}}{\partial t}\cdot \underline{\psi}\, d\Omega + \int_\Omega [(\underline{u}\cdot\nabla)\underline{u}]\cdot \underline{\psi}\, d\Omega + \int_\Omega \mu\nabla\underline{u}\cdot\nabla\underline{\psi}\, d\Omega - \int_\Omega p\nabla\cdot\underline{\psi}\, d\Omega = \int_\Omega \underline{f}\cdot\underline{\psi}\, d\Omega + \int_{\partial\Omega}(\mu\frac{\partial \underline{u}}{\partial n} - p\, n)\cdot\underline{\psi}\, d\Gamma \quad \forall\, \underline{\psi}\in \mathrm{V}\ . \tag{5.9}$$

By multiplying (5.2) by a suitable test function $\varphi \in \mathcal{Q}$, we obtain

$$\int_\Omega \varphi\nabla\cdot\underline{u}\, d\Omega = 0 \quad \forall\, \varphi\in\mathcal{Q}\ . \tag{5.10}$$

The spaces V and $\mathcal{Q}$ can be chosen $\mathrm{V} := [\mathrm{H}_0^1(\Omega)]^2$ and $\mathcal{Q} := \mathrm{L}^2(\Omega)$ respectively.

5.1.2 Characteristic - fractional step and weak formulation

As the momentum equation (5.1) has a non linear convective part, let us make this approximation to the convective term

$$\frac{\partial \underline{u}}{\partial t} + \underline{u}\cdot\nabla\underline{u} = \frac{D\underline{u}}{Dt} \approx \frac{\tilde{\underline{u}}^{n+\frac{1}{2}} - \tilde{\underline{u}}^n(\underline{\tilde{X}})}{\Delta t} \tag{5.11}$$

where the operator $\frac{D(\cdot)}{Dt}$ is the total derivative and $\underline{\tilde{X}}$ is the foot of the trajectory of the characteristic (see Chapter 4 Section 4.3).
Since we wish to use equal-order velocity pressure approximation, the velocity components and the pressure will belong to the same approximation polynomial of degree two space.
The weak formulation of characteristic and fractional step (see chapter 1) is the following:

$$\int_\Omega \frac{\tilde{\underline{u}}^{n+\frac{1}{2}} - \underline{u}^n(\underline{\tilde{X}})}{\Delta t}\cdot\underline{\psi}\, d\Omega - \int_\Omega \mu\Delta\tilde{\underline{u}}^{n+\frac{1}{2}}\cdot\underline{\psi}\, d\Omega = -\int_\Omega \nabla p^n\cdot\underline{\psi}\, d\Omega + \int_\Omega \underline{f}^{n+1}\cdot\underline{\psi}\, d\Omega \quad \forall\,\underline{\psi}\in\mathrm{V} \tag{5.12}$$

From

$$\frac{u^{n+1} - \tilde{u}^{n+\frac{1}{2}}}{\Delta t} + \nabla\tilde{p}^{n+1} = 0 \quad \text{with} \quad \nabla\cdot u^{n+1} = 0 \tag{5.13}$$

it comes

$$-\nabla\cdot\tilde{u}^{n+1} = -\Delta t\ \Delta\tilde{p}^{n+1}$$

which is a Poisson equation for the pressure and its weak formulation is given by

$$-\int_\Omega \nabla\cdot\tilde{\underline{u}}^{n+\frac{1}{2}}\varphi\, d\Omega + \Delta t\int_\Omega \Delta\tilde{p}^{n+1}\varphi\, d\Omega = 0 \quad \forall\,\varphi\in\mathcal{Q}\ . \tag{5.14}$$

where $\tilde{\underline{u}}^{n+\frac{1}{2}}$ is the provisional value of the velocity field and $\tilde{p}^{n+1} = p^{n+1} - p^n$ is the provisional value of pressure field.

Therefore, (5.12) and (5.14) can be rewritten as

$$\int_\Omega \frac{\tilde{\underline{u}}^{n+\frac{1}{2}} - \underline{u}^n(\underline{\tilde{X}})}{\Delta t}\cdot\underline{\psi}\, d\Omega + \mu\int_\Omega \nabla\tilde{\underline{u}}^{n+\frac{1}{2}}\cdot\nabla\underline{\psi}\, d\Omega = -\int_\Omega \nabla p^n\cdot\underline{\psi}\, d\Omega + \int_\Omega \underline{f}^{n+1}\cdot\underline{\psi}\, d\Omega \quad \forall\,\underline{\psi}\in\mathrm{V} \tag{5.15}$$

$$\int_\Omega \nabla\tilde{p}^{n+1}\cdot\nabla\varphi\,d\Omega = -\frac{1}{\Delta t}\int_\Omega \nabla\cdot\underline{\tilde{u}}^{n+\frac{1}{2}}\varphi\,d\Omega \quad \forall\,\varphi\in\mathcal{Q}\ . \tag{5.16}$$

We complete this fractional step by adding the pressure correction

$$p^{n+1} = p^n + \tilde{p}^{n+1} \tag{5.17}$$

and the velocity correction

$$\underline{u}^{n+1} = \underline{\tilde{u}}^{n+\frac{1}{2}} - \Delta t\nabla\tilde{p}^{n+1}\ . \tag{5.18}$$

Rewriting this fractional step with respect to each component of the velocity field $\underline{u}=(u,v)$ and applying the weak formulation in the velocity correction (5.18), it comes with $\underline{\psi}=(\psi_1,\psi_2)$

$$\int_\Omega \frac{\tilde{u}^{n+\frac{1}{2}}-u^n(\underline{\tilde{X}})}{\Delta t}\psi_1\,d\Omega + \mu\int_\Omega \nabla\tilde{u}^{n+\frac{1}{2}}\cdot\nabla\psi_1\,d\Omega = -\int_\Omega \frac{\partial p^n}{\partial x}\psi_1\,d\Omega + \int_\Omega f_u^{n+1}\psi_1\,d\Omega \tag{5.19}$$

$$\int_\Omega \frac{\tilde{v}^{n+\frac{1}{2}}-v^n(\underline{\tilde{X}})}{\Delta t}\psi_2\,d\Omega + \mu\int_\Omega \nabla\tilde{v}^{n+\frac{1}{2}}\cdot\nabla\psi_2\,d\Omega = -\int_\Omega \frac{\partial p^n}{\partial y}\psi_2\,d\Omega + \int_\Omega f_v^{n+1}\psi_2\,d\Omega \tag{5.20}$$

$$\int_\Omega \nabla\tilde{p}^{n+1}\cdot\nabla\varphi\,d\Omega = -\frac{1}{\Delta t}\int_\Omega (\frac{\partial\tilde{u}^{n+\frac{1}{2}}}{\partial x}+\frac{\partial\tilde{v}^{n+\frac{1}{2}}}{\partial y})\varphi\,d\Omega \tag{5.21}$$

$$p^{n+1} = p^n + \tilde{p}^{n+1} \tag{5.22}$$

$$\int_\Omega u^{n+1}\varphi\,d\Omega = \int_\Omega \tilde{u}^{n+\frac{1}{2}}\varphi\,d\Omega - \Delta t\int_\Omega \frac{\partial\tilde{p}^{n+1}}{\partial x}\varphi\,d\Omega \tag{5.23}$$

$$\int_\Omega v^{n+1}\varphi\,d\Omega = \int_\Omega \tilde{v}^{n+\frac{1}{2}}\varphi\,d\Omega - \Delta t\int_\Omega \frac{\partial\tilde{p}^{n+1}}{\partial y}\varphi\,d\Omega \tag{5.24}$$

that is the problem becomes: find $(u^{n+1},v^{n+1})\in(\mathrm{V}_1,\mathrm{V}_2)\in\mathrm{H}_0^1(\Omega)^2$ and $p^{n+1}\in\mathcal{Q}$ such that

$$\begin{aligned}\frac{1}{\Delta t}\int_\Omega \tilde{u}^{n+\frac{1}{2}}\psi_1\,d\Omega + \mu\int_\Omega \nabla\tilde{u}^{n+\frac{1}{2}}\cdot\nabla\psi_1\,d\Omega = &-\int_\Omega \frac{\partial p^n}{\partial x}\psi_1\,d\Omega + \int_\Omega f_u^{n+1}\psi_1\,d\Omega + \\ &\frac{1}{\Delta t}\int_\Omega u^n(\underline{\tilde{X}})\psi_1\,d\Omega \quad \forall\ \psi_1\in\mathrm{V}_1\end{aligned} \tag{5.25}$$

$$\begin{aligned}\frac{1}{\Delta t}\int_\Omega \tilde{v}^{n+\frac{1}{2}}\psi_2\,d\Omega + \mu\int_\Omega \nabla\tilde{v}^{n+\frac{1}{2}}\cdot\nabla\psi_2\,d\Omega = &-\int_\Omega \frac{\partial p^n}{\partial y}\psi_2\,d\Omega + \int_\Omega f_v^{n+1}\psi_2\,d\Omega + \\ &\frac{1}{\Delta t}\int_\Omega v^n(\underline{\tilde{X}})\psi_2\,d\Omega \quad \forall\ \psi_2\in\mathrm{V}_2\end{aligned} \tag{5.26}$$

$$\int_\Omega \nabla\tilde{p}^{n+1}\cdot\nabla\varphi\,d\Omega = -\frac{1}{\Delta t}\int_\Omega (\frac{\partial\tilde{u}^{n+\frac{1}{2}}}{\partial x}+\frac{\partial\tilde{v}^{n+\frac{1}{2}}}{\partial y})\varphi\,d\Omega \quad \forall\ \varphi\in\mathcal{Q} \tag{5.27}$$

$$p^{n+1} = p^n + \tilde{p}^{n+1} \tag{5.28}$$

$$\int_\Omega u^{n+1}\varphi\,d\Omega = \int_\Omega \tilde{u}^{n+\frac{1}{2}}\varphi\,d\Omega - \Delta t\int_\Omega \frac{\partial\tilde{p}^{n+1}}{\partial x}\varphi\,d\Omega \quad \forall\ \varphi\in\mathcal{Q} \tag{5.29}$$

$$\int_\Omega v^{n+1}\varphi\,d\Omega = \int_\Omega \tilde{v}^{n+\frac{1}{2}}\varphi\,d\Omega - \Delta t\int_\Omega \frac{\partial\tilde{p}^{n+1}}{\partial y}\varphi\,d\Omega \quad \forall\ \varphi\in\mathcal{Q}\ . \tag{5.30}$$

5.1.3 Algebraic formulation of the equal-order approximation

Let $\mathcal{I}_h = \{K\}$ be a suitable triangulation of the domain Ω. By using equal order velocity pressure approximation (i.e., $\psi_1 = \psi_2 = \varphi \in \mathrm{X}_h^2$) with second degree polynomial on element (having six nodes on the boundary), the system of equation (5.25) - (5.30) can be rewritten as follows :

$$
\begin{aligned}
\sum_K [\, \frac{1}{\Delta t} \sum_i \tilde{u}_i^{n+\frac{1}{2}} \int_K \varphi_i \varphi_j \, dK + \mu \sum_i \tilde{u}_i^{n+\frac{1}{2}} \int_K \nabla\varphi_i \cdot \nabla\varphi_j \, dK \,] &= \sum_K [\, \sum_i p_i^n \int_K \frac{\partial \varphi_i}{\partial x} \varphi_j \, dK + \\
\sum_i (f_i^{n+1})_u \int_K \varphi_i \varphi_j \, dK + \frac{1}{\Delta t} \int_K u^n(\underline{\tilde{X}}) \varphi_j \, dK \,] & \\
\sum_K [\, \frac{1}{\Delta t} \sum_i \tilde{v}_i^{n+\frac{1}{2}} \int_K \varphi_i \varphi_j \, dK + \mu \sum_i \tilde{v}_i^{n+\frac{1}{2}} \int_K \nabla\varphi_i \cdot \nabla\varphi_j \, dK \,] &= \sum_K [\, \sum_i p_i^n \int_K \frac{\partial \varphi_i}{\partial y} \varphi_j \, dK + \\
\sum_i (f_i^{n+1})_v \int_K \varphi_i \varphi_j \, dK + \frac{1}{\Delta t} \int_K v^n(\underline{\tilde{X}}) \varphi_j \, dK \,] & \\
\sum_K [\, \sum_i \tilde{p}_i^{n+1} \int_K \nabla\varphi_i \cdot \nabla\varphi_j \, dK \,] &= -\frac{1}{\Delta t} \sum_K [\, \sum_i \tilde{u}_i^{n+\frac{1}{2}} \int_K \frac{\partial \varphi_i}{\partial x} \varphi_j \, dK + \\
\sum_i \tilde{v}_i^{n+\frac{1}{2}} \int_K \frac{\partial \varphi_i}{\partial y} \varphi_j \, dK \,] & \\
p^{n+1} &= p^n + \tilde{p}^{n+1} \\
\sum_K [\, \sum_i u_i^{n+1} \int_K \varphi_i \varphi_j \, dK \,] &= \sum_K [\, \sum_i \tilde{u}_i^{n+\frac{1}{2}} \int_K \varphi_i \varphi_j \, dK - \\
\Delta t \sum_i \tilde{p}^{n+1} \int_K \frac{\partial \varphi_i}{\partial x} \varphi_j \, dK \,] & \\
\sum_K [\, \sum_i v_i^{n+1} \int_K \varphi_i \varphi_j \, dK \,] &= \sum_K [\, \sum_i \tilde{v}_i^{n+\frac{1}{2}} \int_K \varphi_i \varphi_j \, dK - \\
\Delta t \sum_i \tilde{p}^{n+1} \int_K \frac{\partial \varphi_i}{\partial y} \varphi_j \, dK \,] &
\end{aligned}
\tag{5.31}
$$

that is in algebraic form

$$
\begin{aligned}
\frac{1}{\Delta t} \mathrm{M} \tilde{u}^{n+\frac{1}{2}} + \mu \mathrm{A} \tilde{u}^{n+\frac{1}{2}} &= -\mathrm{B}^T p^n + \mathrm{M} f_u^{n+1} + \frac{1}{\Delta t} \int_\Omega \underline{u}^n(\underline{\tilde{X}}) \varphi \, d\Omega \\
\frac{1}{\Delta t} \mathrm{M} \tilde{v}^{n+\frac{1}{2}} + \mu \mathrm{A} \tilde{v}^{n+\frac{1}{2}} &= -\mathrm{C}^T p^n + \mathrm{M} f_v^{n+1} + \frac{1}{\Delta t} \int_\Omega \underline{v}^n(\underline{\tilde{X}}) \varphi \, d\Omega \\
\mathrm{A} \tilde{p}^{n+1} &= -\frac{1}{\Delta t} [\, \mathrm{B}^T \tilde{u}^{n+\frac{1}{2}} + \mathrm{C}^T \tilde{v}^{n+\frac{1}{2}} \,] \\
p^{n+1} &= p^n + \tilde{p}^{n+1} \\
\mathrm{M} u^{n+1} &= \mathrm{M} \tilde{u}^{n+\frac{1}{2}} - \Delta t \mathrm{B}^T \tilde{p}^{n+1} \\
\mathrm{M} v^{n+1} &= \mathrm{M} \tilde{v}^{n+\frac{1}{2}} - \Delta t \mathrm{C}^T \tilde{p}^{n+1}
\end{aligned}
\tag{5.32}
$$

where

$$
\mathrm{A} = \sum_K \int_K \nabla\varphi_i \cdot \nabla\varphi_j \, dK \quad \text{is the stiffness matrix,}
$$

$$\mathrm{M} = \sum_K \int_K \varphi_i \varphi_j \, dK \quad \text{is the mass matrix,}$$

$$\mathrm{B} = \sum_K \int_K \frac{\partial \varphi_i}{\partial x} \varphi_j \, dK \quad \text{is the matrix of the component along } x \text{ of the gradient operator}$$

and

$$\mathrm{C} = \sum_K \int_K \frac{\partial \varphi_i}{\partial y} \varphi_j \, dK \quad \text{is the matrix of the component along } y \text{ of the gradient operator.}$$

The algebraic system (5.32) can be written in a more compact form

$$\begin{aligned}
(\frac{1}{\Delta t}\mathrm{M} + \mu \mathrm{A})\tilde{u}^{n+\frac{1}{2}} &= -\mathrm{B}^T p^n + \mathrm{M} f_u^{n+1} + \frac{1}{\Delta t} \int_\Omega \underline{u}^n(\underline{\tilde{X}}) \varphi \, d\Omega \\
(\frac{1}{\Delta t}\mathrm{M} + \mu \mathrm{A})\tilde{v}^{n+\frac{1}{2}} &= -\mathrm{C}^T p^n + \mathrm{M} f_v^{n+1} + \frac{1}{\Delta t} \int_\Omega \underline{v}^n(\underline{\tilde{X}}) \varphi \, d\Omega \\
\Delta t \mathrm{A} \tilde{p}^{n+1} &= -[\, \mathrm{B}^T \tilde{u}^{n+\frac{1}{2}} + \mathrm{C}^T \tilde{v}^{n+\frac{1}{2}} \,] \\
p^{n+1} &= p^n + \tilde{p}^{n+1} \\
\mathrm{M} u^{n+1} &= \mathrm{M} \tilde{u}^{n+\frac{1}{2}} - \Delta t \mathrm{B}^T \tilde{p}^{n+1} \\
\mathrm{M} v^{n+1} &= \mathrm{M} \tilde{v}^{n+\frac{1}{2}} - \Delta t \mathrm{C}^T \tilde{p}^{n+1}
\end{aligned} \tag{5.33}$$

5.2 The inf-sup condition and some stabilization methods

Guermond and Quartapelle in their study [55] affirm that the fractional step method based on Poisson equation for pressure overcomes the inf-sup condition under a suitable limitations for Δt and for P_1/P_1 linear approximation; without respecting suitable temporal limitations or for equal order P_2/P_2 approximation the results are unstable. Since we are using equal-order P_2/P_2 approximation, a stabilization technique is needed in order to complete the formulation (5.33).

Stabilizations methods

Among the stabilized methods widespread, we can quote the Galerkin least square (GLS) technique [49, 47, 61], the least square method for first-order system such as [12] and the least square for second order scheme [46, 126] and other methods [14, 41].

The orthogonal sub-scale stabilization

Codina and Soto [35] built a stabilization technique based on the definition of the "sub scale velocities" that are responsible of the instability due to non respect of the inf-sup condition. The algebraic terms able to govern the effects of the sub scale velocities (SSV) are obtained by:

- Imposing that the SSV satisfy a momentum like equation to which is added a suitable function
- Orthogonalizing the space for the approximation of the SSV to the space V_h in which the physical velocity $\underline{u}_h$ has to be approximated.

This technique allows in particular to deal with convection dominated flows and to use equal-order velocity pressure interpolations. We apply this technique to complete the formulation (5.33).

5.3 New algebraic stabilized method

To stabilize (5.33), we have applied the technique of Codina and Soto [35] and have added the suitable algebraic terms to equations of the momentums and the Poisson equation of the pressure, to get

$$
\begin{aligned}
(\frac{1}{\Delta t}\mathrm{M}+\mu\mathrm{A})\tilde{u}^{n+\frac{1}{2}}+\tau(\mathrm{A}-\mathrm{B}^T\mathrm{M}_L^{-1}\mathrm{B}^T)p^n &= -\mathrm{B}^T p^n+\mathrm{M}f_u^{n+1}+\\
&\frac{1}{\Delta t}\int_\Omega \underline{u}^n(\underline{\tilde{X}})\varphi\, d\Omega\\
(\frac{1}{\Delta t}\mathrm{M}+\mu\mathrm{A})\tilde{v}^{n+\frac{1}{2}}+\tau(\mathrm{A}-\mathrm{C}^T\mathrm{M}_L^{-1}\mathrm{C}^T)p^n &= -\mathrm{C}^T p^n+\mathrm{M}f_v^{n+1}+\\
&\frac{1}{\Delta t}\int_\Omega \underline{v}^n(\underline{\tilde{X}})\varphi\, d\Omega\\
\Delta t\mathrm{A}(p^{n+1}-p^n)+\tau(\mathrm{A}-\mathrm{B}^T\mathrm{M}_L^{-1}\mathrm{B}^T-\mathrm{C}^T\mathrm{M}_L^{-1}\mathrm{C}^T)p^{n+1} &= -[\,\mathrm{B}^T\tilde{u}^{n+\frac{1}{2}}+\mathrm{C}^T\tilde{v}^{n+\frac{1}{2}}\,]
\end{aligned}
\tag{5.34}
$$

where τ is a weight parameter and M_L is the lumped mass matrix.
Rewriting (5.34), we obtain

$$(\mathrm{M}+\mu\Delta t\mathrm{A})\tilde{u}^{n+\frac{1}{2}}=\Delta t\{-\mathrm{B}^T p^n+\mathrm{M}f_u^{n+1}-\tau(\mathrm{A}-\mathrm{B}^T\mathrm{M}_L^{-1}\mathrm{B}^T)p^n\}+\int_\Omega \underline{u}^n(\underline{\tilde{X}})\varphi\, d\Omega \tag{5.35}$$

$$(\mathrm{M}+\mu\Delta t\mathrm{A})\tilde{v}^{n+\frac{1}{2}}=\Delta t\{-\mathrm{C}^T p^n+\mathrm{M}f_v^{n+1}-\tau(\mathrm{A}-\mathrm{C}^T\mathrm{M}_L^{-1}\mathrm{C}^T)p^n\}+\int_\Omega \underline{v}^n(\underline{\tilde{X}})\varphi\, d\Omega \tag{5.36}$$

$$[(\Delta t+\tau)\mathrm{A}-\tau(\mathrm{B}^T\mathrm{M}_L^{-1}\mathrm{B}^T+\mathrm{C}^T\mathrm{M}_L^{-1}\mathrm{C}^T)]p^{n+1}=\Delta t\mathrm{A}p^n-[\,\mathrm{B}^T\tilde{u}^{n+\frac{1}{2}}+\mathrm{C}^T\tilde{v}^{n+\frac{1}{2}}\,]\ . \tag{5.37}$$

We still have the equation for the correction of the velocity fields

$$
\begin{aligned}
\mathrm{M}u^{n+1} &= \mathrm{M}\tilde{u}^{n+\frac{1}{2}}-\Delta t\,\mathrm{B}^T\tilde{p}^{n+1}\\
\mathrm{M}v^{n+1} &= \mathrm{M}\tilde{v}^{n+\frac{1}{2}}-\Delta t\,\mathrm{C}^T\tilde{p}^{n+1}\ .
\end{aligned}
\tag{5.38}
$$

Let us go back in section 5.1 and consider the equation written in terms of the temperature and its corresponding boundary conditions (5.4) - (5.5). This equation is a linear convection diffusion that we have already treated in chapter 4 section 4.3. The formulation of this equation was given by equation (4.81). Thus the final algebraic stabilized formulation of the problem (5.1) - (5.5) is given by the following system (5.39) - (5.44):

$$(\mathrm{M}+\mu\Delta t\mathrm{A})\tilde{u}^{n+\frac{1}{2}}=\Delta t\{-\mathrm{B}^T p^n+\mathrm{M}f_u^{n+1}-\tau(\mathrm{A}-\mathrm{B}^T\mathrm{M}_L^{-1}\mathrm{B}^T)p^n\}+\int_\Omega \underline{u}^n(\underline{\tilde{X}})\varphi\, d\Omega \tag{5.39}$$

$$(\mathrm{M}+\mu\Delta t\mathrm{A})\tilde{v}^{n+\frac{1}{2}}=\Delta t\{-\mathrm{C}^T p^n+\mathrm{M}f_v^{n+1}-\tau(\mathrm{A}-\mathrm{C}^T\mathrm{M}_L^{-1}\mathrm{C}^T)p^n\}+\int_\Omega \underline{v}^n(\underline{\tilde{X}})\varphi\, d\Omega \tag{5.40}$$

$$[(\Delta t+\tau)\mathrm{A}-\tau(\mathrm{B}^T\mathrm{M}_L^{-1}\mathrm{B}^T+\mathrm{C}^T\mathrm{M}_L^{-1}\mathrm{C}^T)]p^{n+1}=\Delta t\mathrm{A}p^n-[\,\mathrm{B}^T\tilde{u}^{n+\frac{1}{2}}+\mathrm{C}^T\tilde{v}^{n+\frac{1}{2}}\,] \tag{5.41}$$

$$\mathrm{M}u^{n+1}=\mathrm{M}\tilde{u}^{n+\frac{1}{2}}-\Delta t\,\mathrm{B}^T\tilde{p}^{n+1} \tag{5.42}$$

$$\mathrm{M}v^{n+1}=\mathrm{M}\tilde{v}^{n+\frac{1}{2}}-\Delta t\,\mathrm{C}^T\tilde{p}^{n+1} \tag{5.43}$$

$$(\,\mathrm{M}+\lambda\Delta t\mathrm{A}\,)\mathtt{T}^{n+1}=\Delta t\mathrm{M}\mathtt{S}^{n+1}+\int_\Omega \underline{\mathtt{T}}^n(\underline{\tilde{X}})\varphi^{n+1}\, d\Omega \tag{5.44}$$

Choice of the weight parameter τ

According to [35], the weight is given by:

$$\tau = (c_1 + \frac{\mu}{h^2} + c_2\frac{|\underline{u}_*|}{h})^{-1}$$

where c_1 and c_2 are suitable constants, μ is the diffusion parameter, h the spatial mesh size and $\underline{u}_*$ is the velocity sub-scale. For suitable choices of c_1 and c_2, the L^2-norm of the solutions remains approximately the same over the elements. Based on our experience, the accuracy of the solution is not strongly dependent of the choice of the τ parameter. Following some numerical tests, we chose $\tau = 10^{-5}$. We note that from [55],the fractional step method is convergent of order $O((\Delta t)^2)$ provided that $\Delta t \geq ch^{l+1}$, l being the velocity interpolation degree, h the spatial mesh size and c a suitable constant.

Numerical procedure

step 1) The velocity, pressure and temperature fields are given at time $t = t^n$

step 2) Compute the intermediate velocity field by equations (5.39) and (5.40)

step 3) Compute the pressure field at time $t = t^{n+1}$ by equation (5.41)

step 4) Correct the intermediate velocity field by equations (5.42) and (5.43)

step 5) Compute the temperature field at time $t = t^{n+1}$ by equation (5.44)

step 6) Go to step 1 and repeat the above procedure until the desired solution is obtained.

Algorithm aspects

The integration over the triangles is performed by means of Gaussian quadrature rule using a seven-points formula as we used the P_2 interpolation. This assures a good evaluation of all scalar products including those which involve the characteristic terms. The Jacobian determinant, the basis functions and the basis function derivatives at Gauss points of all elements are evaluated once and for all at the beginning of the calculation and stored in arrays for subsequent use. The algorithm requires us to solve large sparse linear systems of algebraic equations for velocity, pressure and temperature. The matrices of the linear systems for the provisorial velocity components, pressure Poisson problem, correction of the velocity components and temperature does not change at each time level. The solution of the six linear systems (5.39)-(5.44) is calculated by the iterative Bi-CGSTAB method; as we realized that solving the six linear systems is similar to solving six elliptic equations at each time level, hence we used a Schwarz overlapping additive multi-domains techniques as preconditioner. This technique has been introduced in chapter 3. A suitable software has been developed and its description will be the object of chapter 7.

5.4 Numerical tests: two theoretical problems

In order to verify the efficiencies of the new method, some problems were solved. We consider a two dimensional time dependent problem introduced by Zang et al.[139] in subsection 5.4.1 and in the next paragraph, a two dimensional stationary problem by Shih et al.[128].

5.4.1 Two dimensional time dependent problem

We use the solution by Zang et al [139] for benchmarking the Navier-Stokes equations combined with the time-stepping scheme.

Problem Setting

We consider the following solution to the two dimensional unsteady flow of decaying vortices:

$$
\begin{aligned}
u &= -\cos(\pi x)\sin(\pi y)\exp(-2t) \\
v &= \sin(\pi x)\cos(\pi y)\exp(-2t) \\
p &= -\frac{1}{4}[\,\cos(2\pi x) + \cos(2\pi y)\,]\exp(-4t)
\end{aligned}
$$

where $\underline{u} = (u, v)$.
We choose the domain $\Omega = [0,1]^2$, the viscosity $\mu = 1$ and Dirichlet boundary conditions with $g_D = \underline{u}$ are imposed on the whole boundary $\partial\Omega$. The Reynolds number is $R_e = 1$ and the maximun spatial mesh size is $h_{max} = 0.07$ The time increment used is $\Delta t = 0.005$ and we used 200 time steps to reach the final time instant $t = 1$. The weight parameter is chosen as $\tau = 10^{-5}$. This flow has also been used by previous researchers such as Kim and Moin [73], Patankar [101] and Feraudi and Pennati [48] to test the accuracy of their numerical methods and approximations of boundary conditions. We compute the $\mathrm{L}^\infty-$ error norm of the numerical solution (for the two components of the velocity field and for the pressure) at the final time instant $t = 1$. By denoting e_u, e_v and e_p the $\mathrm{L}^\infty-$ error norm of the two components of the velocity and of the pressure respectively, we reported in Table 5.1 the accuracy obtained.

Table 5.1: $\mathrm{L}^\infty-$ error norm of the numerical solution.

Method	e_u	e_v	e_p
New method	$1.397E-5$	$1.588E-5$	$1.141E-2$

In Figure 5.1 and Figure 5.2 are plotted the streamlines of the flow for the numerical solution $\underline{u}_h$ and for the exact solution $\underline{u}$ respectively. Also in Figure 5.3 and Figure 5.5 and Figure 5.4 and Figure 5.6 are plotted the contours of the numerical solution u_h and v_h and analytical solution u and v respectively.

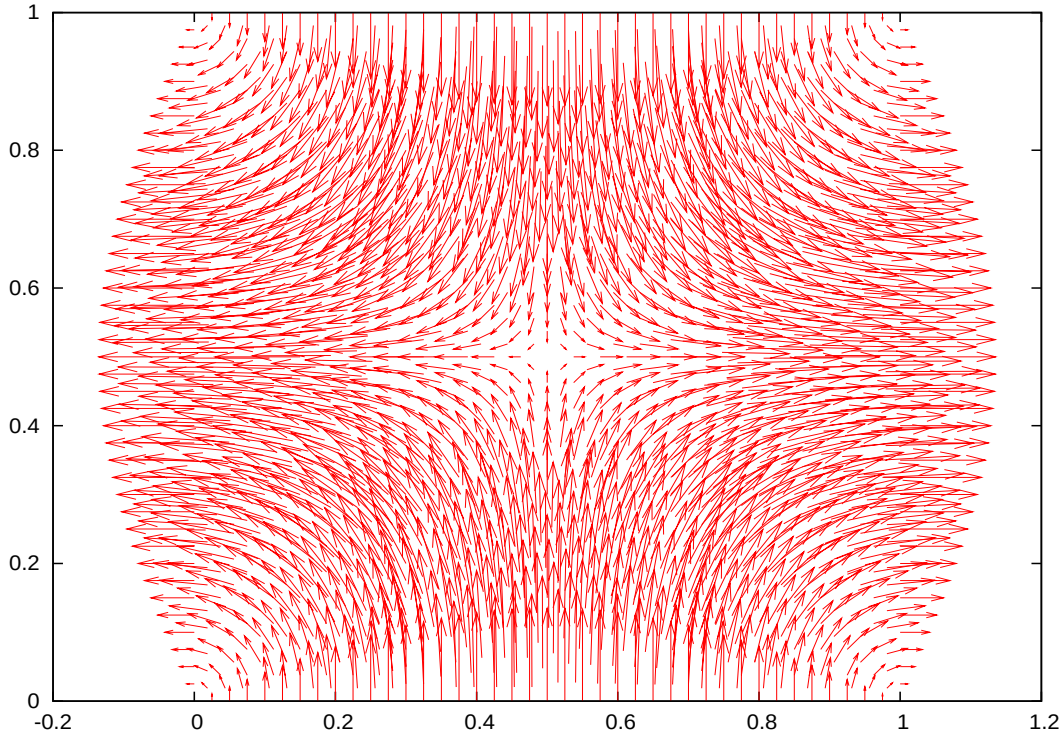

Figure 5.1: Streamlines of the flow of the numerical velocity $\underline{u}_h$.

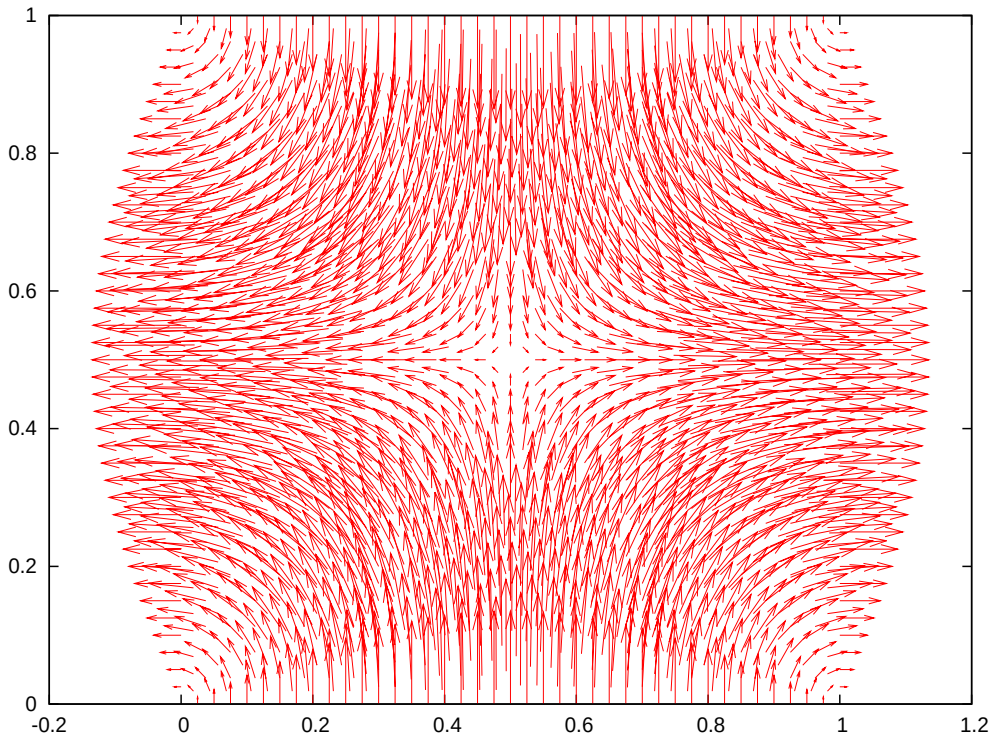

Figure 5.2: Streamlines of the flow of the analytical velocity $\underline{u}$.

In order to check the performance of the Schwarz additive overlapping preconditioner used, we computed and reported in Table 5.2 the condition number of the four matrices (matrices of the provisional velocity, of the updated pressure, of the correction of the velocity and eventually of the temperature) when they are not preconditioned $\mathcal{K}(\mathcal{A}_h)$ and when they are Schwarz

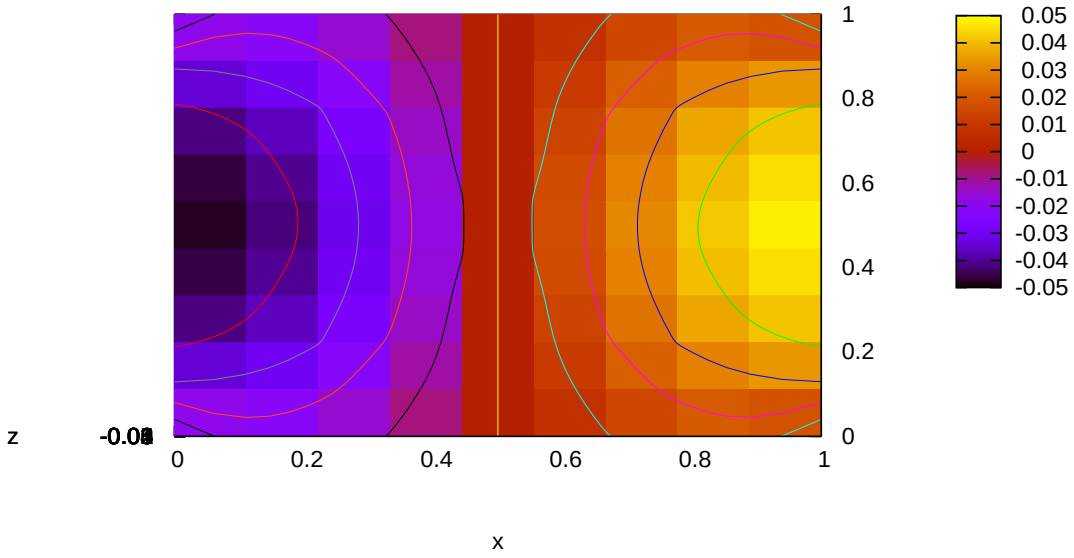

Figure 5.3: Contours of the numerical component u_h.

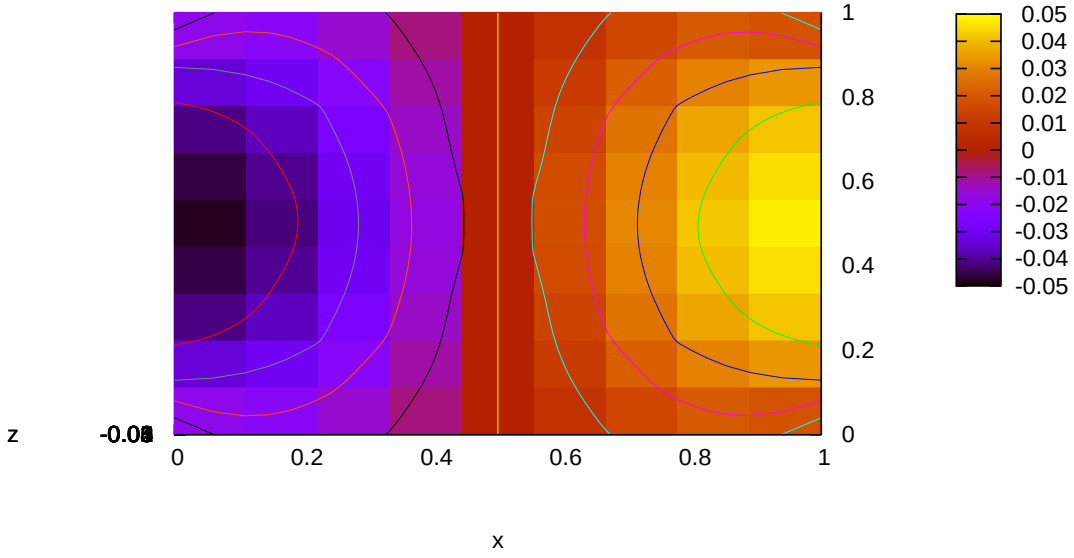

Figure 5.4: Contours of the analytical component u.

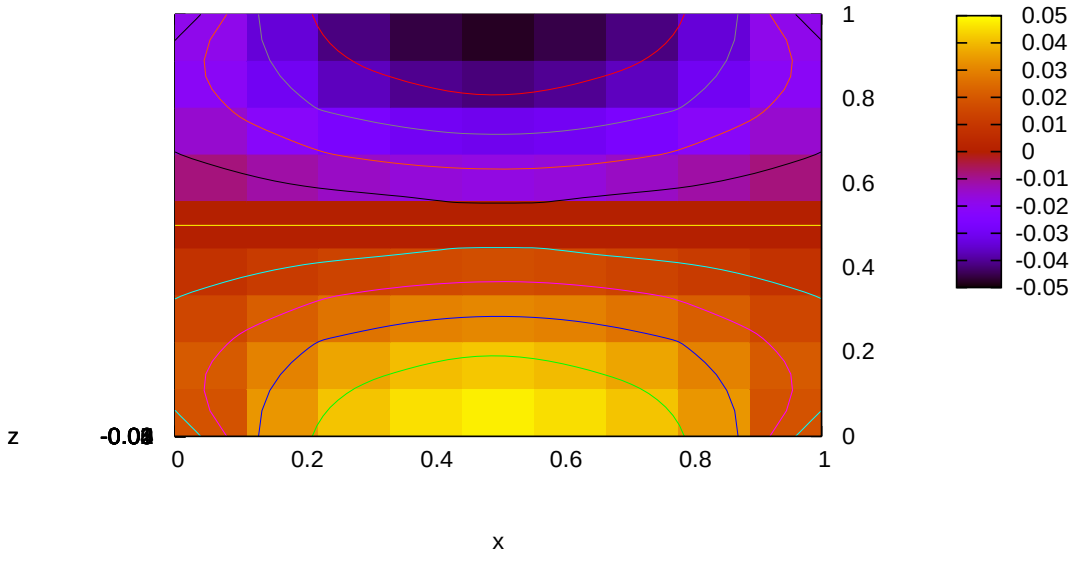

Figure 5.5: Contours of the numerical component v_h.

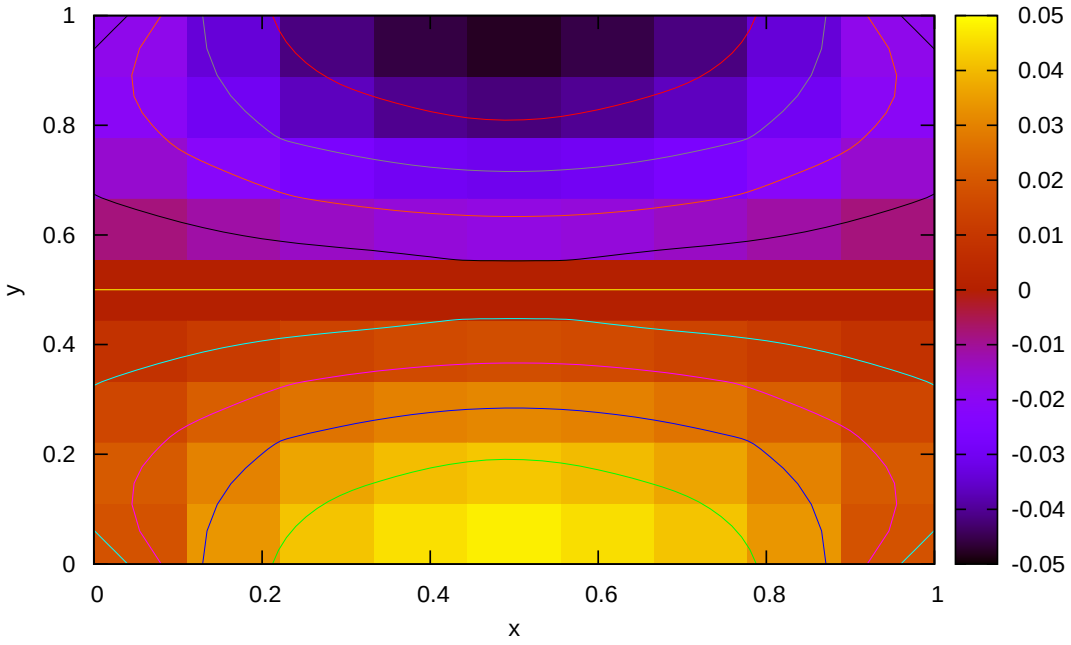

Figure 5.6: Contours of the analytical component v.

Table 5.2: Condition numbers $\mathcal{K}(\mathcal{A}_h)$ and $\mathcal{K}(\mathrm{P}_{cas}\mathcal{A}_h)$ of the not preconditioned and of the preconditioned matrices, and iteration numbers $\neq Iters$ and $\neq Iters - S$ for the solution of the non preconditioned and preconditioned systems.

Matrix of	$\mathcal{K}(\mathcal{A}_h)$	$\mathcal{K}(\mathrm{P}_{\mathrm{cas}}\mathcal{A}_h)$	$\neq Iters$	$\neq Iters - S$
$\tilde{u}^{n+\frac{1}{2}}$, $\tilde{v}^{n+\frac{1}{2}}$	$8.796E+1$	$2.213E+1$	67	8
p^{n+1}	$5.687E+8$	$4.011E+4$	250	22
u^{n+1}, v^{n+1}	$6.238E+0$	$3.623E+0$	16	5

preconditioned $\mathcal{K}(\mathrm{P}_{cas}\mathcal{A}_h)$; moreover, we collected the iteration numbers obtained using the Bi-CGSTAB iterative method when the four aforementioned linear systems are not preconditioned $\neq Iters$ and when they are Schwarz preconditioned $\neq Iters - S$. We note that by $\mathcal{A}_h$, we indicate each of the four matrices.

Results and comments

By Table 5.1, we can see that the errors of the velocity components are very low. In fact there are of 0.01034%, 0.04355% of the maximal value of the analytical solutions. As we can note from the Figures 5.1 - 5.6 in all the cases the numerical solution follows closely the trend of the analytical solution.
In the Table 5.2, we see that the Schwarz additive preconditioner works well in the sense that the condition number $\mathcal{K}(\mathrm{P}_{\mathrm{cas}}\mathcal{A}_h)$ and the iteration number $\neq Iters - S$ diminishes substantially in respect to the not preconditioned counterparts.

5.4.2 Two dimensional stationary problem

We use the solution by Shih et al. [128] for benchmarking of the Navier-Stokes approximation.

Problem setting

We consider the following stationary solution to the two dimensional Navier-Stokes equations similar to the classical lid-driven cavity flow but having known exact solution

$$
\begin{aligned}
u &= 8f(x)g'(y) = 8(x^4 - 2x^3 + x^2)(4y^3 - 2y) \\
v &= -8f'(x)g(y) = -8(4x^3 - 6x^2 + 2x)(y^4 - y^2) \\
p &= \frac{8}{R_e}[\, F(x)g'''(y) + f'(x)g'(y)\,] + 64F_1(x)\{g(y)g''(y) - [\, g'(y)\,]^2\}
\end{aligned}
$$

where

$$
\begin{aligned}
f(x) &= x^4 - 2x^3 + x^2 \\
g(y) &= y^4 - y^2 \\
F(y) &= \int f(x)\,dx = 0.2x^5 - 0.5x^4 + \frac{x^3}{3} \\
F_1(x) &= \int f(x)f'(x)\,dx = 0.5[\, f(x)\,]^2
\end{aligned}
$$

and the primes of $f(x)$ and $g(y)$ denote the differentiation with respect to x and y respectively. We chose the domain $\Omega = [0,1]^2$ and the viscosity $\mu = \frac{1}{R_e}$, where R_e is the Reynolds number. By definition,

$$
R_e = \frac{|\underline{u}|L}{\nu}
$$

Table 5.3: $L^{\infty}-$ error norm of the numerical solution $R_e = 1$.

Method	e_u	e_v	e_p
New method	$6.650E-4$	$4.710E-4$	$6.954E-1$
Shih et al. 4/1 staggered [128]	$1.373E-3$	$1.769E-3$	$2.069E-1$
Shih et al. 5/4 staggered [128]	$5.500E-4$	$7.890E-4$	$2.697E+0$

where L is the characteristic length of Ω and ν is the cinematic viscosity.
The boundary conditions for the velocities u and v are of Dirichlet type: zero everywhere except at top edge where

$$u(x,1) = 16(x^4 - 2x^3 + x^2) \tag{5.45}$$

Equation (5.45) indicates that $u(0,1) = 0$ and $u(1,1) = 0$, which eliminates the ambiguity of specifying the corner velocities as in the classical lid-driven flow problem. The time increment is $\Delta t = 0.005$ and we used 200 time steps to reach the final time $t = 1$. The weight parameter is chosen as $\tau = 10^{-5}$ for all cases.
Since we are in presence of a stationary problem, we have simulated a false transient by taking the initial value of the velocity field as the exact analytical solution and of the pressure field as the trivial value $p = 0$. We have computed the $L^{\infty}-$ error of the numerical solution generated by the new method and have reported it in the Tables 5.3, 5.5 and 5.7 for three different values of Reynolds number (R_e). We also reported the data error of Shih et al. [128].

Results first case: $R_e = 1$

We plotted in Figure 5.7 and Figure 5.8 the streamlines of the numerical velocity and analytical velocity (not scaled), in Figure 5.9 the streamlines of the flow of the numerical solution (scaled), in Figure 5.10 - 5.13 we plotted the contours of the velocity components (numerical and analytical) for $R_e = 1$ at the final time instant $t = 1$.

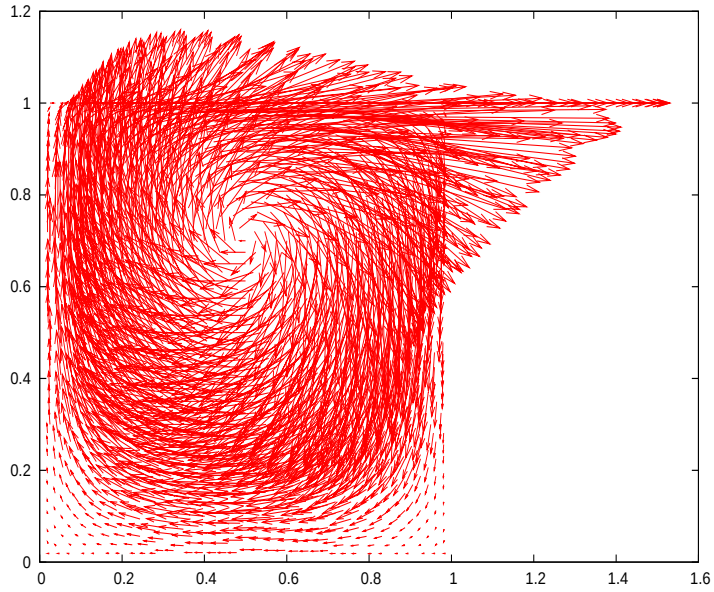

Figure 5.7: Streamlines of the flow for the numerical velocity $\underline{u}_h$ (not scaled) with $R_e = 1$.

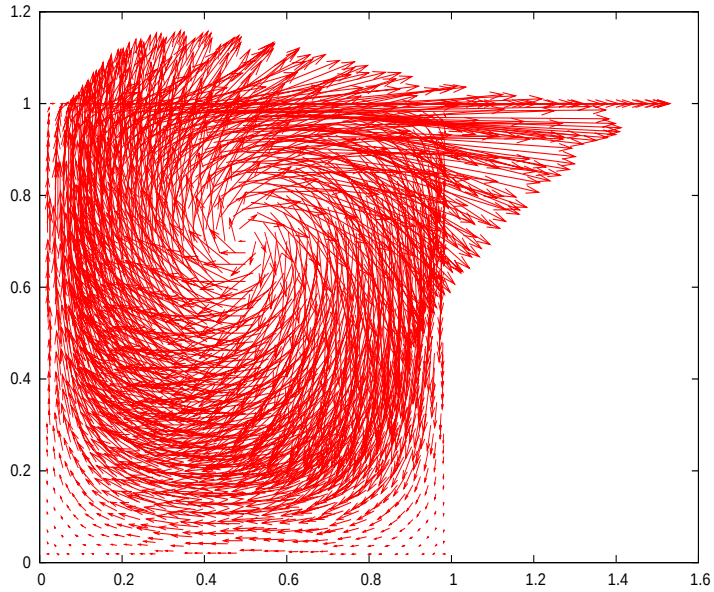

Figure 5.8: Streamlines of the flow for the analytical velocity $\underline{u}$ (not scaled) with $R_e = 1$.

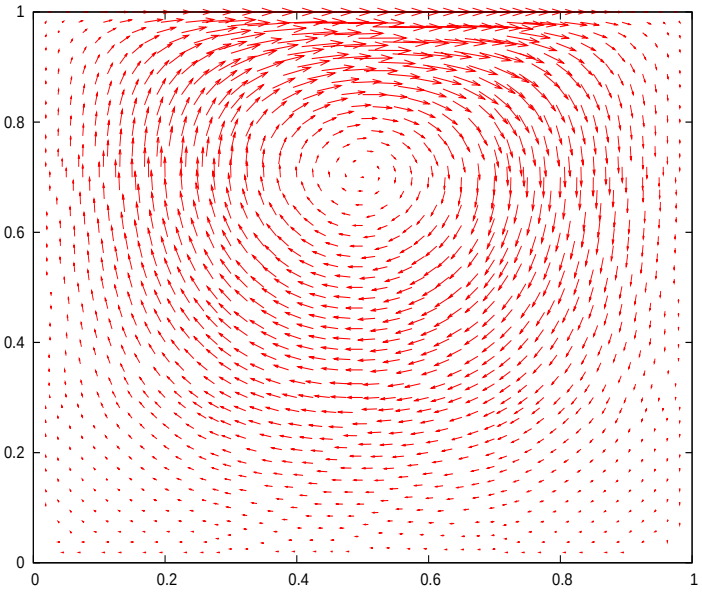

Figure 5.9: Streamlines of the flow for the numerical velocity $\underline{u_h}$ (scaled) with $R_e = 1$.

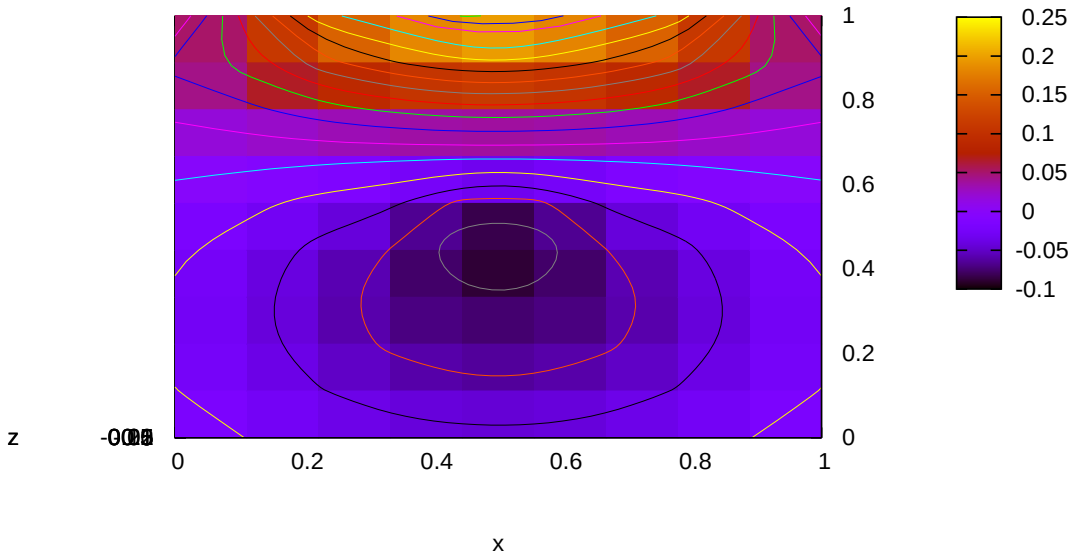

Figure 5.10: Contours of numerical component u_h with $R_e = 1$.

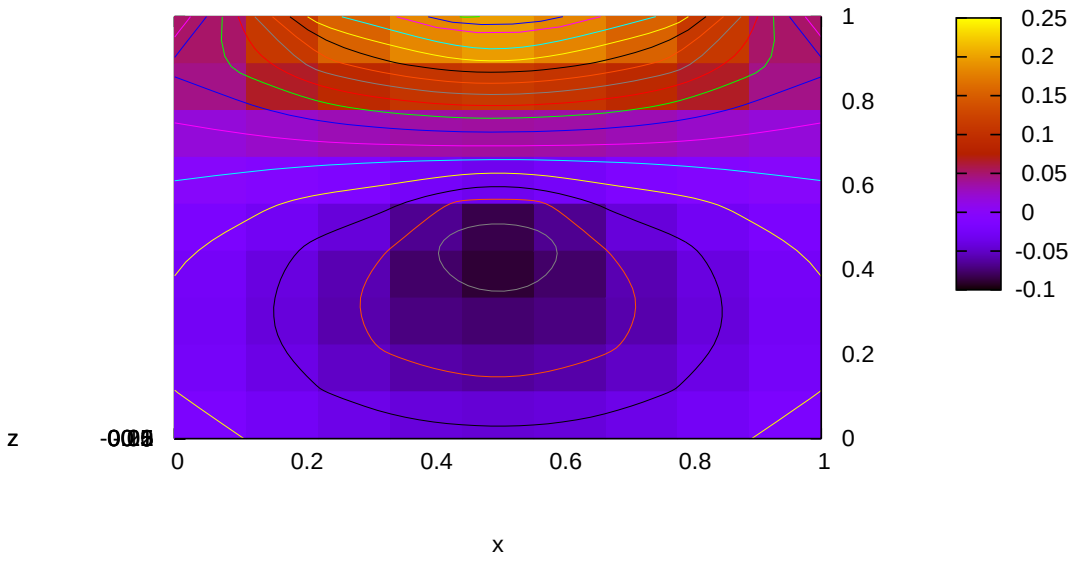

Figure 5.11: Contours of analytical component u with $R_e = 1$.

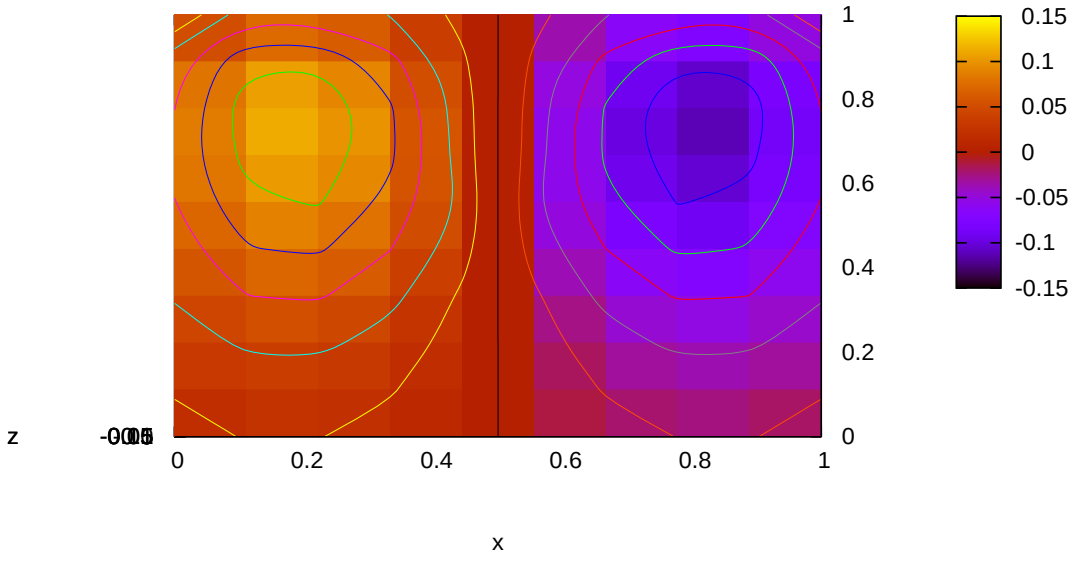

Figure 5.12: Contours of numerical component v_h with $R_e = 1$.

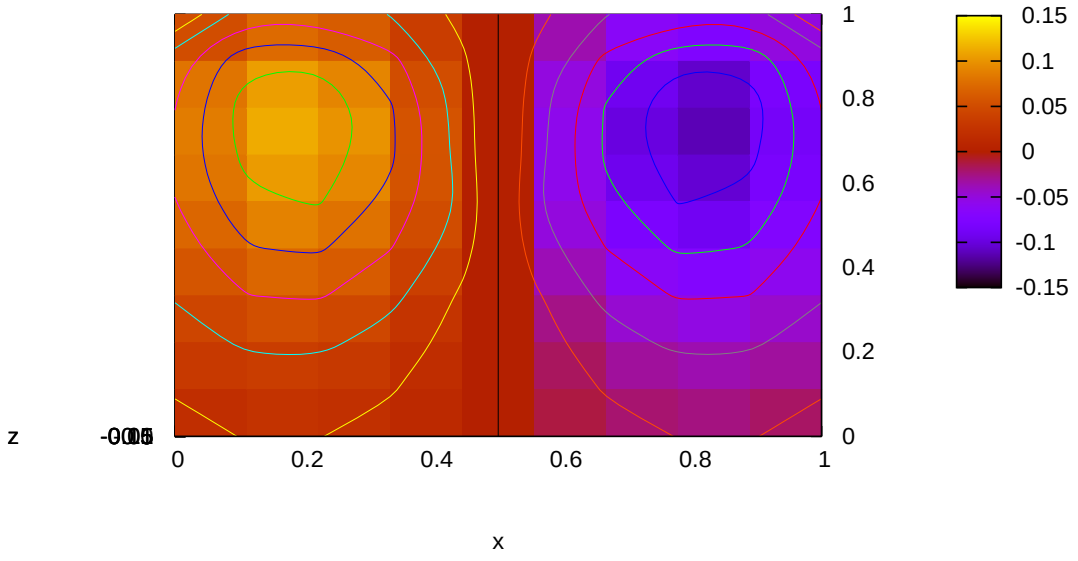

Figure 5.13: Contours of analytical component v with $R_e = 1$.

We reported the performance of the Schwarz overlapping preconditioner in the Table 5.4.

Table 5.4: Condition numbers $\mathcal{K}(\mathcal{A}_h)$ and $\mathcal{K}(\mathrm{P}_{cas}\mathcal{A}_h)$ of the not preconditioned and of the preconditioned matrices, and iteration numbers $\neq Iters$ and $\neq Iters - S$ for the solution of the non preconditioned and preconditioned systems for $R_e = 1$.

Matrix of	$\mathcal{K}(\mathcal{A}_h)$	$\mathcal{K}(\mathrm{P}_{\mathrm{cas}}\mathcal{A}_h)$	$\neq Iters$	$\neq Iters - S$
$\tilde{u}^{n+\frac{1}{2}}$, $\tilde{v}^{n+\frac{1}{2}}$	$8.796E+1$	$2.213E+1$	63	9
p^{n+1}	$5.687E+8$	$4.011E+4$	246	23
u^{n+1}, v^{n+1}	$6.238E+0$	$3.623E+0$	16	5

Table 5.5: $\mathrm{L}^\infty-$ error norm of the numerical solution for $R_e = 10$.

Method	e_u	e_v	e_p
New method	$4.840E-4$	$6.120E-4$	$8.363E-2$
Shih et al. 4/1 staggered [128]	$1.391E-3$	$1.788E-3$	$2.072E-2$
Shih et al. 5/4 staggered [128]	$9.050E-4$	$8.510E-4$	$2.859E-1$

Results second case : $R_e = 10$

For this value of Reynolds number, in Table 5.5 we report the $\mathrm{L}^\infty-$ error norm of the numerical solution at the final instant $t = 1$ and contemporary the error norm of the solution of Shih et al.[128] with two different approximations.

In the Figures 5.14 and 5.15 the streamlines of the numerical and analytical velocity (not scaled), in Figure 5.16 the streamlines of the numerical velocity (scaled), in Figures 5.17 - 5.20 the contours of velocity components (numerical and analytical) for $R_e = 10$ and instant $t = 1$.

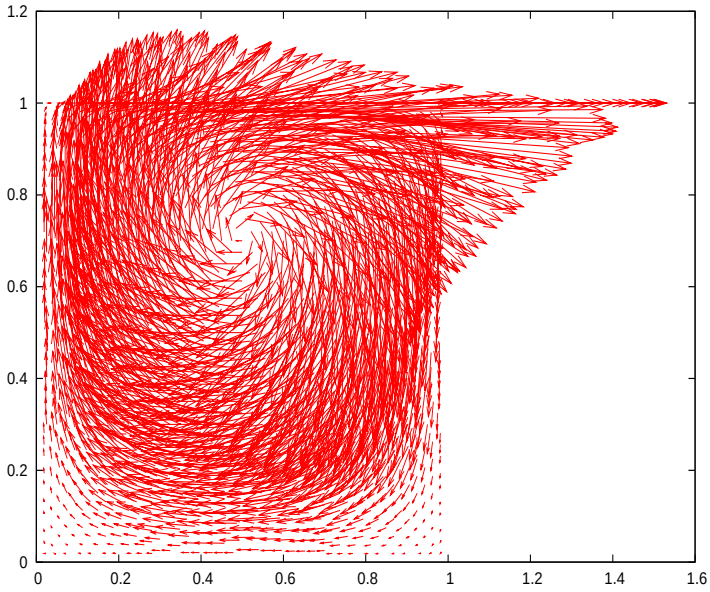

Figure 5.14: Streamlines of the flow for the numerical velocity $\underline{u}_h$ (not scaled) for $R_e = 10$.

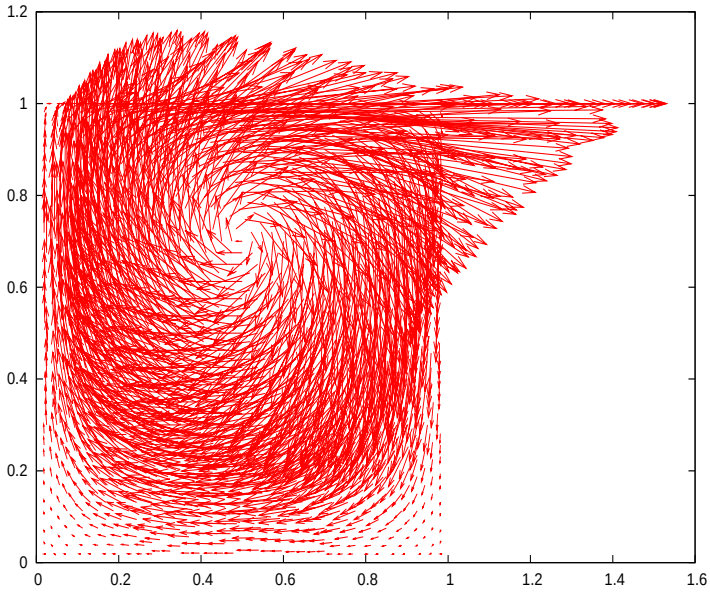

Figure 5.15: Streamlines of the flow of the analytical velocity $\underline{u}$ (not scaled) for $R_e = 10$.

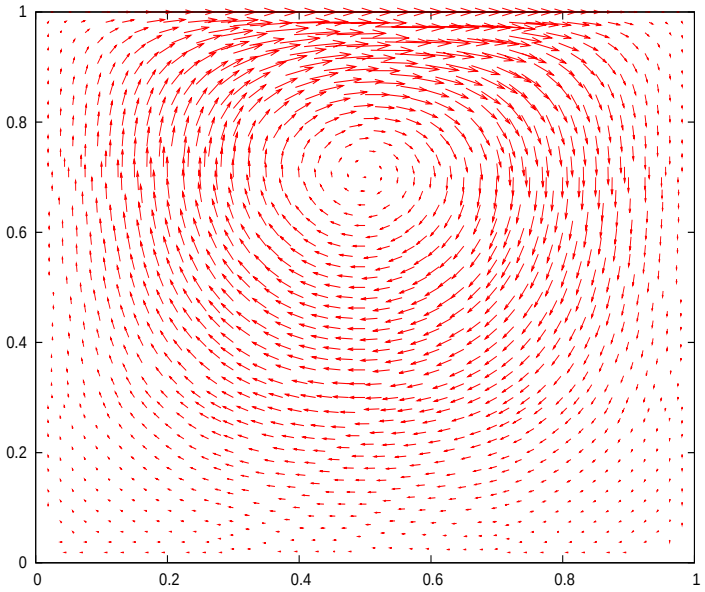

Figure 5.16: Streamlines of the flow for the numerical velocity $\underline{u}_h$ (scaled) with $R_e = 10$.

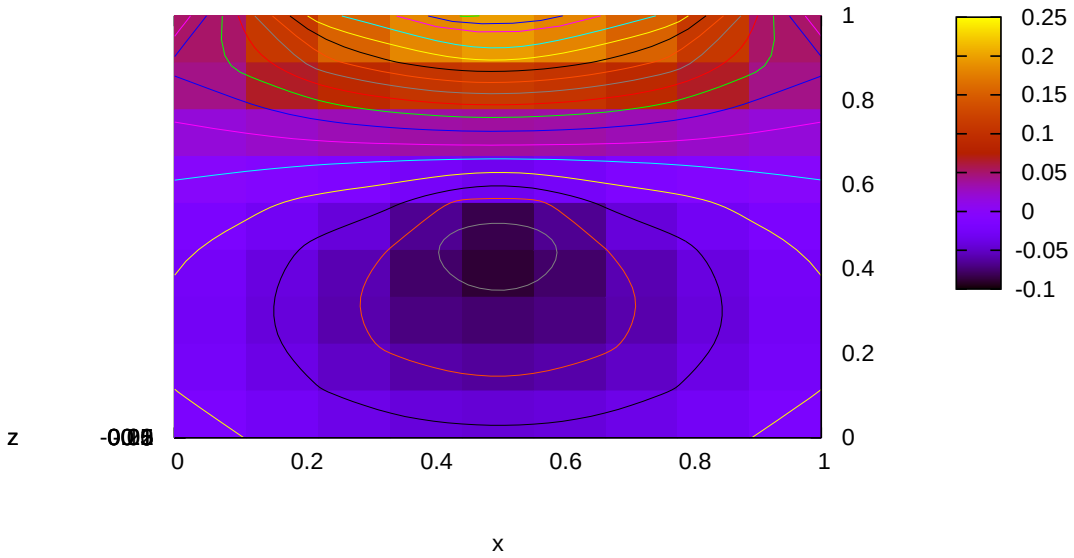

Figure 5.17: Contours of the numerical component u_h with $R_e = 10$.

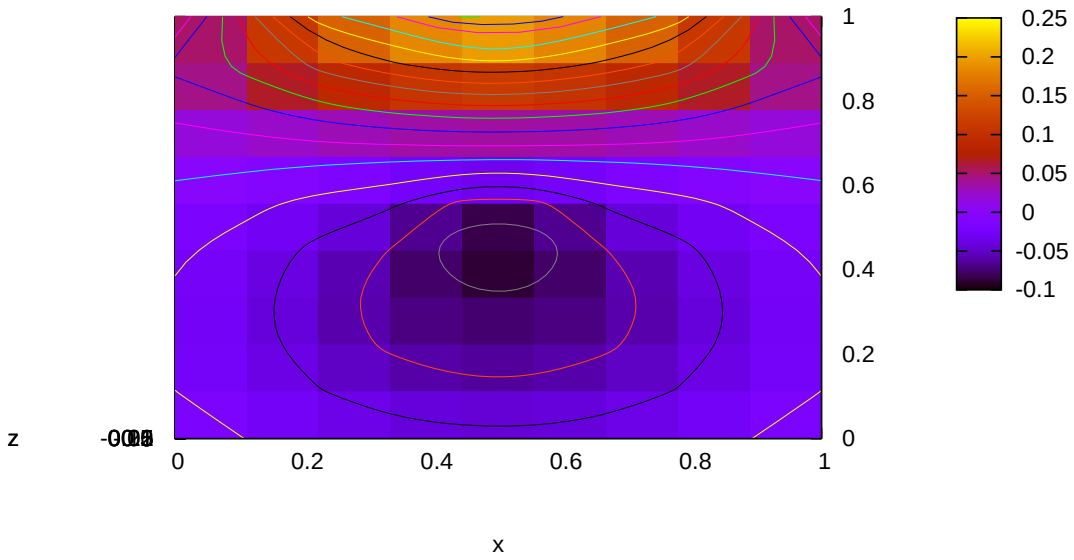

Figure 5.18: Contours of the analytical component u with $R_e = 10$.

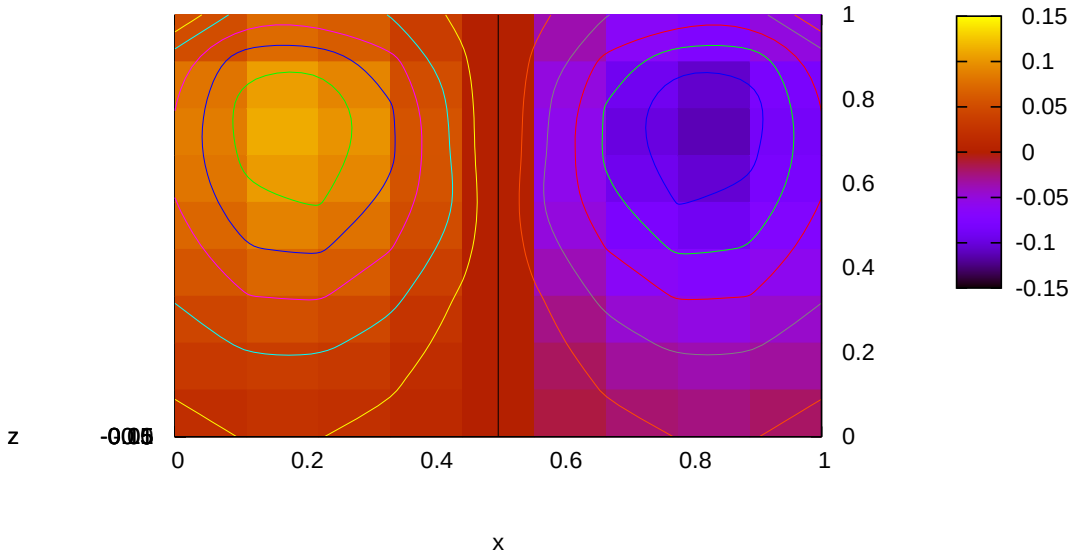

Figure 5.19: Contours of the numerical component v_h with $R_e = 10$.

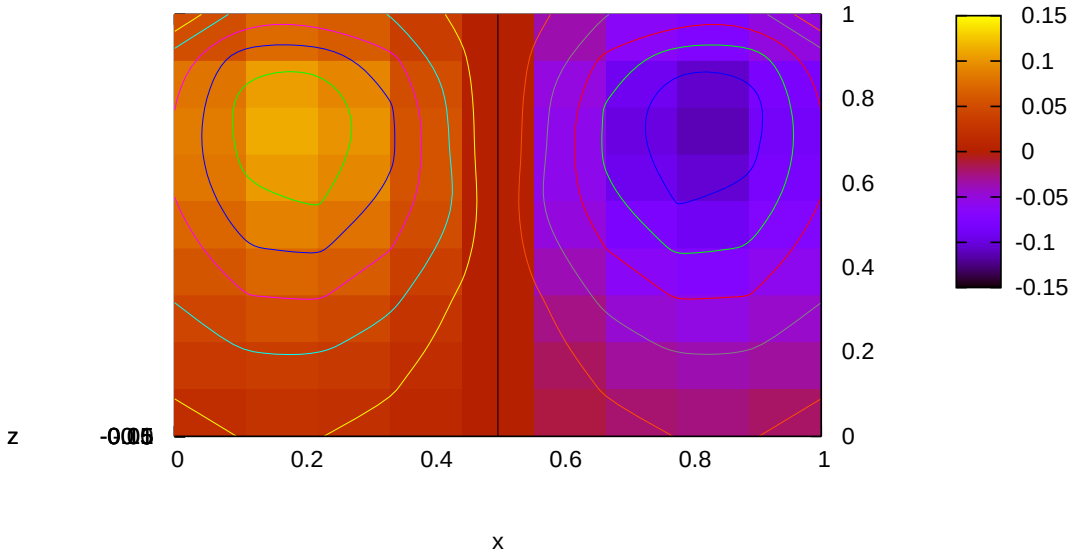

Figure 5.20: Contours of the analytical component v with $R_e = 10$.

We reported the performance of the Schwarz overlapping preconditioner in for $R_e = 10$ in the Table 5.6.

Results third case : $R_e = 100$

For this value of Reynolds number, we report the value of the $L^\infty-$ error norm of the numerical solution in the Table 5.7.

As in the previous case, we have in Figures 5.21 and 5.22 the plot of the streamlines of the numerical and analytical velocity (not scaled), in Figure 5.23 the streamline of the numerical velocity (scaled), in Figure 5.24 and Figure 5.25 - 5.27 the contours of velocity components (numerical and analytical) for $R_e = 100$ at time instant $t = 1$.

Table 5.6: Condition numbers $\mathcal{K}(\mathcal{A}_h)$ and $\mathcal{K}(\mathrm{P}_{cas}\mathcal{A}_h)$ of the not preconditioned and preconditioned matrices and iteration numbers $\neq Iters$ and $\neq Iters - S$ for the solution of the non preconditioned and preconditioned systems for $R_e = 10$.

Matrix of	$\mathcal{K}(\mathcal{A}_h)$	$\mathcal{K}(\mathrm{P}_{\mathrm{cas}}\mathcal{A}_h)$	$\neq Iters$	$\neq Iters - S$
$\tilde{u}^{n+\frac{1}{2}}$, $\tilde{v}^{n+\frac{1}{2}}$	$1.306E+1$	$4.536E+0$	24	5
p^{n+1}	$5.687E+8$	$4.011E+4$	264	23
u^{n+1}, v^{n+1}	$6.238E+0$	$3.623E+0$	16	5

Table 5.7: $\mathrm{L}^\infty-$ error norm of the numerical solution for $R_e = 100$.

Method	e_u	e_v	e_p
New method	$3.087E-3$	$1.941E-3$	$4.000E-2$

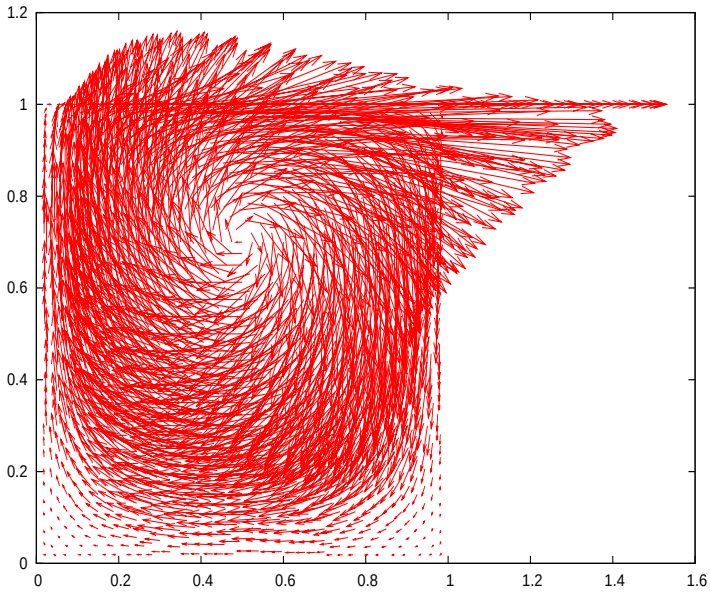

Figure 5.21: Streamlines of the flow for the numerical velocity $\underline{u}_h$ (not scaled) for $R_e = 100$.

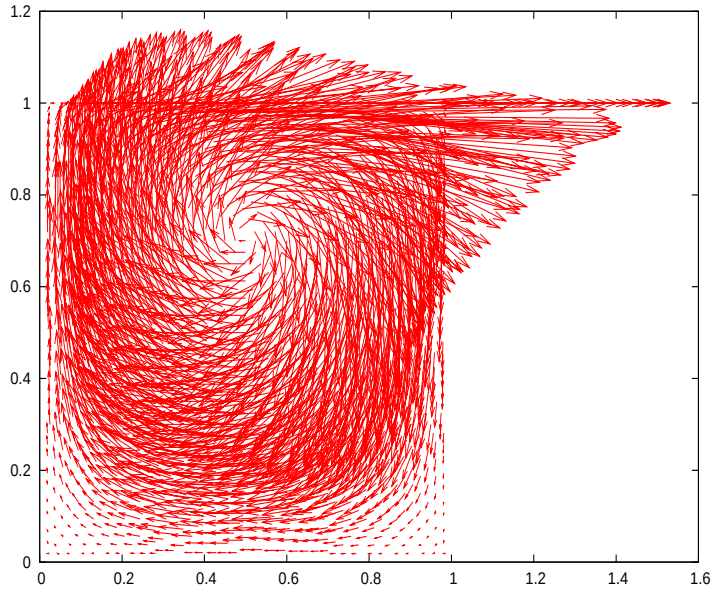

Figure 5.22: Streamlines of the flow of the analytical velocity $\underline{u}$ (not scaled) for $R_e = 100$.

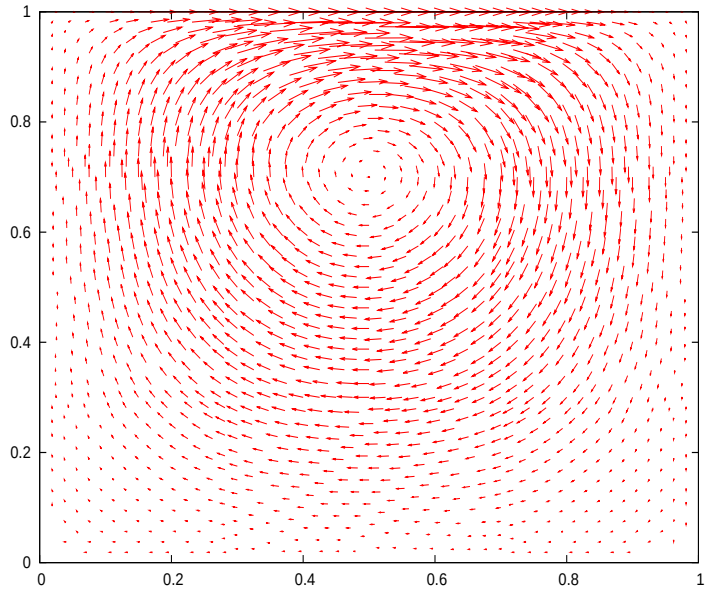

Figure 5.23: Streamlines of the flow for the numerical velocity $\underline{u_h}$ (scaled) with $R_e = 100$.

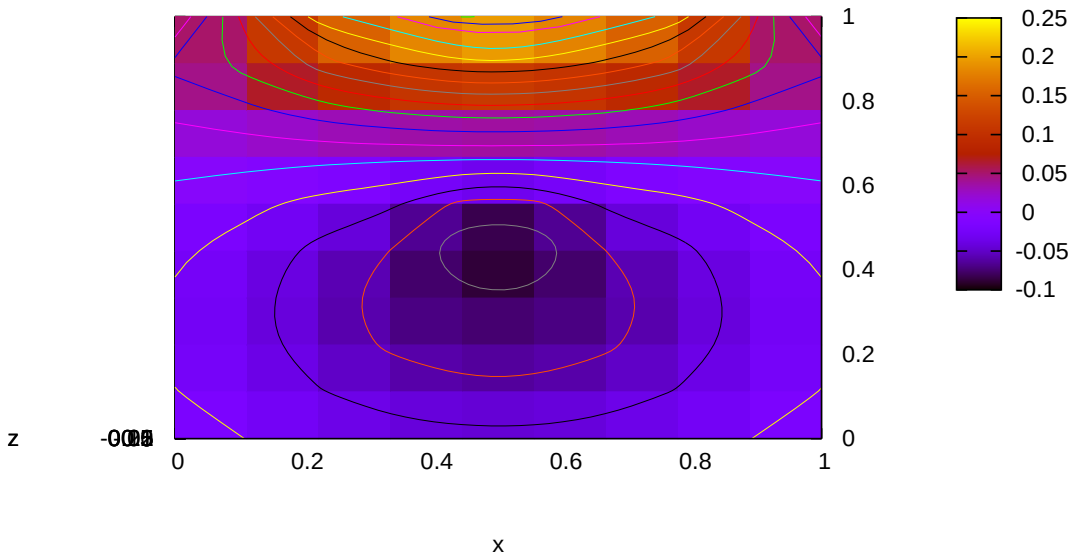

Figure 5.24: Contours of the numerical component u_h for $R_e = 100$.

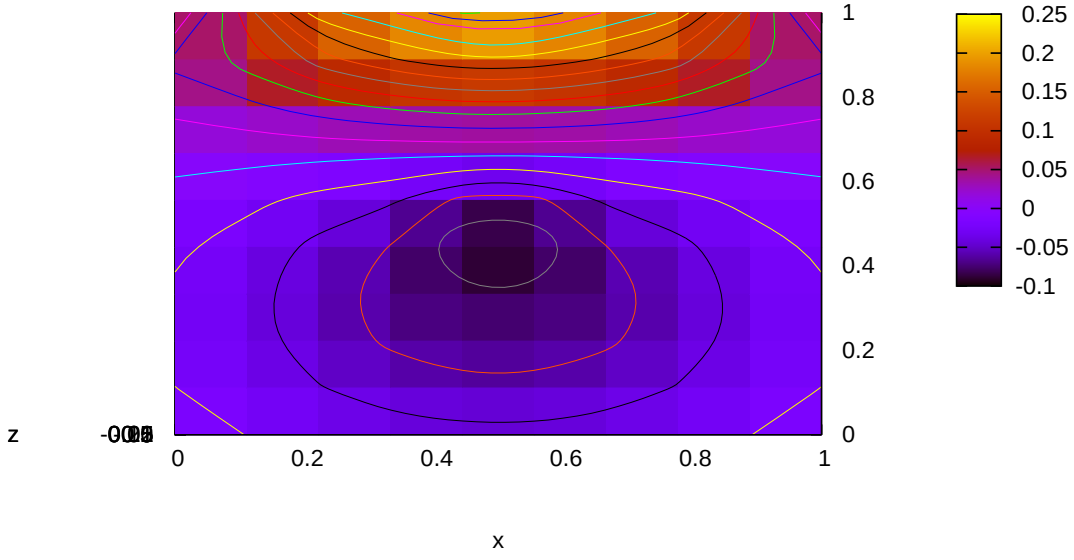

Figure 5.25: Contours of the analytical component u for $R_e = 100$.

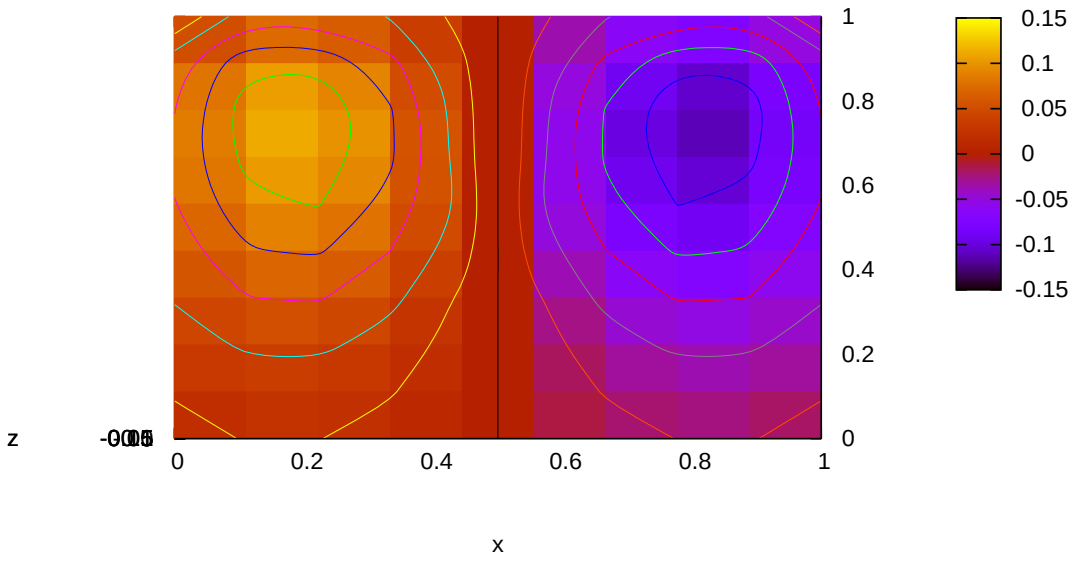

Figure 5.26: Contours of the numerical component v_h for $R_e = 100$.

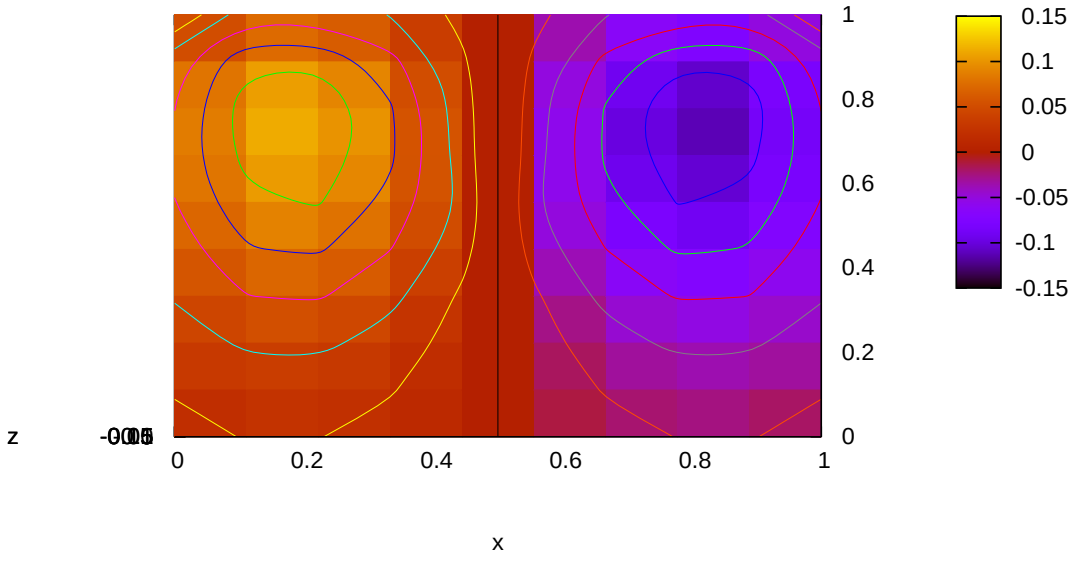

Figure 5.27: Contours of the analytical component v for $R_e = 100$.

Table 5.8: Condition numbers $\mathcal{K}(\mathcal{A}_h)$ and $\mathcal{K}(\mathrm{P}_{cas}\mathcal{A}_h)$ of the not preconditioned and preconditioned matrices and iteration numbers $\neq Iters$ and $\neq Iters - S$ for the solution of the non preconditioned and preconditioned systems for $R_e = 100$.

Matrix of	$\mathcal{K}(\mathcal{A}_h)$	$\mathcal{K}(\mathrm{P}_{\mathrm{cas}}\mathcal{A}_h)$	$\neq Iters$	$\neq Iters - S$
$\tilde{u}^{n+\frac{1}{2}}$, $\tilde{v}^{n+\frac{1}{2}}$	$5.945E+0$	$3.576E+0$	12	4
p^{n+1}	$5.687E+8$	$4.011E+4$	259	23
u^{n+1}, v^{n+1}	$6.238E+0$	$3.623E+0$	16	5

We also report for this choice of Reynolds number the performance of the Schwarz preconditioner in Table 5.8.

Comments

From the Figures 5.7 - 5.27 we can check that in all cases, the numerical solutions are in good harmony with the analytical ones, moreover it can be observed that the clockwise circulation is very similar to the classical lid-driven recirculating flow.

5.5 Numerical tests: a real application

In this section, we consider a benchmark flow problem where no analytic solution is known, but considered very important by the researchers. The aim is to compare our results to some of the well established schemes present in the literature for a natural convection in a square cavity problem.

5.5.1 Natural Convection in a square cavity

De Valh Davis provides in [37] the definition of a large number of test cases involving a two dimensional natural convection in a square enclosure along with values for some reference quantities.

Problem setting

The flow and the boundary conditions for this problem are shown in Figure 5.29. Figure 5.28 shows a 31×31 nonuniform mesh which has been used for all the case considered. We wish to examine the flow of a fluid inside a square cavity for which the top and the bottom walls are kept to be adiabatic and the verticals walls are kept to be isothermal at temperatures $\mathtt{T}_c = -0.5$ and $\mathtt{T}_h = 0.5$ respectively.
Initially the fluid is assumed to be at rest at temperature $\mathtt{T} = 0$; then subsequently the temperature at the vertical walls begins to change and the fluid is subjected to a phenomenon of convection due to the thermal gradient. We also assumed that the fluid is incompressible such that the Boussinesq approximation holds, that is

$$\underline{f} = \rho \underline{g} \alpha (\mathtt{T} - \mathtt{T}_r) \quad . \tag{5.46}$$

where α is the thermal expansion coefficient.

The flow is governed by the Rayleigh number defined as $R_a = \frac{\rho g \alpha L^3 c_p \Delta \mathtt{T}}{K\nu}$ and the Prandtl

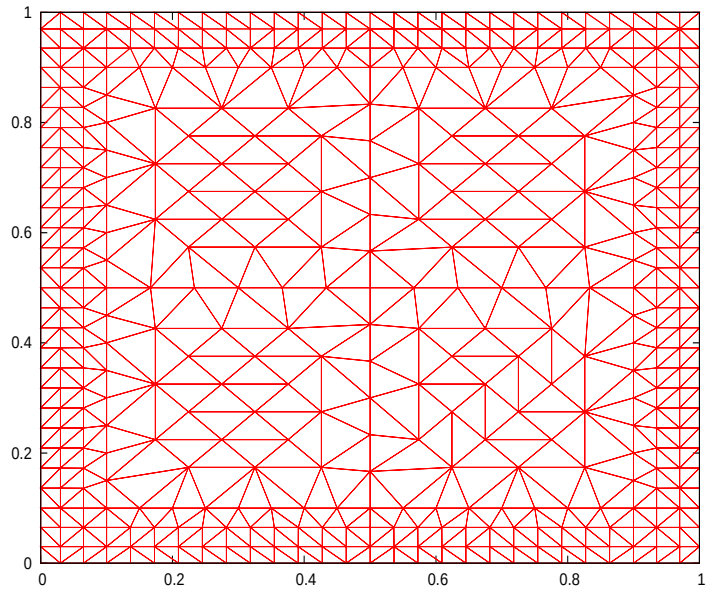

Figure 5.28: Natural convection in a square enclosure: 31×31 non uniform mesh.

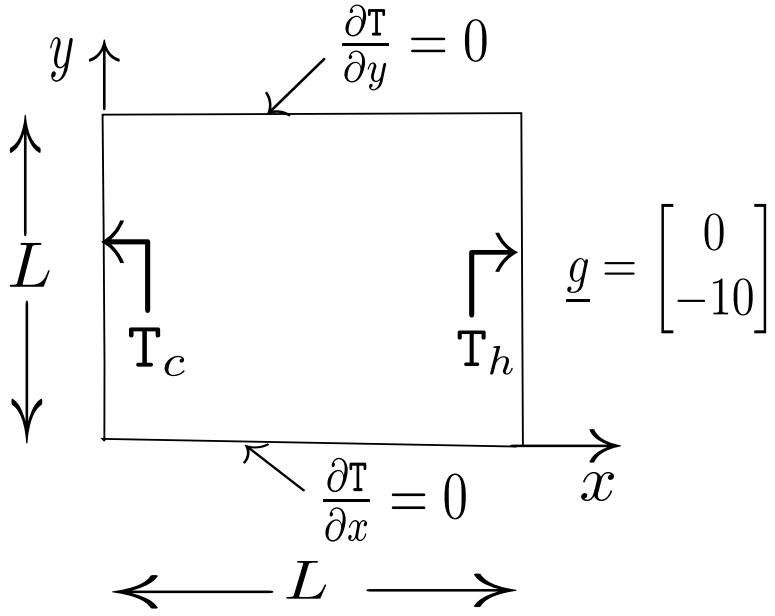

Figure 5.29: Natural convection in a square enclosure: problem setting.

number $P_r = \frac{\mu}{\rho\, k}$ that takes into account the characteristic of the fluid where

$$
\begin{aligned}
\rho &:= \quad \text{density} \\
\underline{g} &:= \quad \text{gravity} \\
\alpha &:= \quad \text{coefficient of the thermal expansion} \\
L &:= \quad \text{length of the cavity} \\
\Delta \mathtt{T} := \mathtt{T}_h - \mathtt{T}_c &\quad \text{variation of the temperature} \\
k &:= \quad \text{coefficient of thermal diffusivity} \\
\nu &:= \quad \text{cinematic viscosity} \\
\mathtt{T}_r &:= \quad \text{reference temperature} \\
K &:= \quad \text{thermal conductivity} \\
c_p &:= \quad \text{specific heat at constant pressure} \ .
\end{aligned}
$$

Table 5.9: Natural convection in a square enclosure: solutions generated by new method for $\tau = 10^{-5}$, by De Vahl Davis [37], by Choi [26] and by Hookey [60] .

Solution	Results	$R_a = 10^3$	$R_a = 10^4$
New method (31×31 mesh grid)	u_{max}	3.614	16.243
	v_{max}	3.670	19.231
De Vahl Davis [37]	u_{max}	3.649	16.178
	v_{max}	3.697	19.617
Hookey and Baliga [60] (31×31 mesh grid)	u_{max}	3.632	16.203
	v_{max}	3.678	19.471
Choi et al. [26] (31×31 mesh grid)	u_{max}	3.644	16.418
	v_{max}	3.726	19.801

Computations have been carried out for $R_a = 10^3$ and 10^4, for the weight parameter $\tau = 10^{-5}$ and $P_r = 0.71$.
A trivial solution ($u = v = p = \mathtt{T} = 0$) has been used as initial guess for $R_a = 10^3$ and $R_a = 10^4$.

Results and comments

The u_{max} and v_{max} data from the benchmark solution of De Vahl Davis [37], Hookey and Baliga [60] and Choi et al. [26] are given in Table 5.9. The corresponding results produced by the new method at the final time ($t = 1$) for the 31×31 mesh grid are also given in Table 5.9. Then follows the graphics obtained.

We note that Davis [37] performed numerical simulations on uniform meshes from 11×11 to 41×41 for $R_a = 10^3$ and $R_a = 10^4$.

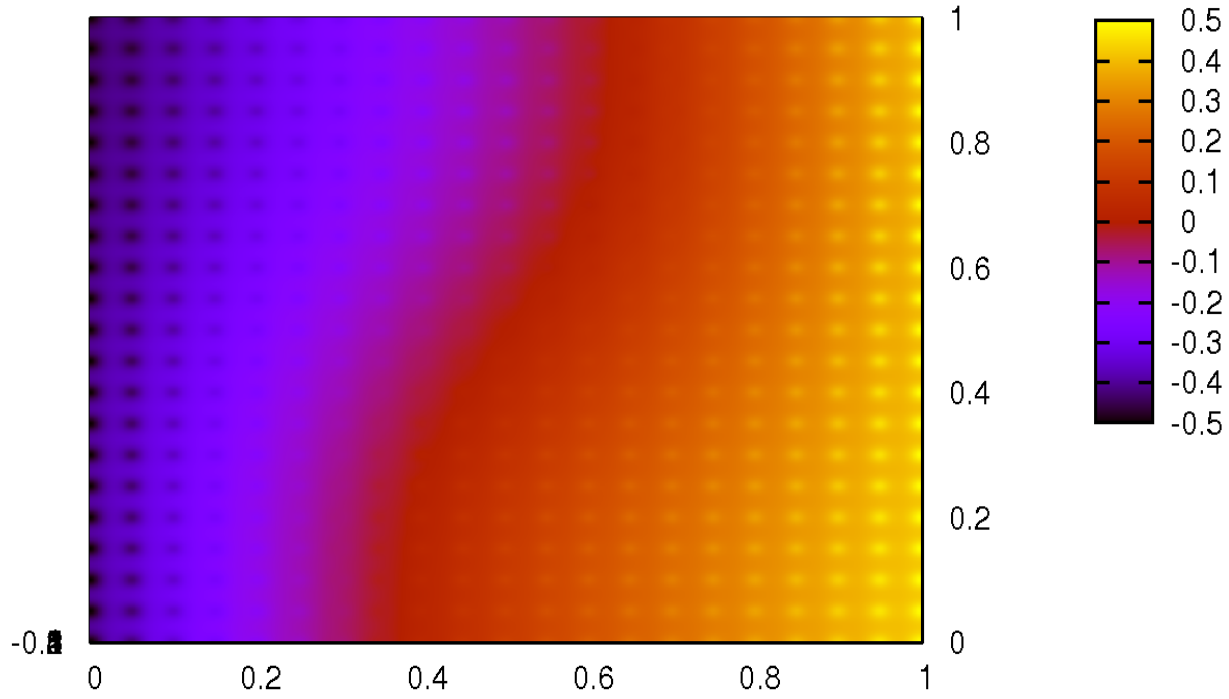

Figure 5.30: Distribution of the temperature for $R_a = 10^3$.

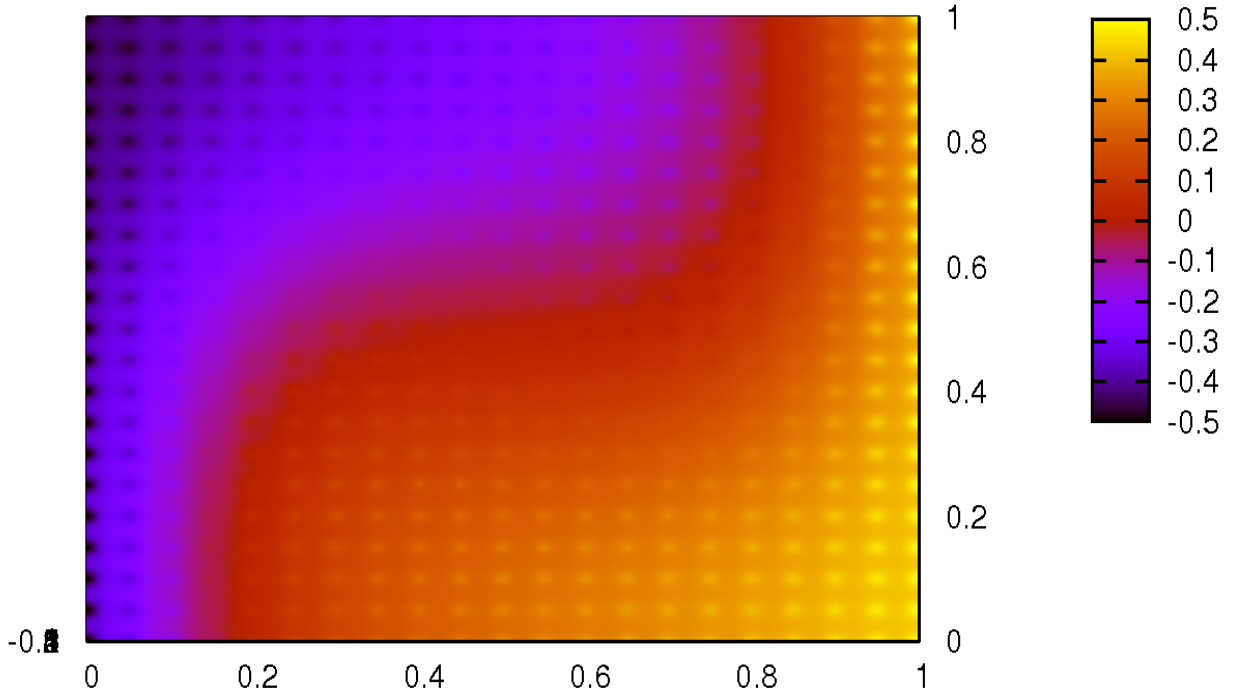

Figure 5.31: Distribution of the temperature for $R_a = 10^4$.

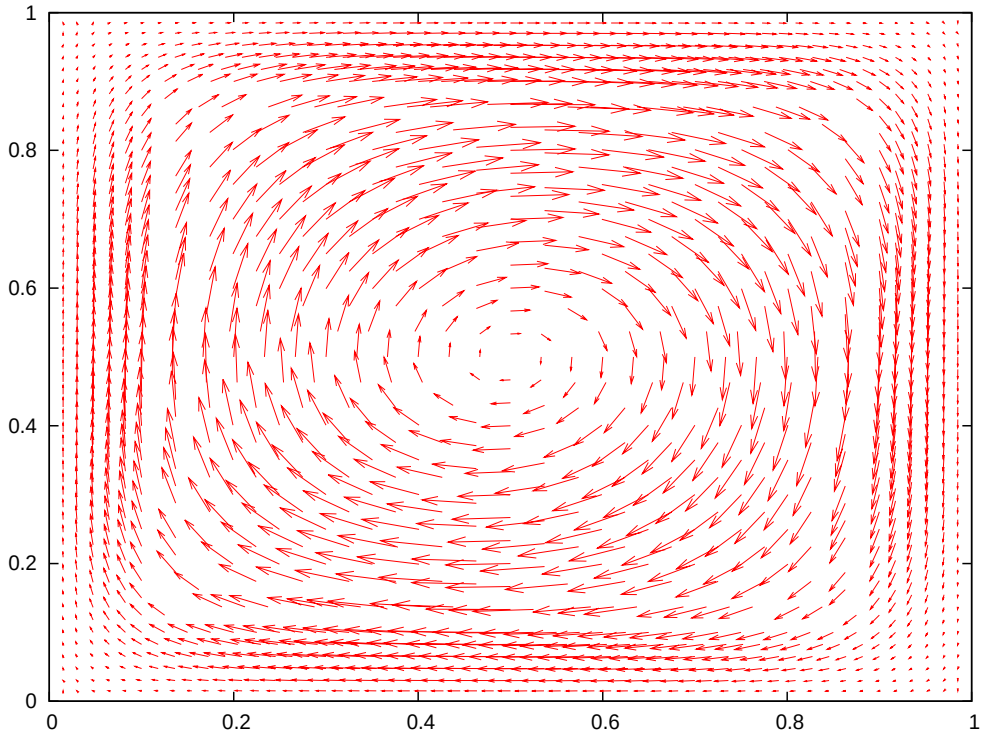

Figure 5.32: Streamlines of the flow (scaled) for $R_a = 10^3$.

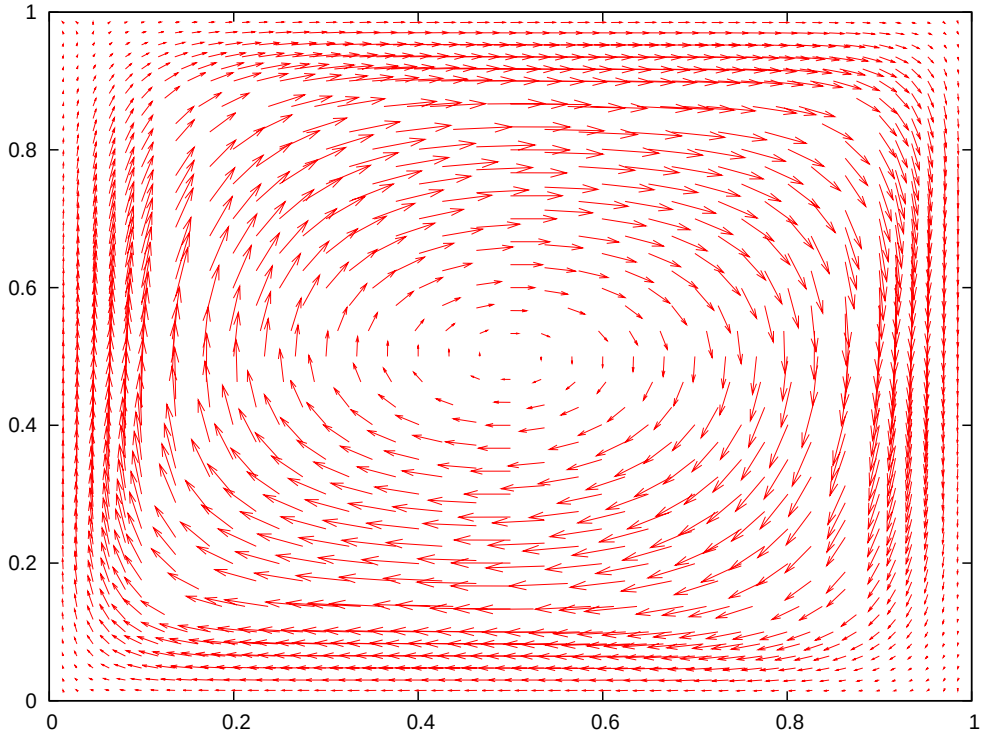

Figure 5.33: Streamlines of the flow (scaled) for $R_a = 10^4$.

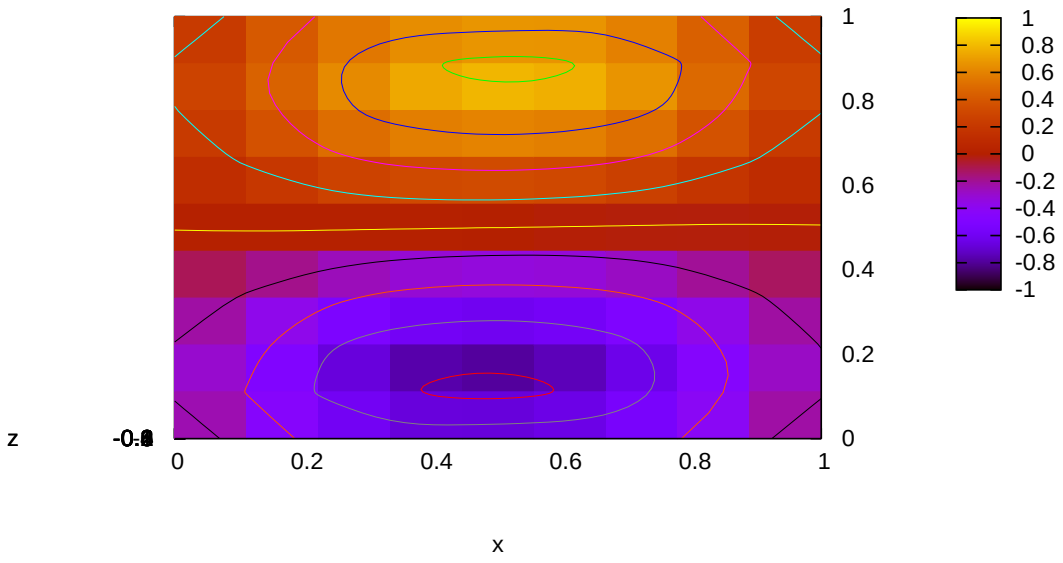

Figure 5.34: Contours of the numerical component u_h for $R_a = 10^3$.

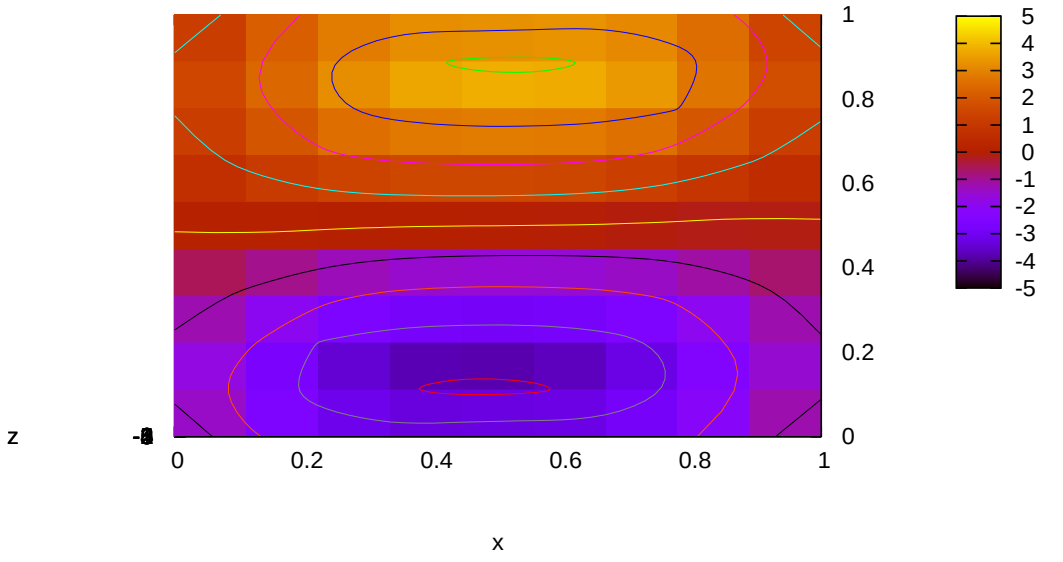

Figure 5.35: Contours of the numerical component u_h for $R_a = 10^4$.

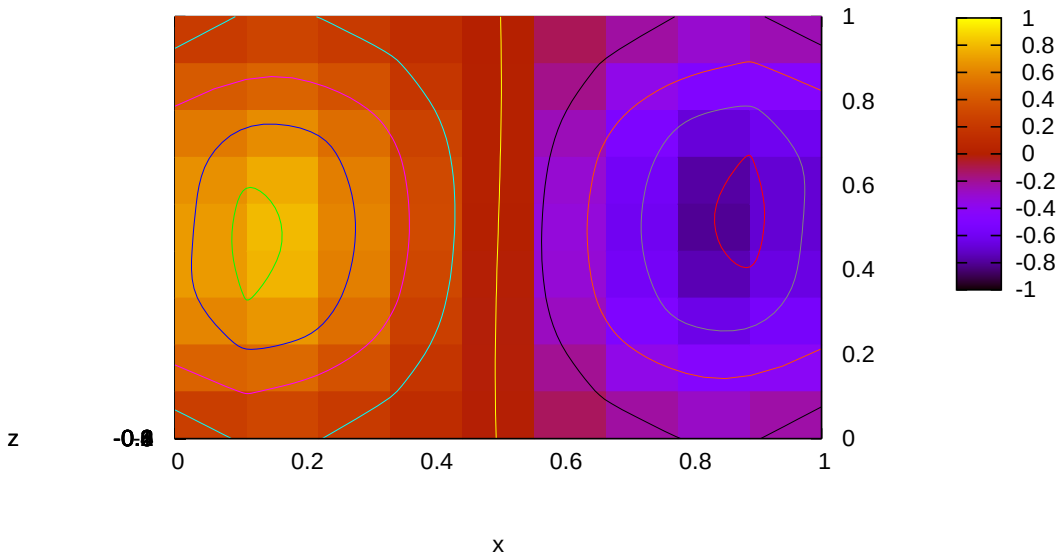

Figure 5.36: Contours of the numerical component v_h for $R_a = 10^3$.

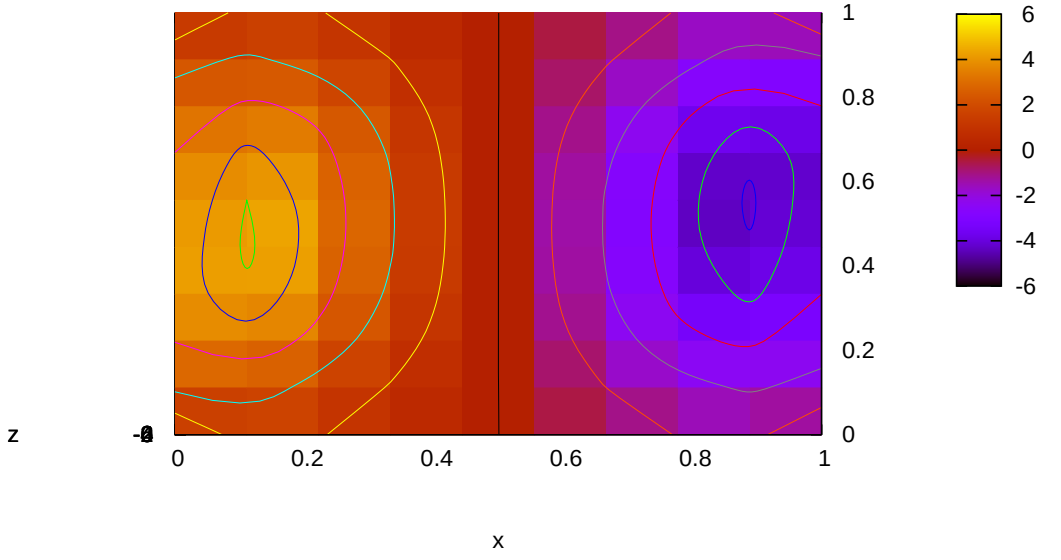

Figure 5.37: Contours of the numerical component v_h for $R_a = 10^4$.

Table 5.10: Condition numbers $\mathcal{K}(\mathcal{A}_h)$ and $\mathcal{K}(\mathrm{P}_{as}\mathcal{A}_h)$ of the not preconditioned and of the preconditioned matrices, and iteration numbers $\neq Iters$ and $\neq Iters - S$ for the solution of the non preconditioned and of the preconditioned systems for $R_a = 10^3$ and $R_a = 10^4$.

Rayleigh number	Matrix of	$\mathcal{K}(\mathcal{A}_h)$	$\mathcal{K}(\mathrm{P}_{\mathrm{as}}\mathcal{A}_h)$	$\neq Iters$	$\neq Iters - S$
10^3	$\tilde{u}^{n+\frac{1}{2}}$, $\tilde{v}^{n+\frac{1}{2}}$	$1.408E+2$	$6.446E+0$	61	6
10^3	p^{n+1}	$1.891E+8$	$3.889E+4$	270	22
10^3	u^{n+1}, v^{n+1}	$2.903E+1$	$3.880E+0$	16	5
10^3	$\mathtt{T}^{n+1}$	$1.909E+2$	$8.638E+0$	86	8
10^4	$\tilde{u}^{n+\frac{1}{2}}$, $\tilde{v}^{n+\frac{1}{2}}$	$1.408E+2$	$6.446E+0$	59	6
10^4	p^{n+1}	$1.891E+8$	$3.889E+4$	273	21
10^4	u^{n+1}, v^{n+1}	$2.903E+1$	$3.880E+0$	16	5
10^4	$\mathtt{T}^{n+1}$	$1.909E+2$	$8.638E+0$	95	8

The new method provides good results for the simulations developed and they are in good agreement with those of the benchmark solutions of De Vahl Davis [37].
In Table 5.10 is reported the performance of the Schwarz overlapping preconditioner relevant to this benchmark problem. As for this particular test we are using also the temperature equation, the performance of Schwarz with respect to this equation are therefore included. All the value where collected at the final time instant ($t = 1$) and we see that the Schwarz overlapping preconditioner works well also for this benchmark problem.

5.6 Conclusions of the chapter

The results of the tests confirm that the new numerical method for solving the 2D Navier-Stokes equations is efficient and accurate like expected (first order in time and second order in space, both for velocity, pressure and temperature). The new technique for advancing in time, based on a fractional step method and characteristics reduces the most expensive computational kernels to the solution of algebraic systems stemming from elliptic problems. In order to reduce as most as possible the computational effort, an iterative method (Bi-CGSTAB), preconditioned by an additive Schwarz preconditioner has been used. An well-established h-adaptive techniques based on the error estimation of the residual could be advantageously used [11].

Chapter 6

Solution of Shallow-Water equations

In this chapter, we present a new numerical method for the solution of one dimensional Shallow-Water problem based on the fractional step scheme seen in the previuos chapter and using P2-P1 finite elements for the spatial approximation.

6.1 An environmental problem

It is well known that under the hypothesis of hydrostatic pressure, it is possible to derive from 3D Navier-Stokes equations a system of partial differential equations named Shallow Water equations (SWE) in which the primitive unknowns are the 2D average velocity components $\underline{u} = (\overline{u}, \overline{v})$ and the elevation ξ. This last unknown represents the variation of the free surface in respect to a reference level (see Figure 6.1).

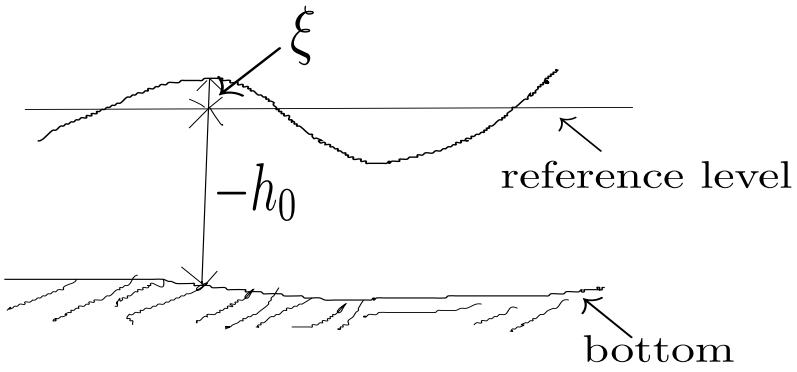

Figure 6.1: Elevation ξ.

The field of application of the SWE is very large; in particular we can recall the study of river and channel flows, the study of tidal water systems, the application of flood waves subsequent to the break down of dams or of bank rivers.
In this section we will present a classical formulation of the 1D SWE aimed to the description of the variation of the concentration relevant to chemical substances transported by the flow of a river.

6.2 The mathematical model

The system of partial differential equations we want to solve in conservative form, is:

$$\begin{cases} \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}(\dfrac{qq}{h}) - \mu\dfrac{\partial^2 q}{\partial x^2} + gh\dfrac{\partial \xi}{\partial x} = f \\ \dfrac{\partial \xi}{\partial t} + \dfrac{\partial q}{\partial x} = 0 \\ \dfrac{\partial c_1}{\partial t} + u\dfrac{\partial c_1}{\partial x} - \Gamma_1\dfrac{\partial^2 c_1}{\partial x^2} + f_1 c_1 = s_1 \\ \dfrac{\partial c_2}{\partial t} + u\dfrac{\partial c_2}{\partial x} - \Gamma_2\dfrac{\partial^2 c_2}{\partial x^2} + f_2 c_2 = s_2 \end{cases} \tag{6.1}$$

where $q \equiv uh$ is the unit discharge; $h = \xi + h_0$ is the total depth water; μ is the dispersion coefficient; g is the gravity; c_1 and c_2 are the concentrations of the chemical substances; Γ_1 and Γ_2 are the diffusion coefficients; f_1 and f_2 are the reactivity laws.

6.3 The numerical model

In the past thirty years, a large amount of papers and books have been written regarding the solution of the SWE; in them, the spatial approximation have been afforded both by FD or FV or FE methods while the advancing in time both by FD schemes or fractional step schemes [2]. Here, we decided to adopt for advancing in time a fractional step method very similar to that we used for solving the Navier-Stokes equations in the previous chapter. The very important idea behind this approach is to split the equations in order to decouple the various physical contributions [8]. For the spatial approximation we decided in order to guarantee the stability to use P1 elements for elevation ξ and P2 elements for unit discharge q and for c_1 and c_2. Actually, at our knowledge, for the SWE does not exist a condition equivalent to the Inf-Sup condition for the Navier- Stokes equations, however numerical experiences carried out in the past showed the presences of instabilities if the polynomial spaces for ξ and q are chosen equal [134]. In detail, the fractional step we used for advancing from instant t^n to instant t^{n+1} is:

- $$u^n = \frac{q^n}{h^n}\ , h^n = h_0 + \xi^n$$

- $$u^{n+\frac{1}{3}} = u^n(\tilde{x}_I) \text{ with } \tilde{x}_I = x - \Delta t u^n(\bar{x}_I) \text{ and } \bar{x}_I = x - \frac{\Delta t}{2}u^n(x),\ q^{n+\frac{1}{3}} = h^n u^{n+\frac{1}{3}}$$

  ($\tilde{x}_I$ is the foot of characteristic relevant to the node with abscissa x)

- $$q^{n+\frac{2}{3}} = q^{n+\frac{1}{3}} + \Delta t\mu\frac{\partial^2 q^{n+\frac{1}{3}}}{\partial x^2} + \Delta t f^{n+\frac{2}{3}}$$

- $$\xi^{n+1} - (\Delta t)^2\frac{\partial}{\partial x}(gh^n\frac{\partial \xi^{n+1}}{\partial x}) + \Delta t\frac{\partial}{\partial x}(\frac{q^{n+\frac{2}{3}}}{h^n}\xi^{n+1}) = \xi^n - \Delta t\frac{\partial}{\partial x}q^{n+\frac{2}{3}} + \Delta t\frac{\partial}{\partial x}(\frac{q^{n+\frac{2}{3}}}{h^n}\xi^n)$$

  this equation for the adjourned value of ξ was derived applying the derivative $\frac{\partial}{\partial x}$ operator to the equation

$$q^{n+1} - q^{n+\frac{2}{3}} + \Delta t g h^n\frac{\partial \xi^{n+1}}{\partial x} - \frac{q^{n+\frac{2}{3}}}{h^n}(\xi^{n+1} - \xi^n) = 0$$

and subtracting the result to the equation

$$\xi^{n+1} - \xi^n + \Delta t \frac{\partial q^{n+1}}{\partial x} = 0$$

- $$q^{n+1} = q^{n+\frac{2}{3}} - \Delta t g h^n \frac{\partial \xi^{n+1}}{\partial x} + \frac{q^{n+\frac{2}{3}}}{h^n}(\xi^{n+1} - \xi^n)$$
- $$u^{n+1} = \frac{q^{n+1}}{h^{n+1}}, \;\; h^{n+1} = h_0 + \xi^{n+1}$$
- $$\tilde{x}_{II} = x - \Delta t u^{n+1}(\bar{x}_{II}) \text{ and } \bar{x}_{II} = x - \frac{\Delta t}{2} u^{n+1}(x)$$

  (by $\tilde{x}_{II}$ we indicate the adjourned value of the characteristic foots)
- $$\frac{c_1^{n+1} - c_1^n(\tilde{x}_{II})}{\Delta t} - \Gamma_1 \frac{\partial^2 c_1^{n+1}}{\partial x^2} + f_1^{n+1} c_1^{n+1} = s_1^{n+1}$$
- $$\frac{c_2^{n+1} - c_2^n(\tilde{x}_{II})}{\Delta t} - \Gamma_2 \frac{\partial^2 c_2^{n+1}}{\partial x^2} + f_2^{n+1} c_2^{n+1} = s_2^{n+1}$$

Remark 6.1. *The use of characteristics for the approximation of the convective terms (both for momentum equation and for the transport equations of chemical substances) requires an interpolation procedure of high order; this is easily obtained because of our $P2$ choice for q, c_1 and c_2 variables (we recall that we use $P1$ elements only for ξ variable).*

6.4 Algebraic formulation

As already said, the spatial approximation is based on the Galerkin FE method; fixing the attention on variables interested we can write:

- For the the provisional discharge $q^{n+\frac{2}{3}}$:

$$\int_\Omega q^{n+\frac{2}{3}} \varphi \, d\Omega = \int_\Omega q^{n+\frac{1}{3}} \varphi \, d\Omega - \Delta t \mu \int_\Omega \frac{\partial q^{n+\frac{1}{3}}}{\partial x} \frac{\partial \varphi}{\partial x} \, d\Omega + \Delta t \int_\Omega f^{n+\frac{2}{3}} \varphi \, d\Omega \tag{6.2}$$

  by which, indicating by $\underline{\mathrm{M}}^q$ the mass matrix, $\underline{\mathrm{A}}^q$ the stiffness matrix, (6.2) becomes:

$$\underline{\mathrm{M}}^q \underline{q}^{n+\frac{2}{3}} = \underline{\mathrm{M}}^q \underline{q}^{n+\frac{1}{3}} - \Delta t \mu \underline{\mathrm{A}}^q \underline{q}^{n+\frac{1}{3}} + \Delta t \underline{\mathrm{M}}^q \underline{f}^{n+\frac{2}{3}} \tag{6.3}$$

- For the elevation ξ^{n+1}:

$$\begin{aligned} \int_\Omega \xi^{n+1} \psi \, d\Omega + \Delta t^2 g h^n \int_\Omega \frac{\partial \xi^{n+1}}{\partial x} \frac{\partial \psi}{\partial x} \, d\Omega + \Delta t \frac{q^{n+\frac{2}{3}}}{h^n} \int_\Omega \frac{\partial \xi^{n+1}}{\partial x} \psi \, d\Omega = \int_\Omega \xi^n \psi \, d\Omega \\ - \Delta t \int_\Omega \frac{\partial q^{n+\frac{2}{3}}}{\partial x} \psi \, d\Omega + \Delta t \frac{q^{n+\frac{2}{3}}}{h^n} \int_\Omega \frac{\partial \xi^n}{\partial x} \psi \, d\Omega \end{aligned} \tag{6.4}$$

  so indicating by $\underline{\mathrm{M}}^\xi$ and $\underline{\mathrm{M}}_L^\xi$ the mass and lumped mass matrices respectively, and by $\underline{\mathrm{A}}^\xi$ the stiffness matrix, by $\underline{\mathrm{B}}^\xi$ the matrix $\int_\Omega \frac{\partial \psi}{\partial x} \psi \, d\Omega$, (6.4) becomes:

$$[\underline{\mathrm{M}}_L^\xi + (\Delta t)^2 g \underline{h}^n \underline{\mathrm{A}}^\xi + \Delta t \frac{\underline{q}^{n+\frac{2}{3}}}{\underline{h}^n} (\underline{\mathrm{B}}^\xi)^T] \underline{\xi}^{n+1} = \underline{\mathrm{M}}_L^\xi \underline{\xi}^n + \Delta t (\frac{\underline{q}^{n+\frac{2}{3}}}{\underline{h}^n} \underline{\mathrm{B}}^\xi)^T \underline{\xi}^n - \Delta t (\underline{\mathrm{B}}^\xi)^T \underline{q}^{n+\frac{2}{3}} \tag{6.5}$$

Remark 6.2. *In this equation, the q values considered are those associated with the boundary nodes of the elements.*

Remark 6.3. *The matrices* $\underline{\mathrm{M}}^{\xi}$, $\underline{\mathrm{A}}^{\xi}$ *and* $\underline{\mathrm{B}}^{\xi}$ *are tridiagonal and given by:*

$$\underline{\mathrm{M}}^{\xi} = \frac{l}{6}\begin{pmatrix} 2 & 1 & & 0 \\ 1 & 4 & 1 & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 2 \end{pmatrix}.$$

$$\underline{\mathrm{A}}^{\xi} = \frac{1}{l}\begin{pmatrix} 1 & -1 & & 0 \\ -1 & 2 & -1 & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 1 \end{pmatrix}.$$

$$\underline{\mathrm{B}}^{\xi} = \frac{1}{2}\begin{pmatrix} -1 & -1 & & 0 \\ 1 & 0 & -1 & \\ & \ddots & \ddots & -1 \\ 0 & & 1 & 1 \end{pmatrix}.$$

where l is the length of the elements.

- For the updated value of the discharge q^{n+1}:

$$\int_{\Omega} q^{n+1}\varphi\, d\Omega = \int_{\Omega} q^{n+\frac{2}{3}}\varphi\, d\Omega - \Delta t g h^n \int_{\Omega} \frac{\partial \xi^{n+1}}{\partial x}\varphi\, d\Omega + \frac{q^{n+\frac{2}{3}}}{h^n}\int_{\Omega}(\xi^{n+1}-\xi^n)\varphi\, d\Omega \qquad (6.6)$$

indicating by $\underline{\mathrm{M}}_L^q$ the mass lumped matrix of discharge and by $\underline{\mathrm{C}}^q$ the matrix $\int_\Omega \frac{\partial \varphi}{\partial x}\varphi\, d\Omega$, (6.6) becomes:

$$\underline{\mathrm{M}}_L^q \underline{q}^{n+1} = \underline{\mathrm{M}}_L^q \underline{q}^{n+\frac{2}{3}} - \Delta t g \underline{h}^n (\underline{\mathrm{C}}^{\xi})^T \underline{\xi}^{n+1} + \frac{\underline{q}^{n+\frac{2}{3}}}{\underline{h}^n}\underline{\mathrm{M}}_L^q(\underline{\xi}^{n+1}-\underline{\xi}^n) \qquad (6.7)$$

Remark 6.4. *In the equation (6.7), the ξ values considered are those calculated in the boundary nodes and in the middle node of each element.*

- For the c_1 pollutant:

$$\int_{\Omega} c_1^{n+1}\varphi\, d\Omega + \Delta t \Gamma_1 \int_{\Omega} \frac{\partial c_1^{n+1}}{\partial x}\frac{\partial \varphi}{\partial x}\, d\Omega + \Delta t \int_{\Omega}(f_1 c_1)^{n+1}\varphi\, d\Omega = \int_{\Omega} c_1^n(\tilde{x})\varphi\, d\Omega + \int_{\Omega} s_1^{n+1}\varphi\, d\Omega \qquad (6.8)$$

by usual, indicating by $\underline{\mathrm{M}}^{c_1}$ and by $\underline{\mathrm{M}}_L^{c_1}$ the mass and lumped mass matrices respectively and by $\underline{\mathrm{A}}^{c_1}$ the stiffness matrix, (6.8) becomes:

$$[\underline{\mathrm{M}}_L^{c_1} + \Delta t \Gamma_1 \underline{\mathrm{A}}^{c_1} + \Delta t \underline{f}_1^{n+1}\underline{\mathrm{M}}_L^{c_1}]\underline{c}_1^{n+1} = \underline{\mathrm{M}}_L^{c_1}\underline{c}_1^n(\tilde{x}_{II}) + \Delta t \underline{\mathrm{M}}_L^{c_1}\underline{s}_1^{n+1} \qquad (6.9)$$

- For the c_2 pollutant, like at the previous point we have :

$$[\underline{\mathrm{M}}_L^{c_2} + \Delta t \Gamma_2 \underline{\mathrm{A}}^{c_2} + \Delta t \underline{f}_2^{n+1}\underline{\mathrm{M}}_L^{c_2}]\underline{c}_2^{n+1} = \underline{\mathrm{M}}_L^{c_2}\underline{c}_2^n(\tilde{x}_{II}) + \Delta t \underline{\mathrm{M}}_L^{c_2}\underline{s}_2^{n+1} \qquad (6.10)$$

Remark 6.5. *The matrices* $\underline{\mathrm{A}}^q$, $\underline{\mathrm{A}}^{c_1}$, $\underline{\mathrm{A}}^{c_2}$ *and* $\underline{\mathrm{M}}_L^q$, $\underline{\mathrm{M}}_L^{c_1}$, $\underline{\mathrm{M}}_L^{c_2}$ *are the same respectively. Thus the only matrices we have to construct are :* $\underline{\mathrm{A}}^q$, $\underline{\mathrm{A}}^{\xi}$, $\underline{\mathrm{M}}_L^q$, $\underline{\mathrm{M}}_L^{\xi}$, $\underline{\mathrm{B}}^{\xi}$ *and* $\underline{\mathrm{C}}^q$.

Remark 6.6. *We would stress that in the approach over presented, the only systems we have to solve are of elliptic kind and are relevant only to the ξ, c_1, c_2 variables.*

Remark 6.7. *Since the final matrix of the systems 6.5 is tridiagonal, the Thomas algorithm has been used for the solution of the algebraic system, while the Bi-CGSTAB solver has been used for the systems 6.7, 6.9 and 6.10.*

6.5 A problem with an analytical solution

In order to check the correctness and the efficiency of the model developed, we solved a problem with a known analytical function. The expression of the elevation and the discharge are similar to that of [132] and are the following:

$$\xi(x,t) = 0.2\sin(\frac{2\pi}{3800}x)\cos(\frac{2\pi}{3800}t)$$
$$q(x,t) = 0.9 - 0.2\cos(\frac{2\pi}{3800}x)\sin(\frac{2\pi}{3800}t)$$
$$h(x,t) = 3 + \xi(x,t)$$

In this problem the source function $f = s(x,t)$:

$$s(x,t) = 5.686ABC + 0.3924ABC^2D + \frac{0.36ADE}{3+0.2DC} - \frac{0.08ADE^2}{3+0.2DC} - \frac{[0.2ABC][0.9-0.2BE]^2}{(3+0.2DC)^2}$$

where

$$A = \frac{2\pi}{3800},\ B = \cos(Ax),\ C = \cos(At),\ D = \sin(Ax),\ E = \sin(At).$$

Two analytical solutions have also be taken for the concentrations c_1 and c_2:

$$c_1(x,t) = 1 + \sin(\frac{4\pi}{3800}x)\cos(\frac{4\pi}{3800}t)$$
$$c_2(x,t) = 1 - cos(\frac{4\pi}{3800}x)\sin(\frac{4\pi}{3800})$$

an the source functions for the concentrations c_1 and c_2 are:

$$s_1(x,t) = -F\sin(Fx)\sin(Ft) + uF\cos(Fx)\cos(Ft) + \Gamma_1F^2\sin(Fx)\cos(Ft) + f_1(1+\sin(Fx)\cos(Ft))$$

$$s_2(x,t) = -F\cos(Fx)\cos(Ft) + uF\sin(Fx)\sin(Ft) - \Gamma_2F^2\cos(Fx)\sin(Ft) + f_2(1-\cos(Fx)\sin(Ft))$$

where $u(x,t) = \frac{q(x,t)}{h(x,t)}$ and $F = \frac{4\pi}{3800}$.
We consider the domain $\Omega = [0, 3800]$, partitioned in 950 elements of equal length; the nodes number is 1901 for the discharge q and the pollutants c_1 and c_2; 951 for the elevation ξ. The transient studied was $10800s$ and $\Delta t = 100s$. The initial and boundary conditions were obtained by the analytical solutions and we imposed Dirichlet at inflow and Neumann at outflow for ξ, c_1 and c_2; while we imposed only Dirichlet at inflow for q. For physical parameters, we choose $\mu = 0$, $g = 9.81$, $\Gamma_1 = \Gamma_2 = 0.001$, $f_1 = 0.0000174$, $f_2 = 0.0115$. In the Table 6.1 are indicated the $\|\cdot\|_\infty$ error norm for elevation ξ, discharge q and pollutants c_1 and c_2 after $7200s$ and $10800s$. We indicated by $Max\ val$ the maximum value of the analytical solution at the corresponding time steps.

In Figures 6.2 - 6.5 are reported the distribution of analytical and numerical solutions of elevation ξ, discharge q and concentrations c_1 and c_2 at time $10800s$ respectively. In Figures 6.6 - 7.2 are reported the time evolution of analytical and numerical solutions of elevation ξ, discharge q and concentration c_1 and c_2 at node $238(x = 237)$ respectively.

Table 6.1: $\|\cdot\|_\infty$ error norm of the elevation, discharge and the concentrations generated by the method after $7200s$ and $10800s$ respectively.

Solution	$\|\cdot\|_\infty$ error norm	$Max\ val$	Number of time steps
ξ	$1.501E-2$	$1.578E-1$	$7200s$
	$1.667E-2$	$1.093E-1$	$10800s$
q	$9.714E-2$	1.022	$7200s$
	$1.309E-1$	1.067	$10800s$
c_1	$9.908E-3$	1.245	$7200s$
	$8.448E-2$	1.401	$10800s$
c_2	$4.445E-2$	1.969	$7200s$
	$8.548E-2$	1.915	$10800s$

6.6 Conclusions of the chapter

The computed velocity field is sub-critical everywhere with maximum Froude number $F_r = \frac{u}{\sqrt{(gh)}} = 0.08$ and the boundary conditions imposed are coherent with the flow. The maximum percentage of errors (see Table 6.1 and Figures 6.2 - 6.3) seem quite large, about 15% for elevation and 12% for unit width discharge, but we have to consider the large time step $\Delta t = 100s$ used (in fact the Courant number is $C = 9.5$). Other numerical experiences with a smaller delta t gave more accurate results.

Finally we would stress the efficiency and the low computational cost of the new numerical model, the transient took $180s$ of elapsed time on a Sony VAIO EA series EA46FMW, that make it a promising tool for the prediction of the distribution of chemical pollutants in a river.

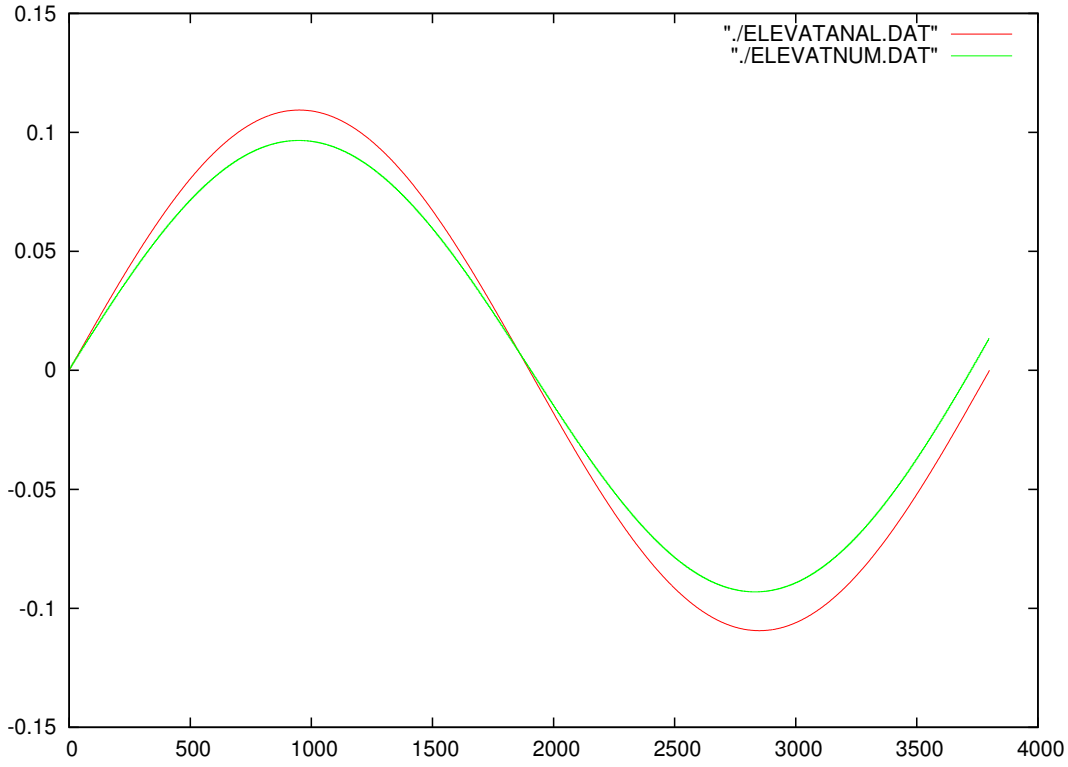

Figure 6.2: Distribution of analytical and numerical elevation ξ at $t = 10800s$.

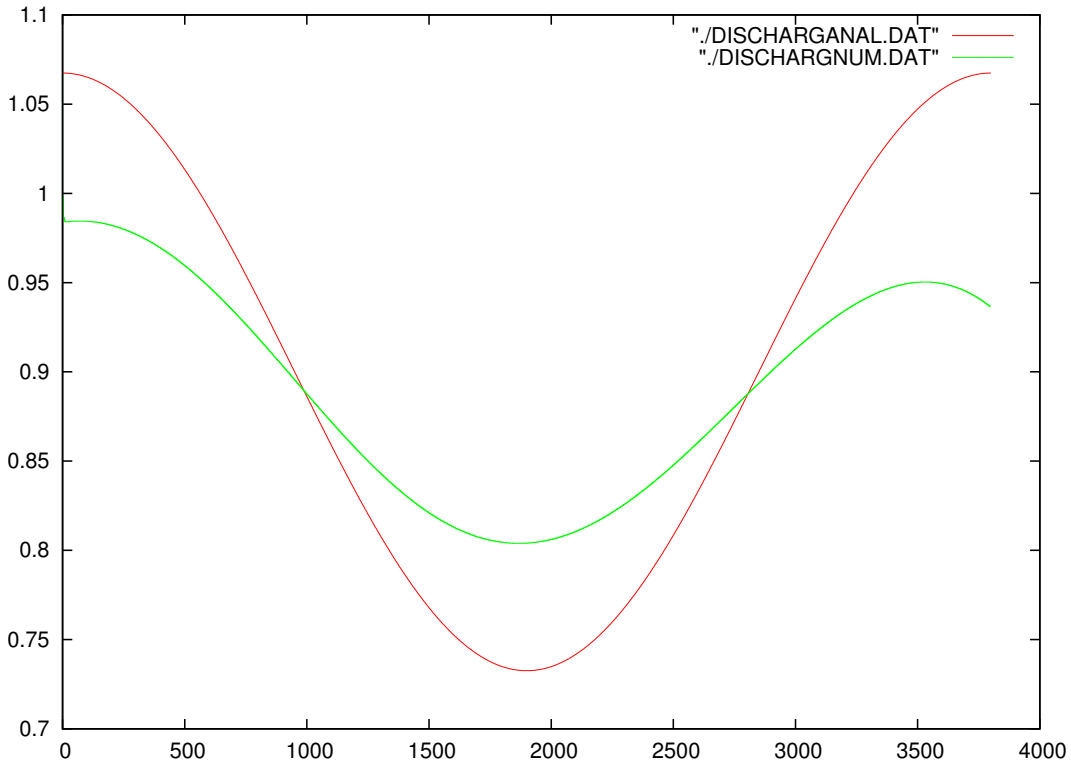

Figure 6.3: Distribution of analytical and numerical discharge q at $t = 10800s$.

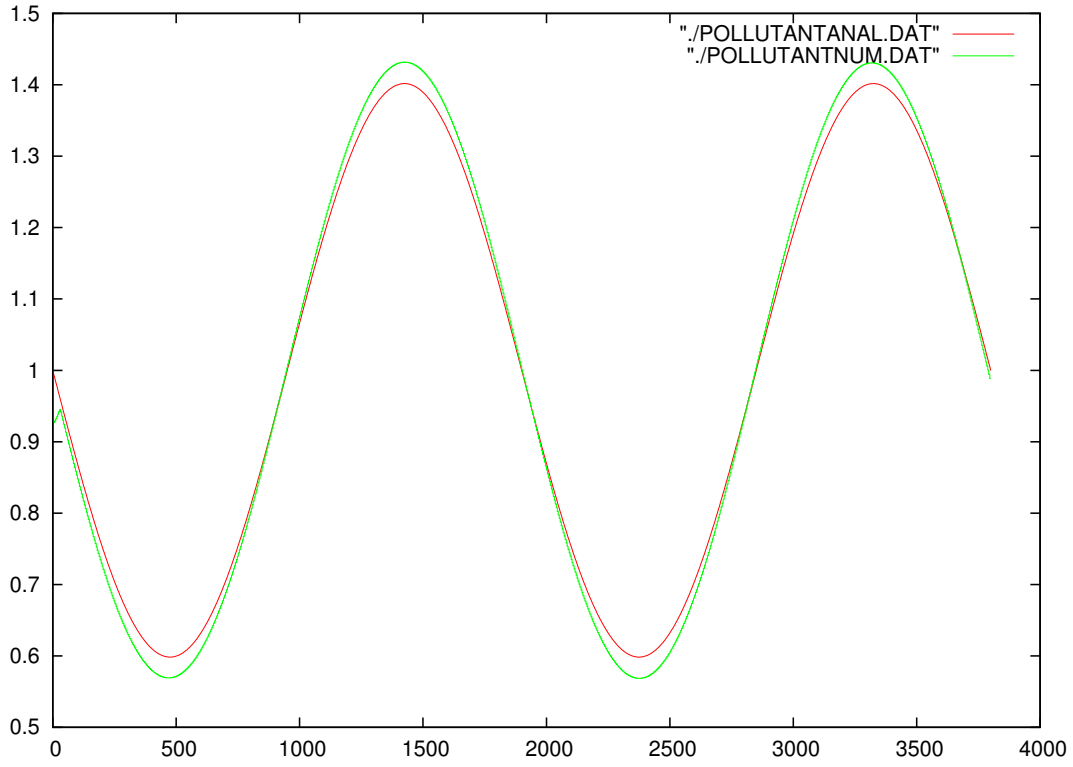

Figure 6.4: Distribution of analytical and numerical concentration c_1 at $t = 10800s$.

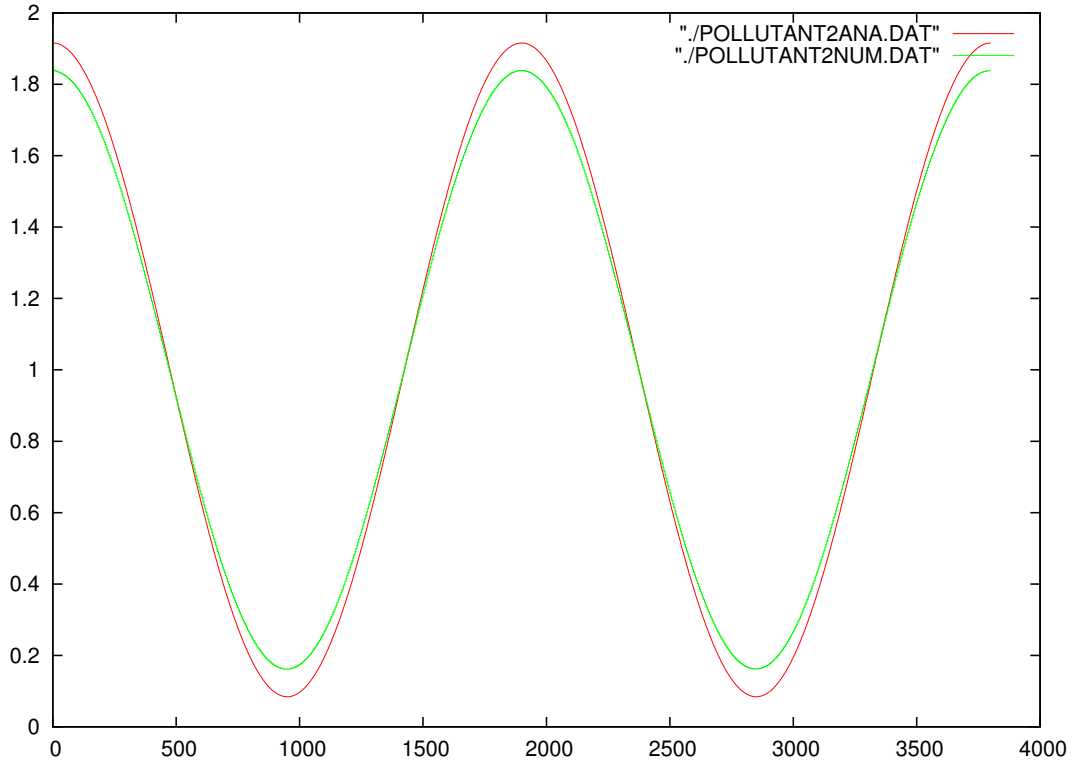

Figure 6.5: Distribution of analytical and numerical concentration c_2 at $t = 10800s$.

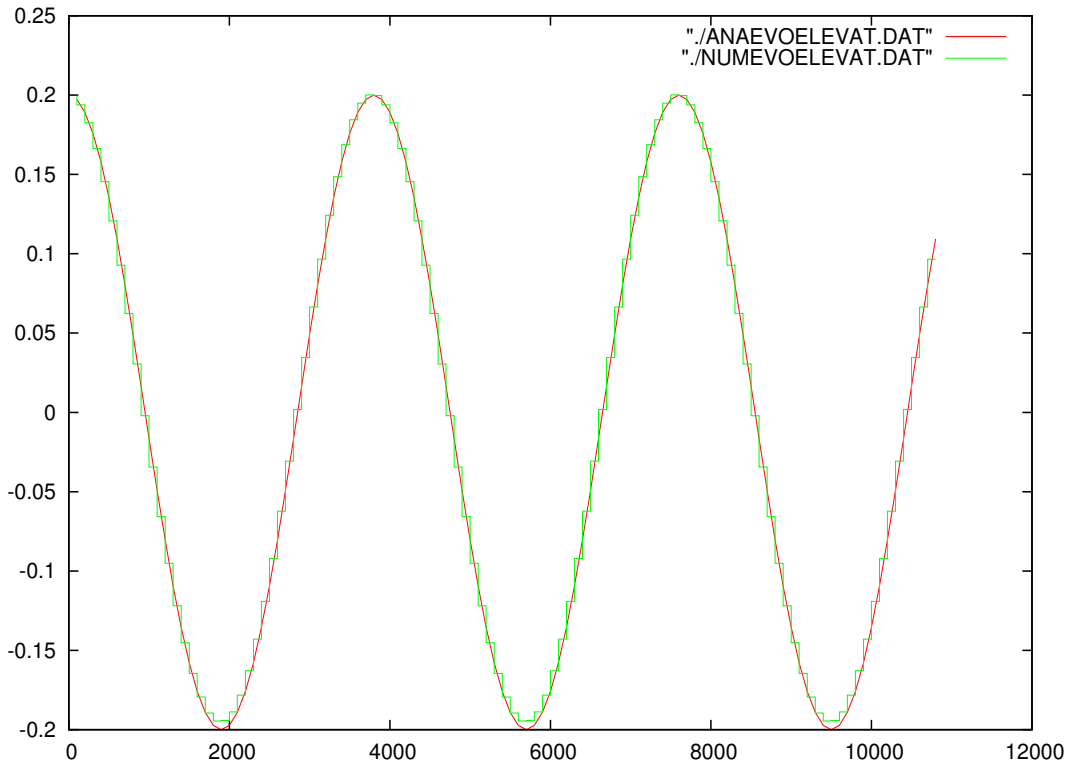

Figure 6.6: Time evolution of analytical and numerical elevation ξ at the node 238 ($x = 237$).

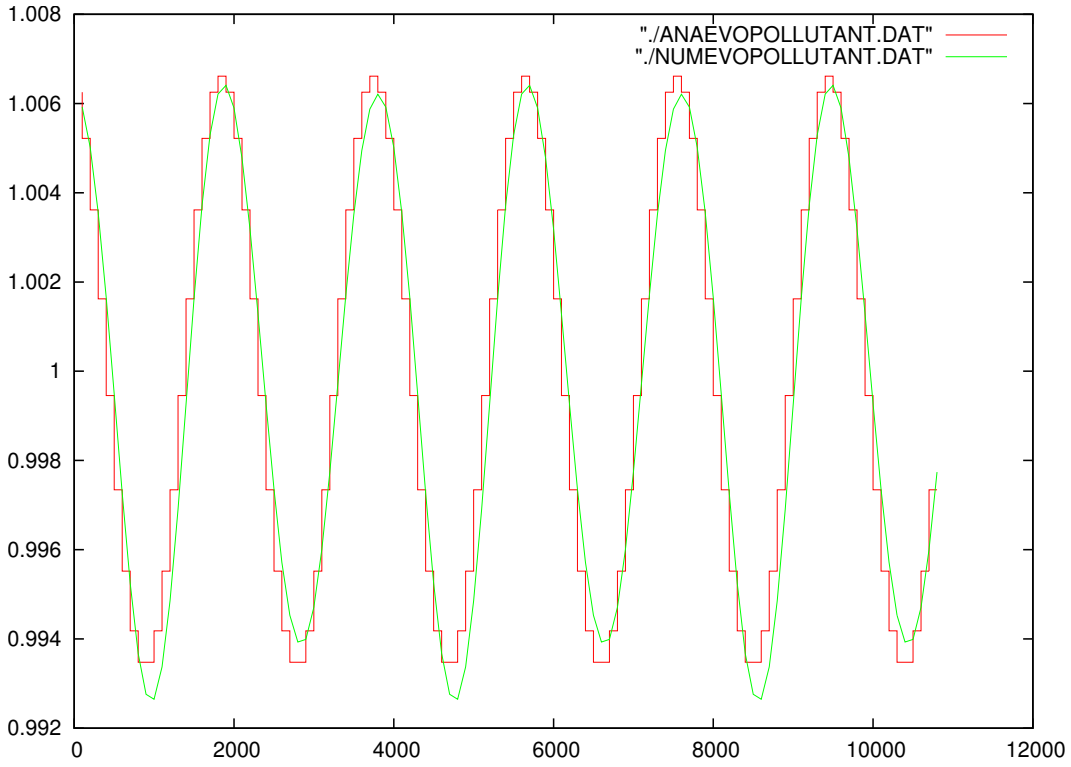

Figure 6.7: Time evolution of analytical and numerical concentration c_1 at node 238 ($x = 237$).

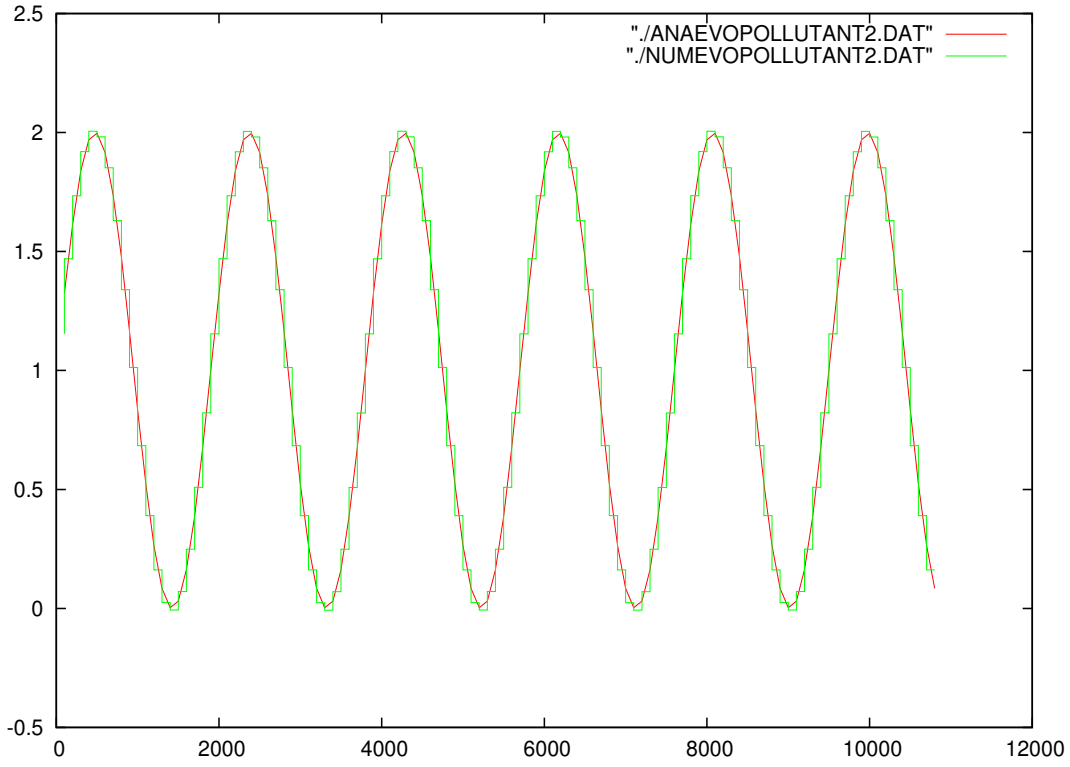

Figure 6.8: Time evolution of analytical and numerical concentration c_2 at node 238 ($x = 237$).

Chapter 7

Software

In this chapter, we shortly present the software TRIANGLE-ANVI-ELFICS-ROTINS-GNUPLOT in which the method and the algorithms from Chapter 2 - 6 where implemented and by which the results of numerical tests were obtained. Actually has been also developed software for the solution of the 1D Shallow-Water equations, but being the algorithms used very similar to those used for the 2D Navier-Stokes problems, hase been considered convenient do not dilate any more. In section 7.1 we give an overview of the preprocessor ANVI and TRIANGLE. The section 7.2 presents the flow chart and a brief description of ELFICS which is a code for the solution of elliptic equations with the two methods: traditional FE and conservative FE eventually preconditioned by a Schwarz overlapping multi-domain approach. In section 7.2, we present the flow chart and a brief description of ROTINS code that solves the Navier-Stokes equations with characteristics and stabilized fractional step and we end this chapter by a short presentation in section 7.4 of the post-processor GNUPLOT used.

7.1 Preprocessor

7.1.1 TRIANGLE

The TRIANGLE package was taken from GEOMPACK [66], a mathematical software package which contains Fortran 77 routines for the generation of two dimensional triangular and three dimensional tetrahedral finite element meshes using efficient geometric algorithms. GEOMPACK currently contains routines for constructing two and three dimensional Delaunay triangulations, decomposing a general polygonal region into simple or convex polygons, constructing the visibility polygon of a simple polygon from a fixed view point and other simpler geometric algorithms. The input data are :

- RGNAME: name of the domain
- TOLIN: numerical tolerance
- ANGSPC: angle spacing parameter in radians used in controlling vertices to be considered as an end point of a separation
- ANGTOL: angle tolerance parameter in radians used in accepting separator(s)
- KAPPA: mesh smoothing parameter belonging to [0.0,1.0]
- NMIN: parameter used to determine if a sufficiently large number of triangles are in polygon
- NTRID: number of desired triangle
- CASE = 1: simple polygon or multiply connected polygonal region

- CASE = 2: general polygonal region with holes and interfaces
- NCUR: number of curves internal to the domain
- MSGLVL: message level, initialize to 0, for no debugging
- NVBC: number of vertices of the boundary

GEOMPACK generates a triangulation whose principal outputs are :

- NVC: total number of triangle vertices
- NTRI: total number of elements
- VCL(1:2;1:NVC): vertices coordinates lists
- A topological table of the elements of the triangulation which is composed by the three vertices defining each element in the triangulation
- A topological table of the adjacent elements of each element that also indicate whose elements have an edge on the physical boundary.

Starting from the output generated by TRIANGLE, we constructed ANVI.

7.1.2 ANVI

The ANVI output was built in order to satisfy the request that each triangle has six nodes (the vertices and the mid-edges) on proper boundary. It takes as input some output of TRIANGLE that we list in section 7.1.1 and generates the following output :

- NNOD: total number of nodes of the domain (vertices and mid-edges included)
- VVCL(1:2; 1:NNOD): vertices and mid-edges coordinates lists that respect a well-defined order
- A first topological table of elements that includes the mid-edge nodes
- A second topological table that identifies if the edges of each elements are internal to the domains or on the physical boundaries.
- Total number of nodes that belong to the physical boundaries and the lists of these nodes.

These ANVI data are input used for running of ELFICS and ROTINS.

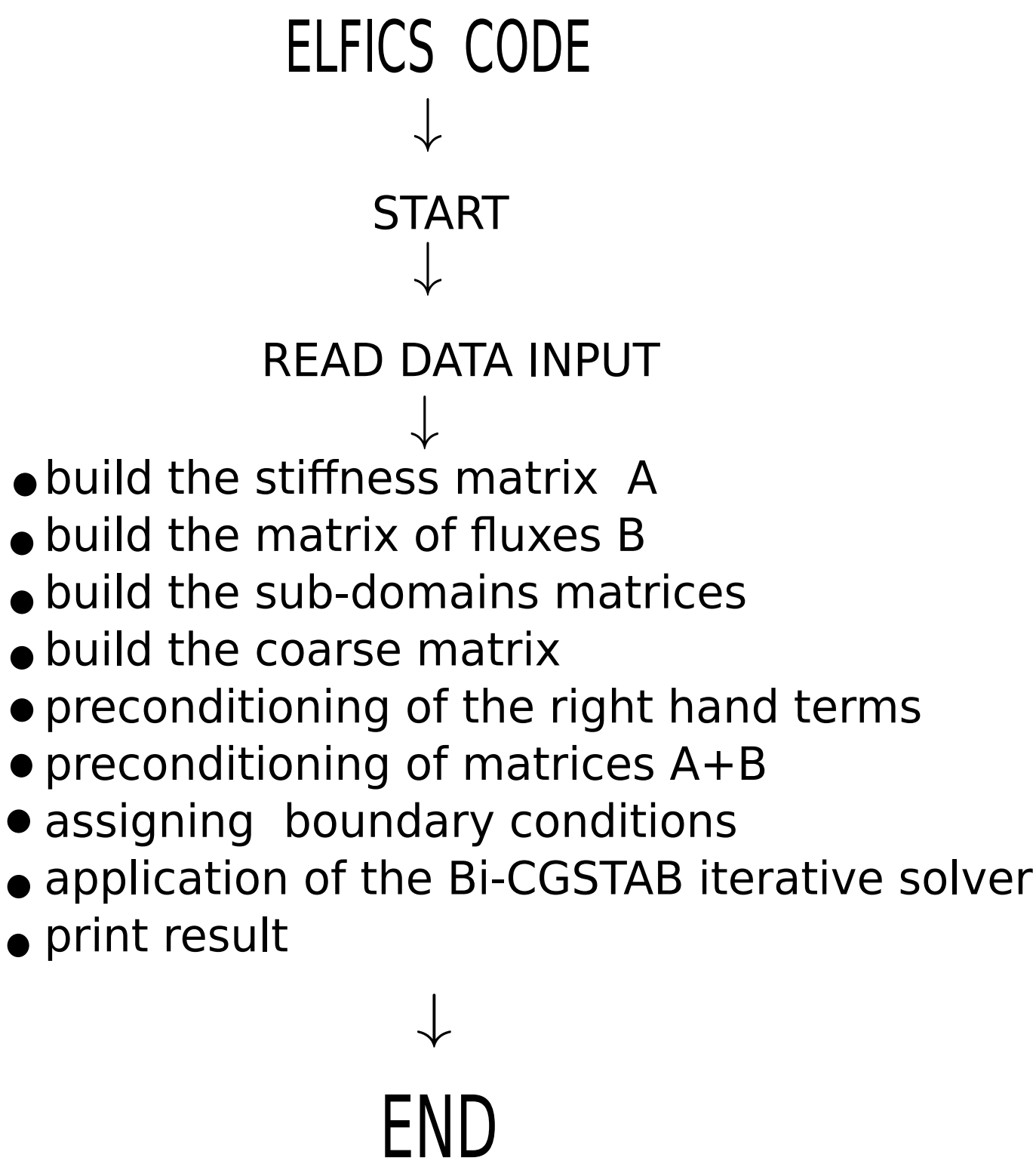

Figure 7.1: ELFICS flow chart.

7.2 ELFICS

ELFICS is a code that was developed in Fortran 90 for the solution of elliptic equations with traditional or conservative FE approach and where the algebraic system is solved with or without the Schwarz additive overlapping multidomain preconditioner. There are many routines in ELFICS and the most important are :

- TRANSJAC: computes the Jacobian of the transformation of each element from the

physical space to the reference space

- STIF: computes the local stiffness matrix
- COORDAXIS: evaluates the six base functions at the seven Gauss points (vertices, mid-edges and barycenter)
- RIGHTVEC: computes the local mass matrix and moreover, for a given source term, it computes the local right hand side vector
- EDGENORCOM: computes the outward normal to the edges of an element
- Q1Q2TRANSFIRST: transforms the flux on the first edge of each triangle from the general space to the first edge of the reference space when in the general space the adjacent element has his first edge common to the element considered
- Q1Q2TRANSSECOND: transforms the flux on the first edge of each triangle from the general space to the first edge of the reference space when in the general space the adjacent element has his second edge common to the element considered
- Q1Q2TRANSTHIRD: transforms the flux on the first edge of each triangle from the general space to the first edge of the reference space when in the general space the adjacent element has his third edge common to the element considered
- Q2Q3TRANSFIRST: transforms the flux on the second edge of each triangle from the general space to the second edge of the reference space when in the general space the adjacent element has his first edge common to the element considered
- Q2Q3TRANSSECOND: transforms the flux on the second edge of each triangle from the general space to the second edge of the reference space when in the general space the adjacent element has his second edge common to the element considered
- Q2Q3TRANSTHIRD: transforms the flux on the second edge of each triangle from the general space to the second edge of the reference space when in the general space the adjacent element has his third edge common to the element considered
- Q3Q1TRANSFIRST: transforms the flux on the third edge of each triangle from the general space to the third edge of the reference space when in the general space the adjacent element has his first edge common to the element considered
- Q3Q1TRANSSECOND: transforms the flux on the third edge of each triangle from the general space to the third edge of the reference space when in the general space the adjacent element has his second edge common to the element considered
- Q3Q1TRANSTHIRD: transforms the flux on the third edge of each triangle from the general space to the third edge of the reference space when in the general space the adjacent element has his third edge common to the element considered
- DIRVALUE: assigns the Dirichlet value with respect to a node belonging to the Dirichlet boundary
- NEWMANCONDI1: approximates the boundary conditions when the first edge of the element considered belongs to the Neumann boundary
- NEWMANCONDI2: approximates the boundary conditions when the second edge of the element considered belongs to the Neumann boundary
- NEWMANCONDI3: approximates the boundary conditions when the third edge of the element considered belongs to the Neumann boundary

The Bi-CGSTAB iterative solver was taken from the Fortran Numerical recipes [141]. The mains routines used are :

- SPRSIN: converts a square matrix $A(1:N, 1:N)$ into a row-indexed sparse storage
- LINBCG: bi-conjugate gradient solution of a sparse system.

The Schwarz overlapping preconditioner can be thought like composed by a number of module as large as the number of the sub-domains used plus one. In fact if $nsub$ is the number of non overlapping sub domains, then we should have $nsub + 1$ submodules where the adjoined submodule is devoted to the treatment of the coarse mesh. The principal routines for Schwarz are:

- VERTPOL: loads the vertices coordinates of the polygon boundary of the sub-domains
- MINRECTPOL: finds the vertices coordinates of the minimum rectangle containing the sub-domains
- TOPOSUB: searches the topological matrix of the elements included in the minimum rectangle containing the sub-domains
- TRANSPOLYGON: computes the matrix of the transformation in a suitable reference space in order to find the elements with one or two edges belonging to the internal boundary of each sub-domains
- DISIREDTOPO: extracts from the general topological matrix the topological matrix of each sub-domains
- OVERLAPTOPO: finds all the elements that belong to the strips of overlapping and builds the final topological matrices of the overlapping sub-domains
- BARYCENTER: computes the barycenter of the big triangles of each sub-domains,
- NODEBARY: finds the closest nodes to the barycenter and keeps these nodes as nodes of the coarse mesh
- LOCALCOARSENODE: finds all the nodes of the coarse mesh and their positions with respect to the global numeration of the nodes of the domains.

To implement the Schwarz overlapping additive preconditioner, it was convenient to use SPARKIT [121] which is a tool package for manipulating and working with sparse matrices and to use also LAPACK [75] which is a linear algebra package. In SPARKIT, we use the FORMATS module for the storage of matrices and the ITSOL module for the incomplete factorization. In LAPACK we were interested in a storage type format called PACKET STORAGE FORMAT. Package routines for Schwarz coming from SPARKIT and LAPACK are the following :

- DNSCR: converts a densely stored matrix into a row oriented compactly sparse matrix
- MSRCSR: converts a compressed matrix using a separated diagonal (modified sparse row format) in the compressed sparse row format
- CSRDNS: converts a row-stored matrix into a dense one
- UPACSTORAGE: transforms an upper triangular matrix to packed storage format
- LOPACSTORAGE: transforms a lower triangular matrix to a packet storage format
- DTPTRI: computes the inverse of a triangular matrix stored in packet-format

- ILU: computes the incomplete LU factorization with dual truncation mechanism
- LPACSTODENSE: transforms a lower triangular matrix stored into a package format into a dense matrix
- UPACSTODENSE: transforms a triangular matrix stored into a package format into a dense matrix.

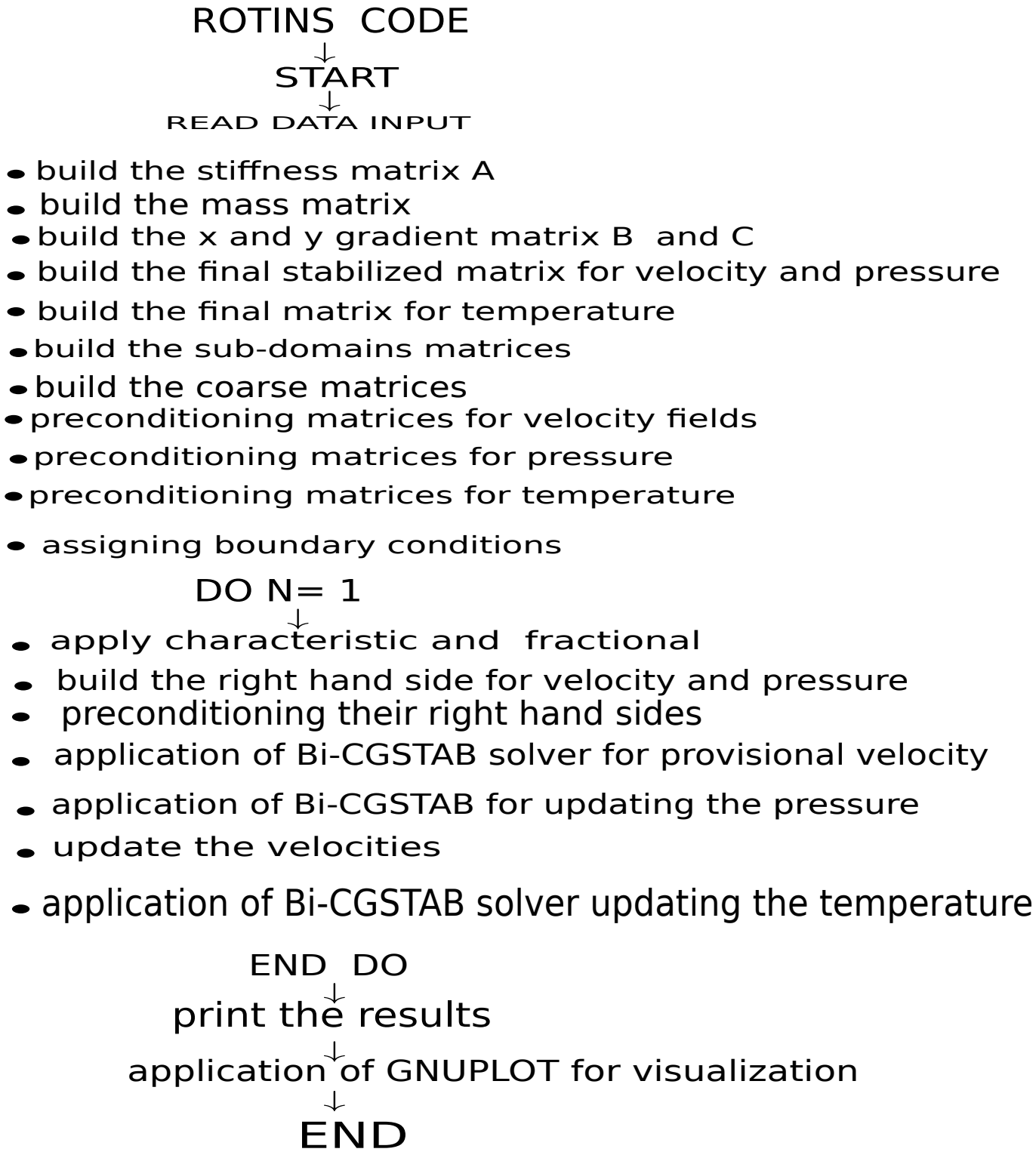

Figure 7.2: ROTINS flow chart.

7.3 ROTINS

ROTINS is the code that was developed for the solution of Navier-Stokes equations by means of characteristics, fractional step and stabilization techniques. The implementation of ROTINS was careful about efficiency issues and it suits parallel computer architectures inherently. It is divided into three modules: the matrix computation module, the solutions of algebraic system without Schwarz preconditioner module and the Schwarz overlapping additive preconditioner module.

7.3.1 MODULE : computation of matrices

In this module, we compute all the matrices involved in the algebraic formulation (5.39)-(5.44) and we store in a suitable file. The main routine is:

- ROTIZGENERAL: computes all the matrices of the equations of the two velocity components, of the updated values of the pressure correction, of the correction of the velocity, and of the updated values of the temperature equations. In order to reduce the computing time, ROTIZGENERAL calls some of the routines that we listed in ELFICS (see Section 7.2) and in this way, all the computations are made once.

7.3.2 MODULE : solution of algebraic system without Schwarz preconditioner

This module takes the matrices that have been computed in the modules of Subsection 7.3.1 and using characteristics plus the fractional step with the stabilized technique, solves at each time instant the six algebraic systems (5.39) - (5.44) by means of iterative Bi-CGTSAB solver without the Schwarz preconditioner. The mains routines are :

- CHARACTERICTICS: computes the foot of characteristics for all the nodes of the elements
- CHARBARYT: computes the foot of characteristics for the barycenter of the elements
- FINDTRIANGLE: finds element that at the time instant $t = t_{n+1}$ contains the foot of characteristics
- ITERPROCESS: interpolates the nodal value at the foot of characteristics using the values obtained at the previous times instant. The values used can be the velocities field or the temperature
- BARYCENTER: computes the barycenter of each element of the triangulations
- INITVAL: computes the initial value of the velocity field, the pressure and the temperature
- PEANROTIZFINALE: uses the previous routines and makes the advancing time, solving each algebraic system by means of the Bi-CGSTAB iterative solver whose routines have been already given in ELFICS.

7.3.3 MODULE : solution of algebraic system with Schwarz overlapping preconditioner

This module is similar to the previous one 7.3.2; the most important difference between them consisting in the use of Schwarz preconditioner. For constructing the Schwarz preconditioners for all the six matrices necessary at the computations of the updated values relevant to the physical variables, the routines listed in ELFICS are used. When the preconditioners are built they are stored in a suitable file. The next step consists in using the routines :

- PEANROTIZFINALEPRECONDITIONED: makes the advancing time and solves iteratively by means of the Bi-CGSTAB solvers at each time step the six algebraic systems that are preconditioned by Schwarz
- DGECO: is taken from [141] and computes the condition number of the six matrices.

The efficiency of ROTINS is such that it can handle easily other types of partial differential equations such as the time dependent parabolic equations and the time dependent convection diffusion equations.

7.4 Post-processor

The graphics produced in this thesis for the test problems of chapter 5 were obtained by means of GNUPLOT [140]. In fact, after the fixed temporal steps, ROTINS outputs the numerical solutions for the velocity, the pressure and the temperature fields which are stored in a suitable format in some data files. Then suitable commands of GNUPLOT are applied to obtain the desired graphics.

Conclusions

In this thesis, a new numerical approach for the solution of 2D incompressible Navier-Stokes equations that requires little computational effort has been addressed. The most important features of the new model are: using polynomials of degree two (both for velocities and pressure) in a FE spatial approximation; advancing in time by a fractional step approach in which the non linear convective terms are approximated by characteristics, adding of suitable stabilizations techniques in order to overcome the instabilities inherent to the equal order choice. The new technique reduces the most expensive computational kernels to the solution of algebraic systems stemming from elliptic problems. In order to reduce as most as possible the computational effort, an iterative method (Bi-CGSTAB), preconditioned by an additive Schwarz preconditioner has been used.

The new model has been tested solving several problems, at first of elliptic, parabolic and convective-diffusive kind, then in some time dependent and stationary Navier-Stokes problems such as the two dimensional unsteady flow of decaying vortices and the lid driven cavity flow having known analytical solutions, and the problem of natural convection in a square cavity. In all the cases the results demonstrate in complete agreement with the theoretical previsions and with the results at disposal from literature, confirming the accuracy and efficiency of the model. The numerical schemes above mentioned (with the obvious modifications due to the specific problem under study) have been applied for the solution of 1D Shallow-Water equations. Actually in this application, the polynomial spaces are not of equal degree, but they are of degree one for elevation and of degree two for the discharge. Also, the model relevant to the shallow-Water has been tested solving a problem with known analytical solution. We would stress that the efficiency and the low computational cost of the new numerical model, make it a promising tool for the prediction of the distribution of chemical pollutants in a river.
A suitable software has been developed for both the fluid dynamics models.

Another potentially interesting tool developed and tested in the thesis, and resembling some techniques recently developed in the framework of discontinuous FE, is a weak formulation of elliptic and parabolic problems able to guarantee the fluxes conservation. Unfortunately, it appears that the conservative finite element method is convergent but not conservative according to the classical definition; however it could be generalized so that a scheme genuinely conservative, like those named finite volume-elements, could be obtained.

Finally, since the computational kernels are of elliptic kind, the feasibility to apply well-established h-adaptivity techniques would be very advantageous.

Acknowledgements

First of all, I want to express my special gratitude to my supervisor Professor Vincenzo Angelo Pennati, for having introduced me in the fascinating world of Numerics and Mathematics of Computation. I have appreciated his patient attitude, affection, availability and precious advices; moreover, his experience and the trust he put in me have been fundamental for my scientific development. I have had the privilege to be his last student since his has closed officially his research and teaching activities. I have enjoyed every second working with him during these 3 years.
I thank Professor François Guibault (Ecole Polytechnique de Montreal) for having accepted to be member of the jury, and for reading this manuscript carefully.
I thank the Chair of the Doctoral programme Professor Stefano Serra Capizzano for having generously putting me in a stimulating research environment. Many thanks to Dr. Matteo Semplice for his comments, suggestions during many discussions.
The financial support of the present work by the Italian Government via the University of Insubria is gratefully acknowledged.
I thank Professor Antonio Di Guardo and Melissa Morselli for their sincere collaboration in the studies of the distribution of pollutants in the NOVELLA river. I am very grateful for their contributions. It is a pleasure for me to thank all members and staffs of the Department of Sciences and High Technology of the University of Insubria for their unfailing assistance.
A special thanks go to Pennati Family, Avalone Family and Vacarro Family. They have had a huge help in my studies.
Finally, my thanks go to my mother Madeleine, my sisters Louisette, Ruth, Ingrid and my fiancée Rolande for the love, support and understanding we share beyond all obstacles.
Last but not the least, I am thanking the Lord for this great achievement in my life.

Bibliography

[1] P.F. Antonietti, B. Ayuso and L. Heltai, *Schwarz domain decompsition preconditioners for interior penalty approximations of elliptic problems*, Tech. Rep. IMATI-CNR, PV, 2005.

[2] V.I. Agoshkov, D. Ambrosi, V. Pennati, A. Quarteroni and F. Saleri, *Mathematical and numerical modeling of shallow water flow*, Computational Mechanics 11, 280-299, 1993.

[3] D.N. Arnold, F. Brezzi, B. Cockburn and L.D. Marini, *Unified Analysis of Discontinuous Galerkin methods for elliptic problems*, SIAM J. Num. Anal., 39, 1749-1779, 2002.

[4] S. Agmon, *Lectures on elliptic boundary values problems*, Van Nostrand, Princeton, NJ, 1965.

[5] D.N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19, 742-760, 1982.

[6] J.P. Aubin, *Approximation des problèmes aux limites non homogènes pour des opérateurs non-linéaires*, J. Math. Anal. Appl., 30, 510-521, 1970.

[7] D.N. Arnold, F. Brezzi and L.D. Marini, *A family of Discontinuous Galerkin methods for the Reissner- Mindlin plate*, J. Sci. Comput., 22, 119-145, 2005.

[8] D. Ambrosi, S. Corti, V. Pennati and F. Saleri, *Numerical simulation of unsteady flow at Po river delta*, Journal of Hydraulic Engineering, 12, 735-743, 1996.

[9] I. Babuška, *The finite element method with penalty*, Math. Comp., 27, 221-228, 1973.

[10] G.A. Baker, *Finite elements methods for elliptic equations using nonconforming elements*, Math. Comp., 31 , 45-59, 1977.

[11] C. Baietti, G. Cerutti and A. Fucci, *Metodi h-adattivi per la risoluzione di problemi ellittici 2D. Exam Report of the course "Soluzione numerica di PDE's"*, a.a. 2010-2011.

[12] P. Bochev, Z. Cai, T.A. Manteuffel and S.F. McCormick, *Analysis of velocity-flux first order system least-squares principles for the Navier-Stokes equations*: Part I, SIAM J. Numer. Anal., 35, 990-1009, 1998.

[13] B.B. Buzbee, F.W. Dorr, J.A. George and G.H. Golub, *The direct solution of discrete Poisson equation on irregular region*, SIAM, J. Numer. Anal., 8, 722-736, 1971.

[14] F. Brezzi and J. Douglas, *Stabilized mixed methods for the Stokes problem*, Numer. Math., 53, 225-235, 1988.

[15] G.A. Baker, V.A. Dougalis and O.A. Karakashian, *On a higher order accurate fully discrete Galerkin approximation to Navier- Stokes equations*, Math. Comput., 39, 339-375, 1982.

[16] A.N. Brooks and T.J.R. Hughes, *Streamline-upwind/ Pretrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Method. Appl. Mech. Engrg., 32, 199-259, 1982.

[17] K. Boukir, Y. Maday, B.Métivet and E. Razafindrakoto, *A high-order characteristics/finite element method for the incompressible Navier-Stokes equations*, Int. J. Numer. Meth. Fluids, 25, 1421-1454, 1997.

[18] C.E. Baumann and J.T. Oden, *A discontinuous hp finite element method for the Navier-Stokes equations*, 10th. International Conference on Finite Element in fluids, 1998.

[19] B.R. Baliga and S.V. Patankar, *A new finite element formulation for convection-diffusion problems*, Numer. Heat Transfer, 3, 393-409, 1980.

[20] J.H. Bramble, J.E. Pasciak and X. Zhang, *Two level preconditioners for 2m'th order elliptic finite element problems*, East-West J. Numer. Math., 4, 99-120, 1996.

[21] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131, 267-279, 1997.

[22] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti and M. Savini, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, 2nd European Conference on turbomachinary Fluid Dynamics and Thermodynamics (Antwerpen, Belgium); R. Decuypere and G. Dibelius eds., Technologish Instituut, 99-108, March 5-7, 1997.

[23] P.R. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comput., 11, 450-481, 1990.

[24] A.J. Chorin, *The numerical solution of the Navier Stokes equations for an incompressible fluid*, Bull. Amer. Math. Soc., 73, 928-931, 1967.

[25] A.J. Chorin, *Numerical solution of the Navier-Stokes equations*, Math. Comput., 22, 745-762, 1968.

[26] H.G. Choi, H. Choi and J.Y. Yoo, *A fractional four-step finite element formulation of the unsteady incompressible Navier-Stokes equations using SUPG and linear equal-order finite element methods*, Comput. Meth. Appl. Mech. Engrg., 143, 333-348, 1997.

[27] Ph.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[28] Ph.G. Ciarlet and J.L. Lions, *Handbook of Numerical Analysis; Finite Element Methods (Part1)*. North-Holland, Amsterdam, 1991.

[29] Z. Cai and S. McCormick, *On the Accuracy of the Finite Volume Element Method for Diffusion Equations on Composite Grids*, SIAM J. Numer. Anal., 26, 3, 636-655, 1990.

[30] Z. Cai, J. Mandel and S. McCormick, *The Finite Volume Element method for diffusion equations on general triangulations*, SIAM J. Numer. Anal., 28, 2, 392-402, 1991.

[31] B. Cockburn and C.W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II*: General framework, Math. Comp., 52, 411-435, 1989.

[32] B. Cockburn and C.W. Shu, *The Runge-Kutta local projection P^1- discontinuous Galerkin method for scalar conservation laws*, RAIRO Modèl. Math. Anal. Numer., 25, 337-361, 1991.

[33] B. Cockburn and C.W. Shu, *The local discontinuous Galerkin finite element method for convection-diffusion systems*, SIAM J. Numer. Anal., 35, 2440-2463, 1998.

[34] B. Cockburn and C.W. Shu, *The Runge-Kutta discontinuous Galerkin finite element method for conservation laws V*: Multidimensional systems, J. Comput. Phys., 141, 199-224, 1998.

[35] R. Codina and O. Soto, *Approximation of the incompressible Navier-Stokes equations using orthogonal subscale stabilization and the pressure segregation on anisotropic finite element meshes*, Comput. Meth. Appl. Mech. Engrg., 193, 1403-1419, 2004.

[36] J. Douglas, Jr. and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, Lecture Notes in Physics, Vol 58, Springer-Verlag, Berlin, 1976.

[37] G. De Vahl Davis, *A natural convection of air in a square cavity: a benchmark numerical solution*, Int. J. Numer. Meth. Fluids, 3, 249-264, 1983.

[38] A. Deponti, V. Pennati and L. De Biase, *A fully 3D finite volume method for incompressible Navier-Stokes equations*, Int. J. Numer. Meth. Fluids, 52, 617-638, 2006.

[39] J. Donea and L. Quartapelle, *An introduction to finite element methods for transient advection problems*, Computer Methods in Applied Mechanics and Engineering, North-Holland 95, 169-203, 1992.

[40] J. Douglas, Jr. and T.F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element and finite difference procedures*, SIAM J. Numer. Anal., 19, 871-885, 1982.

[41] J. Douglas and J. Wang, *An absolutely stabilized finite element method for the Stokes problem*, Math. Computat., 52, 495-508, 1989.

[42] M. Dryja and O.B. Widlund, *Some domain decomposition algorithms for elliptic problems*, in L. Hayes and D. Kincaid, eds., iterative methods for large linear systems, 273-291, Academic Press, San Diego, California, 1989.

[43] M. Dryja and O.B. Widlund, *Additive Schwarz methods for elliptic finite element problems in three dimensions*. In Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, D.E. Keyes et al. eds. SIAM, Philadelphia, 3-18, 1992.

[44] R. Eymard, T. Gallouët and R. Herbin, *Finite Volume Methods*, update of the preprint no 97-19 du LATP, UMR 6632, Marseille 1997, in: Handbook of Numerical analysis, P.G. Ciarlet, J.L. Lions eds, 7, 713-1020, 2006.

[45] R.E. Ewing, T.F. Russell and M.F. Wheeler, *Simulation of miscible displacements using mixed methods and a modified methods of characteristics*, Computer. Meth. Appl. Mech. Engrg., 47, 73-92, 1984.

[46] M. Fortin and S. Boivin, *Iterative stabilization of bilinear velocity-constant pressure element*, Int. J. Numer. Meth. Fluids, 10, 1680-1697, 1990.

[47] L.P. Franca and T.J.R. Hughes, *Convergence analysis of Galerkin least-squares methods for advective-diffusive forms of the Stokes and incompressible Navier-Stokes equations*, Comput. Meth. Appl. Mech. Engrg., 105, 285-298, 1993.

[48] F. Feraudi and V. Pennati, *Transporto del calore in fluidi incomprimibili: un nuovo approcio numerico per problemi bidimensionali non stazionari*, L'Energia Elettrica, 74, 4, 238-251, 1997.

[49] L.P. Franca and R. Stenberg, *Error analysis of some Galerkin least-squares methods for elasticity equations*, SIAM J. Numer. Anal., 28, 1680-1697, 1991.

[50] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[51] M.D. Gunzberger, *Finite element methods for viscous incompressible fluids*, A guide to theory, practice and algorithms, Academic Press, San Diego, 1989.

[52] P. Grisvard, *Behaviour of the solutions of an elliptic boundary value problem in a polygonal or polyhedral domain*, in: Numerical solution of Partial Differential Equations, II, B. Hubbard ed., Academic Press, New York, 207-274, 1976.

[53] P.M. Gresho and S.T. Chan, *On the theory of semi implicity projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix*, Part. 2: Implementation, Int. J. Numer. Meth. Fluids, 11, 621-659, 1990.

[54] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-Splitting. Methods in nonlinear Mechanics*, SIAM, Philadelphia, 1989.

[55] J.L. Guermond and L. Quartapelle, *On stability and convergence of projection methods based on pressure Poisson equation*, Int. J. Numer. Meth. Fluids, 26, 1039-1053, 1998.

[56] V. Girault and P.A. Raviart, *Finite Element Methods for Navier Stokes-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[57] J.L. Guermond and A. Salgado, *A fractional step method based on a pressure Poisson equation for incompressible flow with variable density*, C. R. Acad. Sci. Paris, Ser- I 346, 923-918, 2008.

[58] G.H. Golub and C.F. Van Loan, *Matrix Computations,* 2nd edition, The Johns Hopkins University Press, Baltimore, 1989.

[59] P. Hartman, *Ordinary Differential Equations*, John Wiley and Sons, Baltimore, 1973.

[60] N.A. Hookey and B.R. Baliga, *Evaluation and enhancements of some control volume finite-element methods-part2.* Incompressible fluid flow problems, Numerical Heat Transfert, 14, 273-293, 1988.

[61] T.J.R. Hughes, L.P. Franca and M. Balestra, *Circumventing the Babuška-Brezzi condition: a stable Petrov-Galerkin formulation for the Stokes problem accomodating equal-order interpolations*, A new finite element formulation for computational fluids dynamics: V., Comput. Methods Appl. Mech. Engrg., 59, 85-89, 1986.

[62] T.J.R. Hughes, L.P. Franca and G.M. Hulbert, *The Galerkin/least-squares method for advective-diffusive equations*, A new element formulation for computational fluid dynamics:VIII., Comput. Meth. Appl. Mech. Engrg., 73, 2, 173-189, 1989.

[63] B. Heinrich and S. Nicaise, *The Nitsche mortar finite element method from transmission problems with singularities*, IMA J. Numer. Anal., 23, 331-358, 2003.

[64] J.G. Heywood, R. Rannacher and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations*, Int. J. Numer. Meth. Fluids, 22 (5): 325-352, 1996.

[65] A. Jameson, *Acceleration of transonic potential flow calculations on arbitrary meshes by the multigrid method*, AIAA, paper 79, 1458, 1979.

[66] B. Joe, *GEOMPACK. A software package for the generation of meshes using geometric algorithms*, Adv. Eng. Software, Vol. 13 No 5/6 Combined, 325-331, 1991.

[67] M. Joerg, *Numerical investigations of wall boundary conditions for two fluid flows*, Master's thesis, École Polytechnique Fédérale de Lausanne, 2005.

[68] A. Jameson, W. Schmidt and E. Turkel, *Numerical solutions to Euler equations by finite volume methods using Runge-Kutta marching schemes*, AIAA, paper 81,1259, 1981.

[69] A.C. Kengni Jotsa, *A new method for elliptic problems in CFD resembling the discontinuous FE approach*, Proceedings MASCOT 2010 Special Session, IMACS/ISGG, IMACS Series in Computational and Applied Mathematics, ULPGC, Las Palmas de Gran Canaria, Spain, 219-228, 2010.

[70] Y.A. Kuznetsov, *Matrix iterative methods in subspaces*, Proc. Int. Congress Math., Warsaw, 1983, North-Holland, Amsterdam, 1509-1521, 1989.

[71] P. Knabner and L. Angermann, *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*, Texts in Applied Mathematics 44, Springer, 2000.

[72] L.V. Kantorovich and V.I. Krylov, *Appropriate Methods of Higher Analysis*, P. Noordhoff, Groningen, 1964.

[73] J. Kim and P. Moin, *Application of fractional-step method to incompressible Navier-Stokes equations*, J. Comput. Phys., 59, 308-323, 1985.

[74] A.C. Kengni Jotsa and V.A. Pennati, *Solution of* $2D$ *convection-diffusion transient problems by a fractional-step FE method*, Submitted to Proceedings MASCOT 2011, IMACS/ISGG, IMACS Series in Computational and Applied Mathematics, Rome, 2011.

[75] *LAPACK: Linear Algebra Package*, http://www.netlib.org/lapack/, Univ. of Tennessee, Univ. of California Berkeley and NAG Ltd, November 2006.

[76] L.D. Landau and E.M. Lifshitz, *Fluids mechanics,* translated from the Russian by J.B. Sykes and W.H. Reaid, Course of Theoretical physics, Vol. 6. Pergamon Press, London, 1959.

[77] J. Leray, *Etude de diverses équations intégrales nonlinéaires et de quelques problèmes que pose l'hydrodynamique.* J. Math. Pures, 12, 1-82, 1933.

[78] J. Leray, *Essai sur les mouvements plans d'un liquide visqueux que limitent des parois*, J. Math. Pures Appl., 13, 331-418, 1934.

[79] J. Leray, *Essai sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math., 63, 193-248, 1934.

[80] M. Lesieur, *Turbulence in fluids*, Marinus Nijhoff, Dordrecht, 1987.

[81] M. Lesieur, *Turbulence in fluids*, 2nd edition, Kluwer, Dordrecht, 1990.

[82] J.L. Lions, *Problèmes aux limites non homogènes à données irrégulières: une méthode d'approximation*, Numerical Analysis of Partial Differential Equations (C. I. M. E. 2 Ciclo, Ispra, 1967), Edizioni Cremonese, Rome, 283-292, 1968.

[83] P.L. Lions, *On the Schwarz alternating method I*, in R. Glowinski, H.H. Golub, G.A. Meurant and J. Periaux eds, First International symposium on domain decomposition methods for partial differential equations, SIAM, Philadelphia, PA, 1-42, 1988.

[84] P.L. Lions, *On the Schwarz method II: Stochastic interpretation and order properties.* In domain decomposition methods, T.F. Chan et al. eds, SIAM, Philadelphia, 47-70, 1989.

[85] P.L. Lions, *Mathematical topics in fluid mechanics,* vol.1 and vol.3 of Oxford Lecture. Series in Mathematics and its applications, The Clarendon Press Oxford, New York, 1996.

[86] J.L. Lions and E. Magenes, *Problèmes aux Limites non Homogenes et Applications*, 1, Dunod, Paris, 1968.

[87] J.L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, vol. I, Springer-Verlag, Berlin, 1972.

[88] C. Liu and S. McCormick, *The Finite Volume Element method (FVE) for planar cavity flow*, Proc. 11th Internat. Conf. on CFD, Williamsburg, VA, June 28 - July 2, 1988.

[89] G.I. Marchuk, *Splitting and alternating direction methods,* in: Handbook of Numerical analysis, I, Ph. G. Ciarlet and J. L. Lions eds, North- Holland, Amsterdam, 197-462, 1990.

[90] P.W. McDonald, *The computation of transonic flow through two-dimensional gas turbine cascades*, AMSE, paper 71, GT-89, 1971.

[91] A.M. Matsokin, *Fictitious components and subdomain alternating methods*, in : Vychilisl. Algoritmy v Zadachakh Mat. Fiz. (Ed., V.V. Penenko), Vychisl. Tsentr Sib. Otdel. Acad. Nauk USSR, Novosibirsk, 66-88, (Translated version was issued in Sov. J. Numer. Anal and Math. Modelling), 5, 1990, 53-68, 1972.

[92] G.I. Marchuk, and Y.A. Kuznetsov, *Some problems in iterative methods*, in: Vychislite'nye Metody Lineinoi Algebry (Ed., G.I. Marchuck), Vychisl. Tsentr Sib. Otdel. Acad. Nauk USSR, Novosibirsk, 4-20, 1972.

[93] S.G. Mikhlin, *On the Schwarz algorithm*, Dokl. Akad. Nauk S.S.S.R., 77, 569-571, 1951.

[94] G.I. Marchuk, Y.A. Kuznetsov and A.M. Matsokin, *Fictitious domain and domain decomposition methods*, Sov. J. Numer. Anal and Math. Modelling, 1, 3-35, 1986.

[95] A.M. Matsokin and S.V. Nepomnyaschikh, *A Schwarz alternating method in a subspace*, Sov. Math., 29, 10, 78-84, 1985.

[96] S. McCormick and J. Thomas, *The fast adaptative compositive grid method (FAC) to elliptic boundary value problems*, Math. Comp., 46, 439-456, 1986.

[97] K.W. Morton, A. Priestley and E. Süli, *Stability of the Lagrange-Galerkin method with non-exact integration*, Modélisation mathematiques et analyse numerique, 22, 4, 625-653, 1988.

[98] P. Nithiarasu, R. Codina and O.C. Zienkiewicz, *The Characteristics Based-Split (CBS) scheme- a unified approach to fluid dynamics*, Int. J. Numer. Meth. Engrg., 66, 1514-1546, 2006.

[99] J.A. Nitsche, *Über ein variationsprinzip zur Lösung Dirichlet-Problemen bei Vewendung von Teilräumen, die keinen Randbedingunpen unteworfen sind*, Abh. Math. Sem. Univä. Hamburg, 36, 9-15, 1971.

[100] R.H. Ni, *A multiple grid method for solving the Euler equations*, AIAA, J., 20, 1565-1571, 1982.

[101] S.V. Patankar, *Numerical Heat Transfer and Fluid Flow*, Hemisphere, New York, 1980.

[102] A. Pazy, *Semi-groups of Linear Operator and applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[103] O. Pironneau, *On the transport-diffusion algorithm and its applications to the Navier-Stokes equations*, Numer. Math., 38, 309-332, 1982.

[104] O. Pironneau, *Méthodes des Élements Finis pour les fluides*, Masson, Paris, 1988.

[105] O. Pironneau, J. Liou and T. Tezduyar, *Characteristic-Galerkin and Galerkin/least-squares space-time formulations for the advection-diffusion equation with time-dependent domains*, Comput. Methods in Appl. Mech. Engrg., North-Holland, 100, 117-141, 1992.

[106] A. Priestley, *Evaluating the spatial integrals in the Lagrange-Galerkin method,* in Finite Elements in Fluids, New trends and Applications, K. Morgan, E. Oñate, J. Périaux and J. Peraire eds., Pineridge Press, Swansea, 146-155, 1993.

[107] A. Quarteroni, *Mathematical Aspects of Domain Decomposition Methods*, n.79/P, Dipartimento di Matematica, Politecnico Di Milano, 1992.

[108] A. Quarteroni, *Modellistica numerica per problemi differenziali*, 3^a edizione, Springer, 2006.

[109] A. Quarteroni and R. Ruiz-Baier, *Analysis of a finite volume element method for the Stokes problem*, Numer. Math. Springer-Verlag, 2011.

[110] A. Quarteroni, R. Sacco and F. Saleri, *Numerical Mathematics*, Texts in Applied Mathematics 37, Springer, 1991.

[111] A. Quarteroni, R. Sacco and F. Saleri, *Numerical Mathematics*, 2nd edition, Springer, 2008.

[112] A. Quarteroni and A. Valli, *Domain decomposition for a generalized Stokes problem*, Università degli studi di Trento, Dipartimento di Matematica, U.T.M., 259, 1988.

[113] A. Quarteroni and A. Valli, *Numerical approximation of Partial differential equations*, Springer Series in Computational Mathematics, vol. 23, Springer-Verlag, Berlin, 1997.

[114] A. Quarteroni and A. Valli, *Domain decomposition Methods for partial Differential Equations*, Oxford Scienze Publications, Oxford, 1999.

[115] R. Rannacher, *On Chorin's projection method for the incompressible Navier-Stokes equations,* in the Navier-Stokes Equations II: Theory and Numerical Methods, J.G. Heywood, K. Masuda, R. Rautmann and S.A. Solonnikov, eds., Springer- Verlag, Berlin, 167-187, 1992.

[116] W.H. Reed and T.R. Hill, *Triangular mesh methods for the neutron transport equation*, Tech. Report LA-UR-73-473, Los Alamos Scientific Laboratory, 1973.

[117] B. Rivière, M.F. Wheeler and V. Girault, *Improved energy estimated for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems*, Part I, Teach. Report 99 - 09, TICAM, 1999.

[118] S.E. Razavi, K. Zamzamian and A. Farzadi, *Genuinely multidimensional characteristic-based scheme for incompressible flows*, Int. J. Numer. Meth. Fluids, 57, 929-949, 2008.

[119] J. Shen, *On error estimates of projection methods for Navier-Stokes equations: first order schemes*, SIAM. J. Numer. Anal., 29, 57-77, 1992.

[120] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd edition, 2004.

[121] Y. Saad, *A basic tool kit for sparse matrix computations (Version 2)*, http://www-users.cs.umn.edu/ saad/software/SPARSKIT/index.html, 2009.

[122] J.C. Strikwerda, *Finite difference schemes and partial differential equations*, 2nd edition, SIAM, Philadelphia, 2004.

[123] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 2nd edition, Springer, 1991.

[124] B.F. Smith, P.E. Bjørstad and W.D. Gropp, *Domain Decomposition. Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, 1996.

[125] H.A. Schwarz, *Über einige Abbildungsdufgaben*, Ges. Math. Abh., 11, 65-83, 1869.

[126] P.J. Silvester and N. Kechkar, *Stabilized bilinear-constant velocity pressure finite elements for the conjugate gradient solution of the Stokes problem*, Comput. Methods Appl. Mech. Engrg., 79, 71-86, 1990.

[127] S.L. Sobolev, *Schwarz algorithm in the theory of elasticity*, Dokl. Akad. Nauk S.S.S.R., 4, 235-238, 1936.

[128] T.M. Shih and C.H. Tan, *Effects of grid staggering on numerical schemes*, Int. J. Numer. Meth. Fluids, 9, 193-212, 1989.

[129] S. Sun and M.F. Wheeler, *Mesh Adaptation strategies for Discontinuous Galerkin Methods Applied to Reactive Transport Problems,* in: H.W. Chu, M. Savoie and B. Sanchez, eds, Proceedings of Internationals Conference on Computing, Communications and Control Technologies, 223-228, 2004.

[130] R. Temam, *Sur l'approximation de la solution des equations de Navier-Stokes par la methode de pas fractionaire (II)*, Arch. Rat. Mech. Anal., 33, 377-385, 1969.

[131] R. Temam, *Navier Stokes Equations. Theory and Numerical Analysis*, 3rd edition, North-Holland, Amsterdam, 1984.

[132] G. Vignoli, V.A. Titarev and E.F. Toro, *Ader schemes for the Shallow Water equations in channel with irregular bottom elevation*, Preprint November 2006.

[133] G. Voronoi, *Nouvelles applications des paramètres continus à la thèorie des formes quadratures*, J. Reine Angeu. Math., 134, 198-287, 1908.

[134] A. Walters, *Numerically induced oscillations in the finite element approximations to the Shallow-Water equations*, Int. J. Numer. Meth. Fluids, 3, 591-604, 1983.

[135] M.F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Anal., 15, 152-161, 1978.

[136] K. Yosida, *Functional Analysis*, 4th edition, Springer-Verlag, Berlin, 1974.

[137] O.C. Zienkiewicz, *The Finite Element Method*, 3rd edition, McGraw Hill, London, 1977.

[138] O.C. Zienkiewicz and R. Codina, *A general algorithm for compressible and incompressible flow- part I. The Characteristic-Based Scheme*, Int. J. Numer. Fluids, 20, 869-885, 1995.

[139] Y. Zang, R.L. Street and J.R. Koseff, *A Non-Staggered Grid, Fractional Step Method for Time-Dependent Incompressible Navier-Stokes Equations in Curvilinear Coordinates*, J. Comput. Phys., 114, 18-33, 1994.

[140] *http://www.gnuplot.info.*

[141] *http://sci.ui.ac.ir/ sjalali/nrf/*, Numerical recipes fortran 77, Version 2.07.