
**BASICS OF FLUID MECHANICS AND
INTRODUCTION TO COMPUTATIONAL
FLUID DYNAMICS**

Numerical Methods and Algorithms

Volume 3

Series Editor:

Claude Brezinski

Université des Sciences et Technologies de Lille, France

BASICS OF FLUID MECHANICS AND INTRODUCTION TO COMPUTATIONAL FLUID DYNAMICS

by

TITUS PETRILA

Babes-Bolyai University, Cluj-Napoca, Romania

DAMIAN TRIF

Babes-Bolyai University, Cluj-Napoca, Romania

Springer

eBook ISBN: 0-387-23838-7
Print ISBN: 0-387-23837-9

©2005 Springer Science + Business Media, Inc.

Print ©2005 Springer Science + Business Media, Inc.
Boston

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Springer's eBookstore at:
and the Springer Global Website Online at:

<http://ebooks.springerlink.com>
<http://www.springeronline.com>

Contents

| | |
|---|------|
| Preface | xiii |
| 1. INTRODUCTION TO MECHANICS OF CONTINUA | 1 |
| 1 Kinematics of Continua | 1 |
| 1.1 The Concept of a Deformable Continuum | 1 |
| 1.2 Motion of a Continuum. Lagrangian and Eulerian Coordinates | 4 |
| 1.3 Euler–Lagrange Criterion. Euler’s and Reynolds’ (Transport) Theorems | 13 |
| 2 General Principles. The Stress Tensor and Cauchy’s Fundamental Results | 17 |
| 2.1 The Forces Acting on a Continuum | 17 |
| 2.2 Principle of Mass Conservation. The Continuity Equation | 18 |
| 2.3 Principle of the Momentum Torsor Variation. The Balance Equations | 20 |
| 2.4 The Cauchy Stress Tensor | 21 |
| 2.5 The Cauchy Motion Equations | 23 |
| 2.6 Principle of Energy Variation. Conservation of Energy | 24 |
| 2.7 General Conservation Principle | 25 |
| 3 Constitutive Laws. Inviscid and real fluids | 26 |
| 3.1 Introductory Notions of Thermodynamics. First and Second Law of Thermodynamics | 26 |
| 3.2 Constitutive (Behaviour, “Stresses-Deformations” Relations) Laws | 32 |
| 3.3 Inviscid (Ideal) Fluids | 34 |
| 3.4 Real Fluids | 38 |

| | | |
|-----|--|-----|
| 3.5 | Shock Waves | 43 |
| 3.6 | The Unique Form of the Fluid Equations | 49 |
| 2. | DYNAMICS OF INVISCID FLUIDS | 51 |
| 1 | Vorticity and Circulation for Inviscid Fluids. The Bernoulli Theorems | 51 |
| 2 | Some Simple Existence and Uniqueness Results | 55 |
| 3 | Irrotational Flows of Incompressible Inviscid Fluids. The Plane Case | 59 |
| 4 | Conformal Mapping and its Applications within Plane Hydrodynamics | 64 |
| 4.1 | Helmholtz Instability | 67 |
| 5 | Principles of the (Wing) Profiles Theory | 70 |
| 5.1 | Flow Past a (Wing) Profile for an Incidence and a Circulation “a priori” Given | 70 |
| 5.2 | Profiles with Sharp Trailing Edge. Joukovski Hypothesis | 72 |
| 5.3 | Theory of Joukovski Type Profiles | 74 |
| 5.4 | Example | 77 |
| 5.5 | An Iterative Method for Numerical Generation of Conformal Mapping | 79 |
| 6 | Panel Methods for Incompressible Flow of Inviscid Fluid | 81 |
| 6.1 | The Source Panel Method for Non-Lifting Flows Over Arbitrary Two-Dimensional Bodies | 81 |
| 6.2 | The Vortex Panel Method for Lifting Flows Over Arbitrary Two-Dimensional Bodies | 84 |
| 6.3 | Example | 87 |
| 7 | Almost Potential Fluid Flow | 92 |
| 8 | Thin Profile Theory | 95 |
| 8.1 | Mathematical Formulation of the Problem | 96 |
| 8.2 | Solution Determination | 97 |
| 9 | Unsteady Irrotational Flows Generated by the Motion of a Body in an Inviscid Incompressible Fluid | 100 |
| 9.1 | The 2-Dimensional (Plane) Case | 100 |
| 9.2 | The Determination of the Fluid Flow Induced by the Motion of an Obstacle in the Fluid. The Case of the Circular Cylinder | 102 |
| 9.3 | The 3-Dimensional Case | 103 |

| | | |
|------|---|-----|
| 9.4 | General Method for Determining of the Fluid Flow Induced by the Displacement of an Arbitrary System of Profiles Embedded in the Fluid in the Presence of an “A Priori” Given Basic Flow | 105 |
| 10 | Notions on the Steady Compressible Barotropic Flows | 110 |
| 10.1 | Immediate Consequences of the Bernoulli Theorem | 110 |
| 10.2 | The Equation of Velocity Potential (Steichen) | 113 |
| 10.3 | Prandtl–Meyer (Simple Wave) Flow | 115 |
| 10.4 | Quasi-Uniform Steady Plane Flows | 117 |
| 10.5 | General Formulation of the Linearized Theory | 118 |
| 10.6 | Far Field (Infinity) Conditions | 119 |
| 10.7 | The Slip-Condition on the Obstacle | 120 |
| 10.8 | The Similitude of the Linearized Flows. The Glauert–Prandtl Rule | 121 |
| 11 | Mach Lines. Weak Discontinuity Surfaces | 123 |
| 12 | Direct and Hodograph Methods for the Study of the Compressible Inviscid Fluid Equations | 127 |
| 12.1 | A Direct Method [115] | 128 |
| 12.2 | Chaplygin Hodograph Method. Molenbroek–Chaplygin equation | 129 |
| 3. | VISCOUS INCOMPRESSIBLE FLUID DYNAMICS | 133 |
| 1 | The Equation of Vorticity (Rotation) and the Circulation Variation | 133 |
| 2 | Some Existence and Uniqueness Results | 136 |
| 3 | The Stokes System | 138 |
| 4 | Equivalent Formulations for the Navier–Stokes Equations in Primitive Variables | 140 |
| 4.1 | Pressure Formulation | 140 |
| 4.2 | Pressure-Velocity Formulation | 142 |
| 5 | Equivalent Formulations for the Navier–Stokes Equations in “Non-Primitive” Variables | 143 |
| 5.1 | Navier–Stokes Equations in Orthogonal Generalized Coordinates. Stream Function Formulation | 144 |
| 5.2 | A “Coupled” Formulation in Vorticity and Stream Function | 151 |
| 5.3 | The Separated (Uncoupled) Formulation in Vorticity and Stream Function | 152 |

| | | |
|-----|---|-----|
| 5.4 | An Integro-Differential Formulation | 157 |
| 6 | Similarity of the Viscous Incompressible Fluid Flows | 159 |
| 6.1 | The Steady Flows Case | 162 |
| 7 | Flows With Low Reynolds Number. Stokes Theory | 164 |
| 7.1 | The Oseen Model in the Case of the Flows Past a Thin Profile | 167 |
| 8 | Flows With High (Large) Reynolds Number | 172 |
| 8.1 | Mathematical Model | 173 |
| 8.2 | The Boundary Layer Equations | 174 |
| 8.3 | Probabilistic Algorithm for the Prandtl Equations | 180 |
| 8.4 | Example | 187 |
| 8.5 | Dynamic Boundary Layer with Sliding on a Plane Plaque | 191 |
| 4. | INTRODUCTION TO NUMERICAL SOLUTIONS FOR ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS | 197 |
| 1 | Introduction | 197 |
| 2 | Discretization of a Simple Equation | 203 |
| 2.1 | Using the Finite Difference Method | 203 |
| 2.2 | Using the Finite Element Method | 203 |
| 2.3 | Using the Finite Volume Method | 205 |
| 2.4 | Comparison of the Discretization Techniques | 206 |
| 3 | The Cauchy Problem for Ordinary Differential Equations | 207 |
| 3.1 | Examples | 216 |
| 4 | Partial Differential Equations | 226 |
| 4.1 | Classification of Partial Differential Equations | 226 |
| 4.2 | The Behaviour of Different Types of PDE | 228 |
| 4.3 | Burgers' Equation | 231 |
| 4.4 | Stokes' Problem | 236 |
| 4.5 | The Navier–Stokes System | 239 |
| 5. | FINITE-DIFFERENCE METHODS | 247 |
| 1 | Boundary Value Problems for Ordinary Differential Equations | 247 |
| 1.1 | Supersonic Flow Past a Circular Cylindrical Airfoil | 249 |
| 2 | Discretization of the Partial Differential Equations | 253 |
| 3 | The Linear Advection Equation | 257 |
| 3.1 | Discretization of the Linear Advection Equation | 257 |

| | | |
|-----|--|-----|
| 3.2 | Numerical Dispersion and Numerical Diffusion | 264 |
| 3.3 | Lax, Lax–Wendroff and MacCormack Methods | 266 |
| 4 | Diffusion Equation | 272 |
| 4.1 | Forward-Time Scheme | 272 |
| 4.2 | Centered-Time Scheme | 273 |
| 4.3 | Backward-Time Scheme | 274 |
| 4.4 | Increasing the Scheme’s Accuracy | 275 |
| 4.5 | Numerical Example | 275 |
| 5 | Burgers Equation Without Shock | 277 |
| 5.1 | Lax Scheme | 277 |
| 5.2 | Leap-Frog Scheme | 278 |
| 5.3 | Lax–Wendroff Scheme | 279 |
| 6 | Hyperbolic Equations | 280 |
| 6.1 | Discretization of Hyperbolic Equations | 280 |
| 6.2 | Discretization in the Presence of a Shock | 285 |
| 6.3 | Method of Characteristics | 291 |
| 7 | Elliptic Equations | 295 |
| 7.1 | Iterative Methods | 296 |
| 7.2 | Direct Method | 304 |
| 7.3 | Transonic Flows | 307 |
| 7.4 | Stokes’ Problem | 312 |
| 8 | Compact Finite Differences | 320 |
| 8.1 | The Compact Finite Differences Method (CFDM) | 320 |
| 8.2 | Approximation of the Derivatives | 321 |
| 8.3 | Fourier Analysis of the Errors | 326 |
| 8.4 | Combined Compact Differences Schemes | 329 |
| 8.5 | Supercompact Difference Schemes | 333 |
| 9 | Coordinate Transformation | 335 |
| 9.1 | Coordinate Stretching | 338 |
| 9.2 | Boundary-Fitted Coordinate Systems | 339 |
| 9.3 | Adaptive Grids | 344 |
| 6. | FINITE ELEMENT AND BOUNDARY ELEMENT METHODS | 345 |
| 1 | Finite Element Method (FEM) | 345 |
| 1.1 | Flow in the Presence of a Permeable Wall | 349 |
| 1.2 | PDE-Toolbox of MATLAB | 354 |
| 2 | Least-Squares Finite Element Method (LSFEM) | 356 |
| 2.1 | First Order Model Problem | 356 |

| | | |
|-----|--|-----|
| 2.2 | The Mathematical Foundation of the Least-Squares Finite Element Method | 363 |
| 2.3 | Div-Curl (Rot) Systems | 370 |
| 2.4 | Div-Curl (Rot)-Grad System | 375 |
| 2.5 | Stokes' Problem | 377 |
| 3 | Boundary Element Method (BEM) | 380 |
| 3.1 | Abstract Formulation of the Boundary Element Method | 381 |
| 3.2 | Variant of the Complex Variables Boundary Element Method [112] | 385 |
| 3.3 | The Motion of a Dirigible Balloon | 389 |
| 3.4 | Coupling of the Boundary Element Method and the Finite Element Method | 391 |
| 7. | THE FINITE VOLUME METHOD AND THE GENERALIZED DIFFERENCE METHOD | 397 |
| 1 | ENO Finite Volume Schemes | 398 |
| 1.1 | ENO Finite Volume Scheme in One Dimension | 399 |
| 1.2 | ENO Finite Volume Scheme in Multi-Dimensions | 406 |
| 2 | Generalized Difference Method | 411 |
| 2.1 | Two-Point Boundary Value Problems | 411 |
| 2.2 | Second Order Elliptic Problems | 424 |
| 2.3 | Parabolic Equations | 429 |
| 2.4 | Application | 433 |
| 8. | SPECTRAL METHODS | 439 |
| 1 | Fourier Series | 442 |
| 1.1 | The Discretization | 442 |
| 1.2 | Approximation of the Derivatives | 445 |
| 2 | Orthogonal Polynomials | 447 |
| 2.1 | Discrete Polynomial Transforms | 447 |
| 2.2 | Legendre Polynomials | 450 |
| 2.3 | Chebyshev Polynomials | 452 |
| 3 | Spectral Methods for PDE | 455 |
| 3.1 | Fourier-Galerkin Method | 455 |
| 3.2 | Fourier-Collocation | 456 |
| 3.3 | Chebyshev-Tau Method | 457 |
| 3.4 | Chebyshev-Collocation Method | 458 |
| 3.5 | The Calculation of the Convolution Sums | 459 |
| 3.6 | Complete Discretization | 460 |

| | |
|--------------------------------------|-----|
| <i>Contents</i> | xi |
| 4 Liapunov–Schmidt (LS) Methods | 462 |
| 5 Examples | 472 |
| 5.1 Stokes’ Problem | 472 |
| 5.2 Correction in the Dominant Space | 479 |
| Appendix A | |
| Vectorial-Tensorial Formulas | 483 |
| References | 487 |
| Index | 497 |

Preface

The present book – through the topics and the problems approach – aims at filling a gap, a real need in our literature concerning CFD (Computational Fluid Dynamics). Our presentation results from a large documentation and focuses on reviewing the present day most important numerical and computational methods in CFD.

Many theoreticians and experts in the field have expressed their interest in and need for such an enterprise. This was the motivation for carrying out our study and writing this book. It contains an important systematic collection of numerical working instruments in Fluid Dynamics.

Our current approach to CFD started ten years ago when the University of Paris XI suggested a collaboration in the field of spectral methods for fluid dynamics. Soon after – preeminently studying the numerical approaches to Navier–Stokes nonlinearities – we completed a number of research projects which we presented at the most important international conferences in the field, to gratifying appreciation.

An important qualitative step in our work was provided by the development of a computational basis and by access to a number of expert softwares. This fact allowed us to generate effective working programs for most of the problems and examples presented in the book, an aspect which was not taken into account in most similar studies that have already appeared all over the world.

What makes this book special, in comparison with other similar enterprises?

This book reviews the main theoretical aspects of the area, emphasizes various formulations of the involved equations and models (focussing on optimal methods in CFD) in order to point out systematically the most utilized numerical methods for fluid dynamics. This kind of analysis – leaving out the demonstration details – takes notice of the convergence

and error aspects which are less prominent in other studies. Logically, our study goes on with some basic examples of effective applications of the methods we have presented and implemented (MATLAB).

The book contains examples and practical applications from fluid dynamics and hydraulics that were treated numerically and computationally – most of them having attached working programs. The inviscid and viscous, incompressible fluids are considered; practical applications have important theoretical outcomes.

Our study is not extended to real compressible fluid dynamics, or to turbulence phenomena. The attached MATLAB 6 programs are conceived to facilitate understanding of the algorithms, without optimizing intentions.

Through the above mentioned aspects, our study is intended to be an invitation to a more complete search: it starts with the formulation and study of mathematical models of fluid dynamics, continues with analysis of numerical solving methods and ends with computer simulation of the mentioned phenomena.

As for the future, we hope to extend our study and to present a new more complete edition, taking into account constructive suggestions and observations from interested readers.

We cannot end this short presentation without expressing our gratitude to our families who have supported us in creating this work in such a short time, by offering us peace and by acquitting us from our everyday duties.

The authors

Chapter 1

INTRODUCTION TO MECHANICS OF CONTINUA

1. Kinematics of Continua

1.1 The Concept of a Deformable Continuum

The fluids belong to *deformable continua*. In what follows we will point out the qualities of a material system to be defined as a deformable continuum.

Physically, a material system forms a *continuum* or a *continuum system* if it is “filled” with a continuous matter and every particle of it (irrespective how small it is) is itself a continuum “filled” with matter. As the matter is composed of molecules, the continuum hypothesis leads to the fact that a very small volume will contain a very large number of molecules. For instance, according to Avogadro’s hypothesis, 1cm^3 of air contains $2,687 \times 10^{19}$ molecules (under normal conditions). Obviously, in the study of continua (fluids, in particular) we will not be interested in the properties of each molecule at a certain point (the location of the molecule) but in the average of these properties over a large number of molecules in the vicinity of the respective point (molecule). In fact the association of these averaged properties at every point leads to the concept of continuity, synthesized by the following postulate which is accepted by us: “Matter is continuously distributed throughout the whole envisaged region with a large number of molecules even in the smallest (microscopically) volumes”.

Mathematically, a material system filling a certain region \mathcal{D} of the Euclidean tridimensional space is a continuum if it is a tridimensional material variety (vs. an inertial frame of reference) endowed with a specific measure called *mass*, mass which will be presumed to be absolutely continuous with regard to the volume of \mathcal{D} .

Axiomatically, the notion of mass is defined by the following axioms:

1) There is always an $m : \{\mathcal{M}\} \rightarrow \mathbb{R}_+$, i.e., an application which associates to a material system \mathcal{M} , from the *assembly of all material systems* $\{\mathcal{M}\}$, a real positive number $m(\mathcal{M})$ (which is also a state quantity joined to \mathcal{M}), called the *mass* of the system.

Physically, the association of this number $m(\mathcal{M})$ to a material system \mathcal{M} is made by scaling the physical mass of \mathcal{M} with the mass of another material system considered as unit (i. e. by *measurement*);

2) For any “splitting” of the material system \mathcal{M} in two disjoint subsystems \mathcal{M}_1 and \mathcal{M}_2 ($\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$ and $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$), the application m satisfies the *additivity property*, i.e., $m(\mathcal{M}) = m(\mathcal{M}_1) + m(\mathcal{M}_2)$.

This additivity property attributes to the *mass* application the quality of being a measure. Implicitly, the mass of a material system $m(\mathcal{M})$ is the sum of the masses dm of all the particles (molecules) which belong to \mathcal{M} , what could be written (by using the continuity hypothesis too) as

$$m(\mathcal{M}) = \int_{\mathcal{M}} dm,$$

the integral being considered in the Lebesgue sense;

3) For any material system \mathcal{M} , its mass $m(\mathcal{M})$ does not change during its evolution, i.e., it is constant and consequently $\dot{m} = 0$ (the universal principle of *mass conservation*).

Concerning the hypothesis of absolute continuity of the mass vis a vis the volume of the region \mathcal{D} occupied by the considered material system \mathcal{M} , this hypothesis obviously implies, besides the unity between the material system and the region “filled” by it, that the mass of any material subsystem $P \subset \mathcal{M}$ could become however small if the volume of the region $D \subset \mathcal{D}$, occupied by P , becomes, in its turn, sufficiently small (but never zero, i.e., the principle of the *indestructibility of matter* is observed). More, by accepting that the region \mathcal{D} and all its subregions D , are the closure of certain open sets which contain an infinity of fluid particles occupying positions defined by the corresponding position vectors \mathbf{r} (vs. the inertial frame) and additionally the boundaries of these sets are surfaces (in a finite number) with continuous normal, then according to the Radon–Nycodim theorem, there is a positive numerical function $\rho(\mathbf{r}, t)$, defined a.e. in \mathcal{D} , such that the mass of a part $P \subset \mathcal{M}$ can be expressed by

$$m(P) = \int_{\mathcal{M}} \rho(\mathbf{r}, t) dv,$$

The function $\rho(\mathbf{r}, t)$ is called the *density* or the *specific mass* according to its physical meaning. By using the above representation for the introduction of the density we overtake the shortcomings which could arise by the definition of $\rho(\mathbf{r}, t)$ as a point function through

$$\rho(\mathbf{r}, t) = \lim_{\substack{\text{vol}(D) \rightarrow 0 \\ \text{vol}(D) \neq 0}} \frac{m(P)}{\text{vol}(D)}$$

a definition which, from the medium continuity point of view, specifies ρ only at a discrete set of points.¹ Obvious, the acceptance of the existence of the density is a continuity hypothesis.

In the sequel, the region \mathcal{D} occupied by the continuum \mathcal{M} (and analogously D occupied by the part P) will be called either the *volume support* of \mathcal{D} , or the *configuration* at the respective moment in which the considered continuum appears.

The regularity conditions imposed on \mathcal{D} and on its boundary will support, in what follows, the use of the tools of the classical calculus (in particular the Green formulas).

Obviously, the continuum will not be identified with its volume support or its configuration. We will take for the continuum systems the topology of the corresponding volume supports (configurations), i.e., the topology which has been used in classical Newtonian mechanics. In particular, the distance between two particles of a continuum will be the Euclidean distance between the corresponding positions of the involved particles.

In the study of continua, in general, and of fluids, in particular, time will be considered as an absolute entity, irrespective of the state of the motion and of the fixed or mobile system of reference. At the same time the velocities we will deal with are much less than the velocity of light so that the relativistic effects can be neglected.

In the working space which is the tridimensional Euclidean space — space without curvature — one can always define a Cartesian inertial system of coordinates. In this space we can also introduce another system of coordinates without changing the basic nature of the space itself.

In the sequel, an infinitesimal volume of a continuum (i.e., with a sufficiently large number of molecules but with a mass obviously infinitesimal) will be associated to a geometrical point making a so-called *continuum particle*, a particle which is identified by an ordered triple

¹Since the function ρ defined by this limit cannot be zero or infinite (corresponding to the outside or inside molecule location of the point where the density is calculated), $\text{Vol}(D)$ can never be zero.

of numbers representing, in fact, the coordinates of the point (particle) within the chosen system. The synonymy between particle and material point (geometrical point endowed with an infinitesimal mass) is often used.

An important concept in the mechanics of continua will be that of a “*closed system*” or a “*material volume*”. A *material volume* is an arbitrary entity of the continuum of precise identity, “enclosed” by a surface also formed of continuum particles. All points of such a material volume, boundary points included, move with a respective local velocity, the material volume deforming in shape as motion progresses, with an assumption that there are no mass fluxes (transfers) *in* or *out* of the considered volume, i.e., the volume and its boundary are composed by the *same* particles all the time.

Finally, a continuum is said to be *deformable* if the distances between its particles (i.e., the Euclidean metric between the positions occupied by them) are changing during the motion as a reaction to the external actions. The liquids and gases, the fluids in general, are such deformable continua.

1.2 Motion of a Continuum. Lagrangian and Eulerian Coordinates

To define and make precise the motion of a continuum we choose both a rectangular Cartesian and a general curvilinear reference coordinate systems, systems which can be supposed inertial.

Let \mathbf{R} and \mathbf{r} be, respectively, the position vectors of the continuum particles, within the chosen reference frame, at the initial (reference) moment t_0 and at any (current) time t respectively. We denote by (X_i) and (x_i) , respectively, the coordinates of the two vectors in the rectangular Cartesian system while (X^i) and (x^i) will represent the coordinates of the same vectors in the general curvilinear (nonrectangular) system. Thus \mathbf{r} referring to a rectangular Cartesian system is $\mathbf{r} = x_1\mathbf{i}_1 + x_2\mathbf{i}_2 + x_3\mathbf{i}_3 \equiv x_k\mathbf{i}_k$, where any two repeated indices imply summation, and \mathbf{i}_k are the unit vectors along the x_k axes respectively. For a general system of coordinates (x^1, x^2, x^3) , the same position vector \mathbf{r} will be, in general, a nonlinear function $\mathbf{r}(x^i)$ of these coordinates. However its differential $d\mathbf{r}$ is expressible linearly in dx^i for all coordinates, precisely

$$d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial x^m} dx^m \equiv \mathbf{a}_m dx^m,$$

the vectors \mathbf{a}_m being called the *covariant base vectors*. Obviously if x^m are the Cartesian coordinates $x^m \equiv x_m$ and, implicitly, $\mathbf{a}_m \equiv \mathbf{i}_m$.

Let now χ_t be the mapping which associates to any particle P of the continuum \mathcal{M} , at any time t , a certain position \mathbf{r} obviously belonging to the volume support (configuration) \mathcal{D} , i.e., $P \xrightarrow{\chi_t} \mathbf{r}$. This mapping is called *motion*, the equation $\mathbf{r} = \chi(P, t)$ defining the motion of that particle. Obviously the motion of the whole continuum will be defined by the ensemble of the motions of all its particles, i.e., by the mapping $\chi(\mathcal{M}, t)$, $\chi : \mathcal{M} \xrightarrow{t} \mathcal{D}$, which associates to the continuum, at any moment t , its corresponding configuration.

The motion of a continuum appears then as a sequence of configurations at successive moments, even if the continuum cannot be identified with its configuration $\mathcal{D} = \chi(\mathcal{M}, t)$.

The mapping which defines the motion has some properties which will be made precise in what follows. But first let us identify the most useful choices of the independent variables in the study (description) of the continuum motions. They are the *Lagrangian coordinates (material description)* and the *Eulerian coordinates (spatial description)*.

Within the material description, the continuum particles are “identified” with their positions (position vectors) in a suitable *reference configuration* (like, for instance, the configuration at the initial moment t_0).² These positions in the reference configuration would provide the “fingerprints” of the continuum particle which at any posterior moment t , will be individualized through this position \mathbf{R} belonging to the reference configuration \mathcal{D}_0 .

Under these circumstances, due to the mentioned identification, the equation of the motion is

$$\mathbf{r} = \chi(\mathbf{R}, t), \quad (1.1)$$

the \mathbf{R} coordinates (X^i or X_i), together with t , representing the *Lagrangian* or *material coordinates*, through which all the other motion parameters can be expressed. Hence $\frac{d}{dt}\chi(\mathbf{R}, t)$ and $\frac{d^2}{dt^2}\chi(\mathbf{R}, t)$, with \mathbf{R} scanning the points of the domain \mathcal{D}_0 , will define the velocity field and the acceleration field respectively at the moment t .³

The equation of motion, for an \mathbf{R} fixed and t variable, defines the trajectory (path) of the particle P which occupied the position \mathbf{R} at the initial moment.

Finally, from the same equation of motion but for t fixed and \mathbf{R} variable in the configuration \mathcal{D}_0 , we will have that the corresponding \mathbf{r} is

²In the theory of elasticity one takes as reference configuration that configuration which corresponds to the natural (undeformed) state of the medium.

³We suppose the existence of these fields and their continuity except, possibly, at a finite number of points (surfaces) of discontinuity.

“sweeping” the current configuration (at the time t) $\mathcal{D} \equiv \chi(\mathcal{D}_0, t)$. In this respect (1.1) can be also understood as a mapping of the tridimensional Euclidean space onto itself, a mapping which depends continuously on $t \in \mathcal{T}$ and the motion of the continuum in the whole time interval \mathcal{T} will be defined by the vector function $\chi(\mathbf{R}, t)$ considered on $\mathcal{D}_0 \times \mathcal{T}$.

Now, one imposes some additional hypotheses for the above mapping joined to the equation of motion (1.1). These hypotheses are connected with the acceptance of some wider classes of real motions which confer their validity.

Suppose that \mathbf{r} is a vectorial function of class $C^2(\mathcal{D}_0)$ with respect to the \mathbf{R} components. This means that the points which were neighbours with very closed velocities and accelerations, at the initial moment, will remain, at any time t , neighbours with velocities and accelerations very closed too. Further, we presume that, at any moment t , there is a bijection between \mathcal{D}_0 and \mathcal{D} except, possibly, of some singular points, curves and surfaces. Mathematically this could be written through the condition that, at any time t , the mapping Jacobian $J = \det(\text{grad } \mathbf{r}) \neq 0$ a.e. in \mathcal{D} .

This last hypothesis linked to preserving the particles’ identity (they neither merge nor break) is also known as the *smoothness condition* or the *continuity axiom*. As from the known relation between the elemental infinitesimal volumes of \mathcal{D}_0 and \mathcal{D} , namely $dv = JdV$, one deduces, through $J \neq 0$, that any finite part of our continuum cannot have the volume (measure) of its support zero or infinite, the above hypothesis also implies the *indestructibility of matter principle*.

In the previous hypotheses it is obvious that (1.1) has, at any moment t , an inverse and consequently $\mathbf{R} = \chi^{-1}(\mathbf{r}, t) \in C^2(\dot{\mathcal{D}})$. Summarizing, in our hypotheses, the mapping (1.1) is a *diffeomorphism* between \mathcal{D}_0 and \mathcal{D} .

The topological properties of the mapping (1.1) lead also to the fact that, during the motion, the material varieties (i.e., the geometrical varieties “filled” with material points) keep their order. In other words, the material points, curves, surfaces and volumes don’t degenerate via motion; they remain varieties of the same order. The same topological properties imply that if $C_0(S_0)$ is a material closed curve (surface) in the reference configuration, then the image curve (surface) $C(S)$, at any current time t , will be also a closed curve (surface).

Further, if the material curves (surfaces) $C_0^{(1)}(S_0^{(1)})$ and $C_0^{(2)}(S_0^{(2)})$ are tangent at a point P_0 , then, at any posterior moment, their images will be tangent at the corresponding image point P , etc.

The material description, the adoption of the Lagrangian coordinates, is advisable for those motion studies when the displacements are small

and we may watch the whole motion of the individualized (by their positions in the reference configuration) particles.

In the case of fluids, in general, and of gases, in particular, the molecules are far enough apart that the cohesive forces are not sufficiently strong (in gases, for instance, an average separation distance between the molecules is of the order $3,5 \times 10^{-7} \text{cm}$). As a consequence to follow up such particles during their motion becomes a difficult task, the corresponding displacements being very large (a gas sprayed inside “fills” immediately the respective room).

That is why for fluids, in general, and for gases, in particular, another way to express the parameters of the motion, to choose the independent variable, should be considered. This new type of motion description is known as the *spatial* or *Eulerian description*, the corresponding variables being the *spatial* or *Eulerian coordinates*.

Precisely, as Eulerian coordinates (variables) the components of \mathbf{r} (x_i or x^i) and t are to be considered. In other words, in this description, we focus not on the continuum particles themselves but on their position in the current configuration and we determine the motion parameters of those particles (not the same !) which are locating at the respective positions at that time. Thus to know $\mathbf{v} = \mathbf{v}(\mathbf{r}, t)$, for a fixed \mathbf{r} at $t \in \mathcal{T}$, means to know the velocities of *all* the particles which, in the considered interval of time, pass through the position defined by \mathbf{r} . On the other hand, if we know the velocity field $\mathbf{v}(\mathbf{r}, t)$ on $\mathcal{D} \times \mathcal{T}$, by integrating the differential equation $\frac{d\mathbf{r}}{dt} = \mathbf{v}(\mathbf{r}, t)$, with initial conditions (assuming that the involved velocity field is sufficiently smooth to ensure the existence and uniqueness of the solution of this Cauchy problem) one gets $\mathbf{r} = \chi(\mathbf{R}, t)$, which is just the equation of motion (1.1) from the material (Lagrangian) description. Conversely, starting with (1.1) one could immediately set up $\mathbf{v}(\mathbf{r}, t)$, etc., which establishes the complete equivalence of the two descriptions.

In what follows we calculate the time derivatives of some (vectorial or scalar) fields f expressed either in Lagrangian variables ($f(\mathbf{R}, t)$) or in Eulerian variables ($f(\mathbf{r}, t)$).

In the first case $\dot{f} = \frac{\partial f}{\partial t}$ and this derivative is called a *local* or *material derivative*. Obviously, in this case, $\mathbf{v} = \frac{\partial \mathbf{r}}{\partial t}(\mathbf{R}, t)$ and $\mathbf{a} = \frac{\partial^2 \mathbf{r}}{\partial t^2}(\mathbf{R}, t)$.

But, in the second case, we have $\dot{f} = \frac{\partial f}{\partial t} + (\mathbf{v} \cdot \nabla)f$, where ∇ is, in Cartesian coordinates, the differential operator $\nabla \equiv \text{grad} \equiv \frac{\partial}{\partial x_i} \mathbf{i}_i$. This derivative is designed to be the *total* or *spatial* or *substantive derivative* or *the derivative following the motion*. In particular $\mathbf{a}(\mathbf{r}, t) = \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v}$.

Stokes has denoted this total derivative by $\frac{D}{Dt}$, the operator $\frac{D}{Dt}$ being equal to $\frac{\partial}{\partial t} + \mathbf{v} \cdot \text{grad} = \frac{\partial}{\partial t} + [\text{grad}(\cdot)] \cdot \mathbf{v}$, due to the obvious equality $(\mathbf{v} \cdot \text{grad}) \mathbf{v} = (\text{grad} \mathbf{v}) \cdot \mathbf{v}$, $\mathbf{v} \cdot \text{grad}$ or $[\text{grad}(\cdot)] \cdot \mathbf{v}$ being the so-called convective part of $\frac{D}{Dt}$.

When all the motion parameters, expressed in Eulerian coordinates, do not depend explicitly on time, the respective motion is called *steady* or *permanent*. Obviously, the steady condition is $\frac{\partial}{\partial t} \equiv 0$ or, equivalently, $\frac{D}{Dt} = \mathbf{v} \cdot \text{grad}$.

Conversely, if time appears explicitly, the motion is *unsteady* or *non-permanent*.

Before closing this section we should make precise the notions of trajectories (pathlines), streamlines and streamsurfaces, vortex lines and vortex tubes, circulation and the concept of stream function as well.

1.2.1 Trajectories

In general the *trajectory* (*pathline*) is the locus described by a material point (particle) during its motion. The trajectories will be the integral curves (solutions) of the system

$$\frac{dx_1}{v_1} = \frac{dx_2}{v_2} = \frac{dx_3}{v_3} = dt \quad (\text{in Cartesian coordinates})$$

or of the system

$$\frac{dx^1}{v^1} = \frac{dx^2}{v^2} = \frac{dx^3}{v^3} = dt \quad (\text{in curvilinear coordinates}),$$

where $\mathbf{v} = v_k(x_i, t) \mathbf{i}_k = v^k(x^i, t) \mathbf{a}_k$, v^k being the so-called contravariant components of the velocity \mathbf{v} in the covariant base vectors \mathbf{a}_k of the considered curvilinear system.

Obviously, at every point of a trajectory the velocity vector is necessarily tangent to the trajectory curve. At the same time we will suppose again the regularity of the velocity field $\mathbf{v}(\mathbf{r}, t)$, to ensure the existence of the solution of the above system (in fact the vectorial equation $\frac{d\mathbf{r}}{dt} = \mathbf{v}(\mathbf{r}, t)$). A detailed study of this system, even in the case when some singular points occur (for instance, the “stagnation points” where $\mathbf{v}(\mathbf{r}, t) = 0$), has been done by Lichtenstein [84].

1.2.2 Streamlines and Streamsurfaces

For a *fixed* time t , the *streamlines* and the *streamsurfaces* are the curves and, respectively, the surfaces in the motion field on which the velocity vector is tangent at every point of them. A streamsurface could be considered as a locus of streamlines.

The definition of streamlines (tangency condition) implies that the streamlines should be the integral curves of the differential system

$$\frac{dx_1}{v_1} = \frac{dx_2}{v_2} = \frac{dx_3}{v_3} \quad (\text{in Cartesian coordinates})$$

or

$$\frac{dx^1}{v^1} = \frac{dx^2}{v^2} = \frac{dx^3}{v^3} \quad (\text{in curvilinear coordinates}),$$

where the time t , which appears explicitly in $v_i(x_i, t)$ or $v^i(x^i, t)$, has to be considered as a parameter with a fixed value.

At every fixed moment, the set of the streamlines constitutes the *motion pattern (spectrum)*. These motion patterns are different at different times.

When the motion is steady, the motion spectrum (pattern) is fixed in time and the pathlines and streamlines are the same, the definable differential system becoming identical. This coincidence could be realized even for an unsteady motion provided that the restrictive condition $\mathbf{v} \times \frac{\partial \mathbf{v}}{\partial t} = 0$ is fulfilled. This result can be got, for instance, from the so-called Helmholtz–Zorawski⁴ criterion which states that a necessary and sufficient condition for the lines of a vectorial field $\mathbf{c}(\mathbf{r}, t)$ to become material curves (i.e., locus of material points) is

$$\mathbf{c} \times \left[\frac{\partial \mathbf{c}}{\partial t} + \text{rot}(\mathbf{c} \times \mathbf{v}) + \mathbf{v} \text{div} \mathbf{c} \right] = 0,$$

Identifying $\mathbf{c} \equiv \mathbf{v}$ we get the necessary and sufficient condition that the lines of the \mathbf{v} field (i.e., the streamlines) become material curves (i.e., trajectories), precisely $\mathbf{v} \times \frac{\partial \mathbf{v}}{\partial t} = 0$.

A *stream tube* is a particular streamsurface made by streamlines drawn from every point of a simple closed curve. A stream tube of infinitesimal cross section is called a *stream filament*.

1.2.3 Vortex Lines and Vortex Surfaces

By curl or vorticity or rotation we understand the vector $\boldsymbol{\omega} = \nabla \times \mathbf{v} = \text{rot} \mathbf{v}$. The rationale for such a definition is the fact that, at every point of the continuum motion, the particles rotate about an instantaneous axis and the vector $\boldsymbol{\omega}$ has the direction of this axis, the value of the rotation being also $\frac{1}{2}\boldsymbol{\omega}$.

⁴See [33]

For a fixed time t , by a *vortex (vorticity, rotation) line (surface)* we understand those curves (surfaces) whose tangents, at every point of them, are directed along the local vorticity (curl, rotation) vector.

Of course the particles distributed along a vortex line rotate about the tangents to the vortex line at their respective positions.

A *vortex (vorticity, rotation) tube* is a vortex surface generated by vortex lines drawn through each point of an arbitrary simple closed curve (there is a diffeomorphism between the continuum surface enclosed by this simple curve and the circular disk).

If the vortex tube has a very small (infinitesimal) sectional area it is known as a *vortex filament*.

1.2.4 **Circulation**

The *circulation* along an arc AB is the scalar $\Gamma(AB) = \int_{AB} \mathbf{v} \cdot d\mathbf{r}$. The following result is a direct consequence of the Stokes theorem [110]⁵: “The circulation about two closed contours on a vortex tube at a given instant t , — closed contours which lie on the vortex tube and encircle it once, in the same sense — are the same” (this result of pure kinematic nature is known as the “first theorem of Helmholtz”).

The invariance of the circulation vis-a-vis the contour C which encircles once the vortex tube supports the introduction of the concept of *the strength of the vortex tube*. More precisely, this strength would be the circulation along the closed simple contour (C) which encircles once, in a direct sense, the tube.

The constancy of this circulation, which is equal to the rotation flux through the tube section bounded by the contour (C), leads to the fact that, within a continuum, both vortex and filament lines cannot “end” (the vanishing of the area bounded by (C) or of the vortex would imply, respectively, the unboundedness of the vorticity or the mentioned area, both cases being contradictions).

That is why the vortex lines and filaments either form rings in our continuum or extend to infinity or are attached to a solid boundary. (The smoke rings from a cigarette make such an example).⁶

⁵The circulation of a vector \mathbf{u} , from a continuous derivable field, along the simple closed contour (L), is equal to the flux of $\text{rot } \mathbf{u}$ through a surface (Σ) bounded by (L), i.e. $\int_L \mathbf{u} \cdot d\mathbf{r} = \iint_{(\Sigma)} \text{rot } \mathbf{u} \cdot \mathbf{n} d\sigma$, provided that the reference frame (system), made by the positively

oriented tangent at a point $P \in (L)$, the outward normal \mathbf{n} to (Σ) at a point M and the vector \mathbf{MP} , for any points M and P , is a right-handed system.

⁶For a *line vortex* (which is distinct from a vortex line and which is a mathematical idealization of a vortex filament assumed to converge onto its axis, i.e. a vortices locus) the same assertion, often made, is false ($\text{rot } \mathbf{v}$ could have zeros within the continuum in motion!)

Obviously, of great interest is how the circulation along a material closed simple contour changes while the contour moves with the continuum. To analyze this aspect let us evaluate $\frac{D}{Dt} \int_A^B \mathbf{v} \cdot d\mathbf{r}$, i.e., the rate of change (in time) of the circulation about a material contour joining the points A and B as it moves with the medium. Considering then $\mathbf{r} = \mathbf{r}(s, t)$, for $0 \leq s \leq l$, we have

$$\begin{aligned} \frac{D}{Dt} \int_A^B \mathbf{v} \cdot d\mathbf{r} &= \frac{D}{Dt} \int_0^l \mathbf{v} \cdot \frac{d\mathbf{r}}{ds} ds = \int_0^l \frac{D}{Dt} \left(\mathbf{v} \cdot \frac{d\mathbf{r}}{ds} \right) ds \\ &= \int_A^B \frac{D\mathbf{v}}{Dt} \cdot d\mathbf{r} + \int_A^B \mathbf{v} \cdot d\mathbf{v} = \int_A^B \mathbf{a} \cdot d\mathbf{r} + \frac{1}{2} (v_B^2 - v_A^2), \end{aligned}$$

where $v = |\mathbf{v}|$. If A and B coincide so as to form a simple closed curve (C) in motion, obviously $\frac{D}{Dt} \oint_C \mathbf{v} \cdot d\mathbf{r} = \oint_C \mathbf{a} \cdot d\mathbf{r}$, i.e., the rate of change of circulation of velocity is equal to the circulation of acceleration along the same closed contour (C). If the acceleration comes from a potential, i.e., $\mathbf{a} = \text{grad } U$, then the circulation of the velocity along the closed contour does not change as the curve moves, the respective motion being called circulation preserving.

For the fluids, under some additional hypotheses a very important result connected with the circulation conservation will be given later on (the Thompson Kelvin theorem).

1.2.5 Stream Function for Plane and Axially Symmetric (Revolution) Motions

By extending the already given kinematic definition to the dynamics case, a motion is supposed to be steady (permanent) if all the (kinematic, kinetic, dynamic) parameters characterizing the medium state and expressed with Euler variables x_1, x_2, x_3, t , are not (explicitly) dependent on t .

All the partial time derivatives of the mentioned parameters being zero ($\frac{\partial}{\partial t} \equiv 0$), we have (from the continuity equation) that $\text{div}(\rho\mathbf{v}) = 0$, i.e., the vector field $\rho\mathbf{v}$ is conservative (solenoidal).

The above equation allows us to decrease the number of the unknown functions to be determined; we will show that in the particular, but ex-

tremely important case, of the plane and axially symmetric (revolution) motions.

A continuum motion is said to be *plane*, parallel with a fixed plane (P), if, at any moment t , the velocity vector (together with other vectors which characterize the motion) is parallel with the plane (P) and all the mechanical (scalar or vectorial) parameters of the motion are invariant vs. a translation normal to (P). We denote by x and y the Cartesian coordinates in (P) so that $x_1 = x$, $x_2 = y$, the variable x_3 not playing a role. In the same way, we denote $v_1 = u$, $v_2 = v$, ($v_3 = 0$), \mathbf{k} being the unit vector normal to (P) and oriented as x_3 axis.

One says that a motion is *axially symmetric* vs. the fixed axis Ox , if, at any moment t , the velocity vector's supports (and of supports of other vectors characterizing the motion) intersect the Ox axis and all the mechanical parameters associated to the motion are rotation (vs. Ox) invariants. We denote by Ox and Oy the orthogonal axes in a meridian half-plane (bounded by Ox), by \mathbf{k} the unit vector which is directly orthogonal to Ox and Oy and by u and v the respective components of the vectors \mathbf{v} obviously located in this half-plane.

Now let be, at a fixed instant t , a contour (C) drawn in Oxy and let (Σ_C) be the corresponding surface generated by:

- a) a translation motion, parallel to \mathbf{k} and of unit amplitude, in the case of plane motions or
- b) an Ox -rotation motion of a 2π -amplitude, in the case of revolution (axially symmetric) motions.

Let m be a number which equals 0, in the case of a plane motion and equals 1, in the case of a revolution motion. Hence

$$\iint_{(\Sigma_C)} \rho \mathbf{v} \cdot \mathbf{n} d\sigma = \int_{(C)} (2\pi y)^m \rho \mathbf{v} \cdot \mathbf{n} ds = \int_{(C)} (2\pi y)^m \rho (udy - vdx),$$

(with the remark that $d\sigma = 2\pi y ds$), the (C) orientation being that obtained by a rotation from \mathbf{n} with $+\frac{\pi}{2}$ and ds is the elemental arc length on (C).

If the motion is steady⁷ and (C) is a closed curve bounding the area (σ) from Oxy , the above expressions vanish⁸ and, by using the divergence (Green) theorem, we get

⁷The result keeps its validity even for unsteady motion provided that the continuum is incompressible; in these hypotheses the function ψ which will be introduced in the sequel, depends on the time t too.

⁸We have an exact total differential due to the condition $div(\rho \mathbf{v}) = 0$.

$$\iint_{(\sigma)} \left[\frac{\partial}{\partial x} (\rho y^m u) + \frac{\partial}{\partial y} (\rho y^m v) \right] dx dy = 0$$

for any (σ) of Oxy . Following the fundamental lemma (given by the end of the next section) we could write

$$\frac{\partial}{\partial x} (\rho y^m u) + \frac{\partial}{\partial y} (\rho y^m v) = 0,$$

a relation which is equivalent with the above continuity equation for the plane or axially symmetric motions.

As the last relation expresses that $\rho y^m (udy - vdx)$ is an exact total differential, there is a function $\rho_0 \psi(x, y)$ (ρ_0 being a positive constant), defined within an arbitrary additive constant, such that

$$\rho y^m (udy - vdx) = \rho_0 d\psi,$$

i.e., we can write

$$u = \frac{\rho_0}{\rho y^m} \frac{\partial \psi}{\partial y}, v = -\frac{\rho_0}{\rho y^m} \frac{\partial \psi}{\partial x}$$

and hence

$$\mathbf{v} = -\frac{\mathbf{k}\rho_0}{\rho y^m} \times \text{grad } \psi.$$

The function $\psi(x, y)$ is, by definition, the *stream function* of the considered steady (plane or axially symmetric) motion.

The above formulas show that the unknown functions u and v could be replaced by the unique unknown function ψ . The curves $\psi = \text{const}$ are the streamlines in Oxy . Generally, (C) being an arc joining the points A and B from the same plane, $(2\pi)^m \rho_0 [\psi(B) - \psi(A)]$ represents the mass flow rate through (Σ_C) , the sense of \mathbf{n} along (C) being determined by the $-\frac{\pi}{2}$ rotation of the (C) tangent (oriented from A to B).

1.3 Euler–Lagrange Criterion. Euler’s and Reynolds’ (Transport) Theorems

Let us consider a material volume (closed system) $\mathcal{D}(t)$ whose surface $\mathcal{S}(t)$ is formed of the same particles which move with the local continuum velocity being thus a material surface. We intend to obtain a necessary and sufficient condition, for an arbitrary boundary surface $\mathcal{S}(t)$ of equation $f(\mathbf{r}, t) = 0$, to be a material surface, i.e., to be, during the motion, a collection of the same continuum particles of fixed identity.

Following Kelvin, if a material point (particle) belonging to $\mathcal{S}(t)$ moves along the unit external normal $\mathbf{n} = \frac{\text{grad}f}{|\text{grad}f|}$, with a velocity u_n , then its infinitesimal displacement $\delta\mathbf{r}$, in an infinitesimal interval of time $(t, t + \delta t)$, will be $\delta\mathbf{r} = \mathbf{n}u_n\delta t$. As this particle should remain on $\mathcal{S}(t)$ (to be a material surface) we would obviously have $f(\mathbf{r} + \delta\mathbf{r}, t + \delta t) = 0$. Keeping only the first two terms of the Taylor's expansion which is backed by the infinitesimal character of the displacement $\delta\mathbf{r}$ (and correspondly of the time δt), we get

$$\frac{\partial f}{\partial t} + u_n(\mathbf{n} \cdot \text{grad}f) = 0.$$

But, on the other side, any material point (particle) of the surface $\mathcal{S}(t)$ should move with the continuum velocity at that point, i.e., necessarily, $u_n = \mathbf{v} \cdot \mathbf{n}$ and thus we get the necessary condition

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \text{grad}f \equiv \frac{Df}{Dt} = 0.$$

To prove also the sufficiency of this condition we should point out that (for instance) if this condition is fulfilled, then there will be at the initial moment a material surface \mathcal{S}_0 , such that our surface $\mathcal{S} \equiv \chi(\mathcal{S}_0, t)$, i.e., it is the image of \mathcal{S}_0 , through the motion mapping at the instant t . But then, due to the conservation theorem of material surfaces, it comes out immediately that $\mathcal{S}(t)$ should be a material surface.

Now let us attach to the first order partial differential equation $\frac{\partial f}{\partial t} + v_i \frac{\partial f}{\partial x_i} = 0$ its characteristic system, i.e., let us consider the differential system

$$\frac{dx_1}{v_1} = \frac{dx_2}{v_2} = \frac{dx_3}{v_3} = dt,$$

It is known that if $\varphi_\alpha(\mathbf{r}, t) = X_\alpha$, X_α being constants ($\alpha = 1, 2, 3$), is a fundamental system of first integrals of our characteristic differential system, the general solution of the above partial differential equation is $f = \Phi(\varphi_1, \varphi_2, \varphi_3) = \Phi(X_1, X_2, X_3)$, where Φ is an arbitrary function of class C^1 . But, then, the particles of coordinates X_α ($\alpha = 1, 2, 3$) which fulfil the equation $\Phi(X_1, X_2, X_3) = 0$ will also fulfil $f = 0$, i.e., at the time t , they will be on the material surface of equation $f = 0$ (in other words, the surface $\mathcal{S}(t)$ is the image, at the moment t , of the material surface $\Phi(X_1, X_2, X_3) = 0$ from a reference configuration).

This result, which gives the necessary and sufficient condition for an (abstract) surface to be material is known as the *Euler–Lagrange criterion*.

Obviously a rigid surface Σ (for instance a wall), which is in contact with a moving continuum, is a particles locus i.e., it is a material surface. Using the above criterium we will have, on such a surface of equation $f(\mathbf{r}, t) = 0$, the *necessary condition* $\mathbf{v} \cdot \mathbf{n}|_{\Sigma} = - \frac{-\partial_t f}{|\text{grad} f|} \Big|_{\Sigma}$ and when the rigid surface is fixed then, $\mathbf{v} \cdot \mathbf{n}|_{\Sigma} = 0$, so that the continuum velocity is tangent at this surface.

The *Euler theorem* establishes that the total derivative of the motion Jacobian $J = \det(\text{grad } \mathbf{r})$, is given by $\dot{J} = J \text{div } \mathbf{v}$.

The proof of this result uses the fact that the derivative of a determinant J is the sum of the determinants J_i which are obtained from J by the replacement of the “ i ” line with that composed by its derivative vs. the same variable.

In our case, for instance,

$$J_1 = \begin{vmatrix} \frac{\partial v_1}{\partial X_1} & \frac{\partial v_1}{\partial X_2} & \frac{\partial v_1}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{vmatrix} = \begin{vmatrix} \frac{\partial v_1}{\partial x_j} \frac{\partial x_j}{\partial X_1} & \frac{\partial v_1}{\partial x_j} \frac{\partial x_j}{\partial X_2} & \frac{\partial v_1}{\partial x_j} \frac{\partial x_j}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{vmatrix} \\ = \frac{\partial v_1}{\partial x_j} J_{1j} = \frac{\partial v_1}{\partial x_1} J,$$

because $J_{11} = J$ and $J_{12} = J_{13} = 0$.

Hence, by identical assessments of J_2 and J_3 , we get the result we were looking for $\dot{J} = J \text{div } \mathbf{v}$. Using this result together with the known relation between the elemental infinitesimal volumes from \mathcal{D} and \mathcal{D}_0 , i.e., $dv = JdV$, we can calculate the total derivative of the elemental infinitesimal volume, at the moment t (that means from \mathcal{D}). Precisely we have

$$\dot{dv} = \dot{J}dV + J\dot{dV} = J\text{div } \mathbf{v}dV = dv\text{div } \mathbf{v}$$

(dV being fixed in time).

Reynolds' (transport) theorem is a quantizing of the rate of change of an integral of a scalar or vectorial function $F(\mathbf{r}, t)$, integral evaluated on a material volume $\mathcal{D}(t)$. As the commutation of the operators of total time derivative and of integration will not be valid any more, the integration domain depending explicitly on time, we have to consider, first, a change of variables which replaces the integral material volume $\mathcal{D}(t)$, depending on time, by a fixed integral domain \mathcal{D}_0 and so the

derivative operator could then commute with that of integration. More precisely we will perform the change of variables given by the equation of motion expressed in Lagrangian coordinates, i.e., $\mathbf{r} = \chi(\mathbf{R}, t)$, the new integration domain becoming the fixed domain \mathcal{D}_0 from the initial configuration and then we could come back to the current domain $\mathcal{D}(t)$.

More exactly, taking into account both Euler's and Green's theorems we have

$$\begin{aligned} \frac{D}{Dt} \int_{\mathcal{D}(t)} F(\mathbf{r}, t) dv &= \frac{D}{Dt} \int_{\mathcal{D}_0} F(\chi(\mathbf{R}, t)) J dV = \frac{D}{Dt} \int_{\mathcal{D}_0} (\dot{F} J + F \dot{J}) dV \\ &= \int_{\mathcal{D}(t)} (\dot{F} + F \operatorname{div} \mathbf{v}) dv = \int_{\mathcal{D}(t)} \left(\frac{\partial F}{\partial t} + \mathbf{v} \operatorname{grad} F + F \operatorname{div} \mathbf{v} \right) dv \\ &= \int_{\mathcal{D}(t)} \left[\frac{\partial F}{\partial t} + \operatorname{div}(F \mathbf{v}) \right] dv = \int_{\mathcal{D}(t)} \frac{\partial F}{\partial t} dv + \int_{\mathcal{S}(t)} F \mathbf{v} \cdot \mathbf{n} d\sigma, \end{aligned}$$

where \mathbf{n} is the unit external normal.

This transport formula will be useful in establishing the equations of motion for continua (under the so-called conservation form).

Analogously, one establishes equivalent formulas for the total derivatives of the curvilinear or surface integrals when the integration domains depend upon time.

Thus

$$\frac{D}{Dt} \int_{\mathcal{S}(t)} \mathbf{F} \cdot \mathbf{n} d\sigma = \int_{\mathcal{C}(t)} \left[\frac{\partial \mathbf{F}}{\partial t} + \operatorname{rot}(\mathbf{F} \times \mathbf{v}) + \mathbf{v} \operatorname{div} \mathbf{F} \right] \cdot \mathbf{n} d\sigma,$$

where $\mathcal{C}(t)$ is the contour enclosing the surface [52].

From this formula comes the necessary and sufficient condition for the flux of a field \mathbf{F} , through a material surface $\mathcal{S}(t)$, to be constant, which condition is

$$\frac{\partial \mathbf{F}}{\partial t} + \operatorname{rot}(\mathbf{F} \times \mathbf{v}) + \mathbf{v} \operatorname{div} \mathbf{F} = 0 \quad (\text{Zorawski condition}).$$

In the formulation of the general principles of the motion equations under a differential form (usually nonconservative), an important role is taken by the following

LEMMA: *Let $\varphi(\mathbf{r})$ be a scalar function defined and continuous in a domain \mathcal{D} and let D be an arbitrary subdomain of \mathcal{D} . If $\int_D \varphi(\mathbf{r}) dv = 0$, for every subdomain $D \subset \mathcal{D}$, then the function $\varphi(\mathbf{r}) \equiv 0$ in \mathcal{D} .*

The proof is immediate by using "reductio ad absurdum" and the continuity of φ [110].

The result is still valid even in the case when instead of the scalar function φ , a vectorial function of the same \mathbf{r} is considered (it is sufficient to use the previous assertion on each component). At the same time the conclusion will remain the same if the above condition takes place only on a set of subdomains (E) with the property that in any neighborhood of a point from \mathcal{D} , there is at least a subdomain from the set (E).

2. General Principles. The Stress Tensor and Cauchy's Fundamental Results

2.1 The Forces Acting on a Continuum

Let us consider a material subsystem P of the continuum \mathcal{M} , a subsystem imagined at a given moment in a certain configuration $D = \chi(P, t)$, which is enclosed in the volume support \mathcal{D} of the whole system \mathcal{M} . On this subsystem P of the continuum \mathcal{M} , two types of actions are exerted:

(i) *contact (surface) actions*, of local (molecular) nature, exerted on the surface S of the support D of the subsystem P by the “complementary” system $\mathcal{M} \setminus P$ (as the “pressure or pull” of the boundary, the “pushing” action through friction on the boundary, etc.)

(ii) *distance (external) actions*, of an extensive character, exerted on the bulk portions of the continuum P and arising due to some external cause (such as gravity, electromagnetic, centrifugal actions, etc.)

But the mechanics principles are formulated, all of them, in the language of forces and not of actions. To “translate” the above mentioned actions into a sharp language of forces we will introduce the so-called *Cauchy's Principle (Postulate)* which states:

“Upon the surface S there exists a distribution of contact forces, of density \mathbf{T} , whose resultant and moment resultant are equipollent to the whole contact action exerted by $\mathcal{M} \setminus P$.”

At the same time there is a distribution of external body or volume forces of density \mathbf{f} , exerted on the whole P or D and whose resultant and moment resultant are completely equivalent (equipollent) with the whole distance (external) action exerted on P .”

The contact forces introduced by this principle are called *stresses*. These stresses, of surface density \mathbf{T} , at a certain moment t , will depend upon the point where they are evaluated and the orientation of the surface element on which this point is considered, orientation characterized by the outward normal unit vector \mathbf{n} on this surface, such that $\mathbf{T} = \mathbf{T}(\mathbf{r}, \mathbf{n}, t)$.

Concerning the external body or volume forces (the gravity forces are body forces while the electromagnetic forces are volume forces, etc.), of density \mathbf{f} , at a certain time t , they depend only on the position vector \mathbf{r}

of the point of application, i.e., $\mathbf{f} = \mathbf{f}(\mathbf{r}, t)$. To avoid ambiguity we will suppose, in this sequel, that all the external forces we work with are body forces (gravity forces being the most important in our considerations). To postulate the existence of the densities \mathbf{T} and \mathbf{f} (continuity hypotheses) is synonymous with the acceptance of the absolute continuity of the whole contact or external (body) actions with respect to the area or the mass respectively. Then, by using the same Radon–Nycodim theorem, the total resultant of the stresses and of the external body forces could be written

$$\mathbf{R}^C = \int_S \mathbf{T} da, \quad \mathbf{R}^d = \int_P \mathbf{f} dm = \int_D \rho \mathbf{f} dv,$$

representations which are important in the general principles formulation.

In the sequel we will formulate the general principles for continua by expressing successively, in mathematical language, the three basic physical principles:

- (i) mass is never created or destroyed (mass conservation);
- (ii) the rate of change of the momentum tursor is equal to the tursor of the direct exerted forces (Newton's second law);
- (iii) energy is never created or destroyed (energy conservation).

2.2 Principle of Mass Conservation. The Continuity Equation

Mass conservation, postulated by the third axiom of the definition of the mass, requires that the mass of *every* subsystem $P \subset \mathcal{M}$ remains constant during motion. Evaluating this mass when the subsystem is located in both the reference configuration (i.e., for $t = 0$) D_0 and the current configuration at the moment t , mass conservation implies that

$$m(P) = \int_{D_0} \rho_0(\mathbf{R}_0) dV = \int_D \rho(\mathbf{r}, t) dv = \int_{D_0} \rho[\chi(\mathbf{R}, t)] J dV,$$

the last equality being obtained by reverting to the current reference configuration.

In the continuity hypothesis of continuum motion ($\rho, \mathbf{v} \in C^1$), as the above equalities hold for every subsystem P (and so for every domain D_0), the fundamental lemma, from the end of sub-section 1.1, leads to

$$\rho_0(\mathbf{R}) = \rho(\chi(\mathbf{R}, t)) J$$

which represents the *equation of continuity* in Lagrangian coordinates.

In spatial (Eulerian) coordinates, by making explicit the third axiom from the mass definition, i.e., $\dot{m} = 0$, we get

$$0 = \dot{m}(P) = \frac{D}{Dt} \int_D \rho(\mathbf{r}, t) dv = \int_D (\dot{\rho} + \rho \operatorname{div} \mathbf{v}) dv,$$

where the Reynolds transport theorem has been used. Backed by the same fundamental lemma, the following forms of the continuity equation can also be obtained:

$$\dot{\rho} + \rho \operatorname{div} \mathbf{v} = 0 \quad (\text{the nonconservative form})$$

or

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0 \quad (\text{the conservative form}).$$

We remark that if in the theoretical dynamics of fluids, the use of nonconservative or conservative form does not make a point, in the applications of computational fluid dynamics it is crucial which form is considered and that is why we insist on the difference between them.

2.2.1 Incompressible Continua

A continuum system is said to be incompressible if the volume (measure) of the support of any subsystem of it remains constant as the continuum moves.

By expressing the volume (measure) of the arbitrary system P at both the initial and the current moment, we have

$$\int_{D_0} dV = \int_D dv = \int_{D_0} J dV,$$

i.e., the incompressibility, in Lagrangian coordinates, implies that $J = 1$ and consequently the equation of continuity becomes

$$\rho_0(\mathbf{R}) = \rho(\chi(\mathbf{R}, t)).$$

We can arrive at the same result, in Eulerian coordinates, if we write

$$0 = \frac{D}{Dt} \int_D dv = \int_D (\dot{1} + \operatorname{div} \mathbf{v}) dv,$$

which leads to $\operatorname{div} \mathbf{v} = 0$ and, from the continuity equation, to $\frac{D\rho}{Dt} = 0$. We conclude that for incompressible continua, the (mass) density

remains constant as the particles are followed while they move (i.e., on any pathline), but the value of this constant could be different from trajectory to trajectory.

If the medium is *homogeneous*, i.e., ρ is constant with respect to the spatial variables, then it is incompressible if and only if ρ is constant vs. the time too.

We note that if a continuum is homogeneous at the moment $t = 0$, it could become nonhomogeneous later on. In fact a continuum remains homogeneous if and only if it is incompressible.

Within this book we will deal only with incompressible homogeneous media (continua).

2.3 **Principle of the Momentum Torsor Variation. The Balance Equations**

According to this principle of mechanics, applied within continua for any material subsystem $P \subset \mathcal{M}$, at any configuration of it $D \equiv \chi(P, t)$, the time derivative of the momentum torsor equals the torsor of the (direct) acting forces.

As the torsor is the pair of the resultant and the resultant moment, while the (linear) momentum of the subsystem P is $\mathbf{H}(P) = \int_P \mathbf{v} dm = \int_D \rho \mathbf{v} dv$ and the angular (kinetic) momentum is $K_0(P) = \int_P \mathbf{r} \times \mathbf{v} dm = \int_D \mathbf{r} \times \rho \mathbf{v} dv$ (O being an arbitrary point of E_3), the stated principle can be written as

$$\frac{D}{Dt} \int_D \rho \mathbf{v} dv = \int_S \mathbf{T} da + \int_D \rho \mathbf{f} dv,$$

respectively

$$\frac{D}{Dt} \int_D \mathbf{r} \times \rho \mathbf{v} dv = \int_S \mathbf{r} \times \mathbf{T} da + \int_D \mathbf{r} \times \rho \mathbf{f} dv,$$

the right members containing the direct acting forces resultant (i.e., the sum of the stresses resultant and of the external body forces), respectively the moment resultant of these direct forces (moment evaluated vs. the same point O).

But, by using the continuity equation, we remark that $\frac{D}{Dt} \int_D \rho \mathbf{v} dv = \int_D \rho \mathbf{a} dv$. In fact, on components, we have

$$\frac{D}{Dt} \int_D \rho v_i dv = \int_D (\dot{\rho} v_i + \rho a_i + \rho v_i \operatorname{div} \mathbf{v}) dv = \int_D \rho a_i dv.$$

Under these circumstances, the above equations become

$$\int_D \rho \mathbf{a} dv = \int_S \mathbf{T} da + \int_D \rho \mathbf{f} dv$$

and

$$\int_D \mathbf{r} \times \rho \mathbf{a} dv = \int_S \mathbf{r} \times \mathbf{T} da + \int_D \mathbf{r} \times \rho \mathbf{f} dv,$$

both equalities being valid for any subsystem $P \subset \mathcal{M}$ and implicitly for any domain $D \subset \mathcal{D}$.

A direct application of the momentum variation principle is *Cauchy's lemma* which establishes that, at any moment and at any point \mathbf{r} from a surface element of orientation given by \mathbf{n} , the stress vector \mathbf{T} , supposed continuous in \mathbf{r} , satisfies the action and reaction principle, i.e., [33] $\mathbf{T}(\mathbf{r}, \mathbf{n}, t) = -\mathbf{T}(\mathbf{r}, -\mathbf{n}, t)$.

2.4 The Cauchy Stress Tensor

As the stress vector \mathbf{T} , evaluated at a point \mathbf{r} , does not depend only on \mathbf{r} and t but also on the orientation of the surface element where the point is considered (i.e., on \mathbf{n}), this vector *cannot* define the stress state at the respective point. In fact, at the same point \mathbf{r} , but considered on differently oriented surface elements, the vectors \mathbf{T} could also be different. This inconvenience could be overcome by the introduction, instead of an unique vector \mathbf{T} , of a *triplet* of stress vectors \mathbf{T}_j whose components with respect to the coordinates axes will form a so-called *tensor of order 2*. This *stress tensor*, introduced by Cauchy, is the first tensor quantity reported by science history.

The triplet of stress vectors thus introduced will be associated, at every moment, to the same point \mathbf{r} but considered on three distinct surface elements having, respectively, the outward normal parallel with the unit vectors \mathbf{i}_j of the reference system, namely $\mathbf{T}_j = \mathbf{T}(\mathbf{r}, \mathbf{i}_j, t)$ ($j = 1, 2, 3$). Let us denote by τ_{ij} ($i = 1, 2, 3$) the components on the axes Ox_j of the vector \mathbf{T}_j , i.e., $\mathbf{T}_j = \tau_{ij} \mathbf{i}_i$.

We will show, in what follows, that the stress state at a point \mathbf{r} , at every moment t , will be characterized by the triplet of these vectors \mathbf{T}_j or, synonymously, by the set of the nine scalars τ_{ij} ($i = 1, 2, 3; j = 1, 2, 3$) which depend only on \mathbf{r} . Precisely, we will show that the

stress \mathbf{T} , evaluated for the considered moment at a point \mathbf{r} , situated on a surface element of normal $\mathbf{n}(n_j)$, can be expressed by the relation $\mathbf{T}(\mathbf{r}, \mathbf{n}, t) = \mathbf{T}_j(\mathbf{r}, t)\mathbf{n}_j$, known as *Cauchy's theorem*.

The proof is backed by the theorem (principle) of momentum applied to a tetrahedral continuum element with its vertex at \mathbf{r} , the lateral faces being parallel to the planes of coordinates, its base is parallel to the plane which is tangent to the surface element where the point \mathbf{r} is located. Considering then that the volume of the tetrahedron tends to zero and using the mean theorem for each of the coordinates, we get Cauchy's theorem. The detailed proof can be found, for instance, in [33].

Let us now consider, for any moment t , the linear mapping $[\mathbf{T}]$ of the Euclidean space E_3 into itself, a mapping defined by the collection of the nine numbers $\tau_{ij}(\mathbf{r}, t)$, i.e., $[\mathbf{T}]\mathbf{i}_j = \tau_{ij}\mathbf{i}_i$. Such a mapping which, in general, is called a *tensor* will be, in our case, just the *Cauchy stress tensor*, a second order tensor in E_3 . We will see that by knowing the tensor $[\mathbf{T}]$ which depends, for any instant t , only on \mathbf{r} , we have the complete determination of the stress state at the point \mathbf{r} .

Precisely we have

$$\mathbf{T}(\mathbf{r}, \mathbf{n}, t) = \mathbf{T}_j(\mathbf{r}, t)\mathbf{n}_j = \tau_{ij}\mathbf{i}_i\mathbf{n}_j = [\mathbf{T}]\mathbf{i}_j\mathbf{n}_j = [\mathbf{T}](\mathbf{r}, t)\mathbf{n}.$$

This fundamental relation shows that \mathbf{T} depends linearly on \mathbf{n} and, consequently, it will always be continuous with respect to \mathbf{n} .

It is also shown that the tensor $[\mathbf{T}]$ is an objective tensor, i.e., at a change of a spatio-temporal frame, change defined by the mapping $[\mathbf{Q}]$ or by the orthogonal proper matrix $Q_{ij} = \mathbf{i}'_i\mathbf{i}_j$, the following relation holds:

$$[\mathbf{T}]'(\mathbf{r}', t') = [\mathbf{Q}](t)[\mathbf{T}](\mathbf{r}, t)[\mathbf{Q}]^T, t' = t + \tau.$$

(the proof could be found, for instance, in [33]).

It is also proved that $[\mathbf{T}]$ is a symmetric tensor, i.e., $[\mathbf{T}] = [\mathbf{T}]^T$ [33]. This result, besides the fact that it decreases the number of parameters which define the stress state (from 9 to 6), will also imply the existence, at every point, of three orthogonal directions, called *principal directions*, and vs. them the normal stresses ($\mathbf{T} \cdot \mathbf{n}$) take extreme values which are also the eigenvalues of the tensor (mapping) $[\mathbf{T}]$.

The stress tensor symmetry is also known as "*the second Cauchy's theorem (law)*".

2.5 The Cauchy Motion Equations

Cauchy's theorem allows us to rewrite in a different form the principle of the momentum tensor variation, that means of the linear momentum and of the angular momentum variation.

Precisely, it is known that

$$\frac{D}{Dt} \int_D \rho \mathbf{v} dv = \int_S [\mathbf{T}] \mathbf{n} da + \int_D \rho \mathbf{f} dv$$

and

$$\frac{D}{Dt} \int_D \mathbf{r} \times \rho \mathbf{v} dv = \int_S \mathbf{r} \times [\mathbf{T}] \mathbf{n} da + \int_D \mathbf{r} \times \rho \mathbf{f} dv.$$

Obviously, in the conditions of the continuous motions (which correspond to the parameters field of class $C^1(\mathcal{D})$), by using the extension of Green's formulas for tensors of order greater than 1 [Appendix A] together with the fundamental lemma, from the (linear) momentum theorem one gets

$$\rho a_i = \tau_{ij,j} + \rho f_i, (i = 1, 2, 3),$$

relations known as *Cauchy's equations* or "*the first Cauchy's law (theorem)*".

These equations could be established under different forms too. Thus, starting with the formulas for the total derivative of both the momentum $\frac{D}{Dt}(\rho \mathbf{v}) = \frac{\partial}{\partial t}(\rho \mathbf{v}) + (\mathbf{v} \cdot \nabla) \rho \mathbf{v}$ and the volume (depending on time) integral, we have

$$\int_D \left[\frac{\partial}{\partial t}(\rho \mathbf{v}) + (\mathbf{v} \cdot \nabla) \rho \mathbf{v} + \rho \mathbf{v} \operatorname{div} \mathbf{v} \right] dv = \int_D (\operatorname{div}[\mathbf{T}] + \rho \mathbf{f}) dv.$$

As $(\mathbf{v} \cdot \nabla) \rho \mathbf{v} + \rho \mathbf{v} \operatorname{div} \mathbf{v} = \operatorname{div}(\rho \mathbf{v} \otimes \mathbf{v})$, the symbol \otimes designating the dyadic product [Appendix A], the above equation could be rewritten in the form

$$\int_D \frac{\partial}{\partial t}(\rho \mathbf{v}) dv + \int_S (\rho \mathbf{v} \otimes \mathbf{v} - [\mathbf{T}]) \mathbf{n} da = \int_D \rho \mathbf{f} dv,$$

known also as the *transport equation of (linear) momentum* and which could be used, in fluid dynamics, for evaluation of the global actions exerted on the immersed bodies.

Then, by using the fundamental lemma, one gets the so-called conservative form of Cauchy's equations

$$\frac{\partial}{\partial t} (\rho \mathbf{v}) + \operatorname{div} (\rho \mathbf{v} \otimes \mathbf{v} - [\mathbf{T}]) = \rho \mathbf{f},$$

which, on components, leads to

$$\frac{\partial}{\partial t} (v_i) + (\rho v_i v_j - \tau_{ij})_{,j} = \rho f_i \quad (i = 1, 2, 3).$$

Concerning the writing of Cauchy's equations in Lagrangian coordinates this requires the introduction of some new tensors as, for instance, the Piola–Kirchoff tensor [33].

Concerning the objectivity (frame invariance) of the Cauchy equations we remark that these equations are *not* frame invariants. Really while the forces which correspond to the contact or distance direct actions are essentially objective (frame invariants) as well as $\mathbf{n} = \frac{\operatorname{grad} f}{|\operatorname{grad} f|}$ and $\operatorname{div}[\mathbf{T}]$ (these together with $\operatorname{grad} f$ and $[\mathbf{T}]$ respectively), the acceleration vector which obviously depends on the frame of reference, is *not* objective.

An objective form of these equations obtained by the introduction of some new vectors but without a physical meaning can be found in [33].

With respect to the mathematical “closure” of the Cauchy system of equations (3 equations with 10 unknowns), this should be established by bringing into consideration the *specific behaviour*, the connection between stresses and deformations, i.e., the “constitutive law” for the continuum together with a thermodynamic approach to the motion of this medium.

2.6 Principle of Energy Variation. Conservation of Energy

The fact that the energy of a material system does not change while the system moves, i.e., the so-called “energy conservation”, will lead to another equation which characterizes the motion of the material medium.

Obviously, by introduction of some thermodynamic considerations later on, this energy equation will be rewritten in a more precise form.

Let us assess the elemental work done per unit time (*the power*) of the forces exerted on a material subsystem P of the deformable continuum \mathcal{M} and whose configuration is D , i.e.,

$$\frac{\delta L}{dt} = \int_S \mathbf{v} \cdot [\mathbf{T}] \mathbf{n} da + \int_D \rho \mathbf{f} \cdot \mathbf{v} dv.$$

Using then the equality $\mathbf{v} \cdot [\mathbf{T}]\mathbf{n} = [\mathbf{T}]\mathbf{v} \cdot \mathbf{n}$, a consequence of the definition of the transposed tensor and of the symmetry of the stress tensor, precisely

$$\mathbf{v} \cdot [\mathbf{T}]\mathbf{n} = [\mathbf{T}]^T \mathbf{v} \cdot \mathbf{n} = [\mathbf{T}]\mathbf{v} \cdot \mathbf{n} \quad [\text{Appendix A}],$$

the first integral of the right side, $\int_S \mathbf{v} \cdot [\mathbf{T}]\mathbf{n} da$, becomes

$$\int_S \mathbf{v} \cdot [\mathbf{T}]\mathbf{n} da = \int_D div[\mathbf{T}]\mathbf{v} dv = \int_D (\tau_{ij} v_i)_{,j} dv.$$

Since $\int_D \rho \mathbf{f} \cdot \mathbf{v} dv = \int_D (\rho a_i v_i - \tau_{ij,j} v_i) dv$, from the Cauchy equations, taking into account that the second order tensor $[\mathbf{G}] = grad \mathbf{v}$ (of components $v_{i,j}$) can be split as a sum of a symmetric tensor $[\mathbf{D}]$ of components $D_{ij} = \frac{1}{2}(v_{i,j} + v_{j,i})$ (the rate-of-strain tensor) and a skew-symmetric tensor $[\mathbf{\Omega}]$ of components $\Omega_{ij} = \frac{1}{2}(v_{i,j} - v_{j,i})$ (the rotation tensor) while $\int_D [\mathbf{G}] \cdot [\mathbf{T}] dv = \int_D [\mathbf{D}] \cdot [\mathbf{T}] dv$, we finally have

$$\frac{\delta L}{dt} = \int_D [\mathbf{D}] \cdot [\mathbf{T}] dv + \frac{D}{Dt} \frac{1}{2} \int_D \rho v^2 dv = W + \dot{E}_C,$$

where W is the internal(deformation) energy whose existence is correlated with the quality of our continuum to be deformable (for rigid bodies obviously $W = 0$) while E_C is the kinetic energy of the system.

Usually a *specific deformation energy* w is defined by $2w = [\mathbf{T}] \cdot [\mathbf{D}] = tr([\mathbf{T}][\mathbf{D}])$ and then $W = 2 \int_D w dv$.

Part of the work done, contained in W , may be recoverable but the remainder is the *lost work*, which is destroyed or dissipated as heat due to the internal friction.

So we have, in the language of deformable continua, the result of energy conservation which states that the work done by the forces exerted on the material subsystem P is equal to the rate of change of kinetic energy E_C and of internal energy W .

2.7 General Conservation Principle

The integral form of mass conservation, momentum torsor and energy principle as established in the previous section respectively, can all be joined together into a unique general conservation principle. Precisely, for any material subsystem $P \subset \mathcal{M}$, which occupies the configuration $D \subset \mathcal{D}$ whose boundary is S , at any moments t_1 and t_2 , we have the following common form for these principles:

$$\int_{D(t_2)} Adv - \int_{D(t_1)} Adv + \int_{t_1}^{t_2} \int_{S(t)} B \cdot \mathbf{n} da dt = \int_{t_1}^{t_2} \int_{D(t)} C dv dt.$$

Obviously if all considered variables (i.e., the motion) are assumed continuous in time, the general conservation principle becomes

$$\frac{D}{Dt} \int_{D(t)} Adv + \int_{S(t)} B \cdot \mathbf{n} da = \int_{D(t)} C dv,$$

where \mathbf{n} is the unit outward vector drawn normal to the surface S .

The above relation states that for a volume support D , the rate of change of what is contained in D , at moment t , plus the rate of flux out of S , is equal to what is furnished to D . The quantities A, B, C are tensorial quantities, A and C having the same tensorial order. If $B \neq 0$, then it is a tensor whose order is one unity higher than A .

If we use the Reynolds transport theorem for the first integral and the Gauss divergence theorem for the second integral, we have

$$\int_{D(t)} \left[\frac{\partial A}{\partial t} + \text{div} \mathbf{f} - C \right] dv = 0,$$

where $\mathbf{f} = A\mathbf{v} + B$.

Since the above result is valid for any material subsystem P of the deformable continuum (i.e., for any D) the fundamental lemma and the same hypothesis on the motion continuity allows us to write

$$\frac{\partial A}{\partial t} + \text{div} \mathbf{f} = C,$$

which is the unique general differential equation, in conservative form, associated to the studied principles.

3. Constitutive Laws. Inviscid and real fluids

3.1 Introductory Notions of Thermodynamics. First and Second Law of Thermodynamics

Thermodynamics is concerned with the behaviour of different material systems from the point of view of certain *state* or *thermodynamic variables parameters*. The considered thermodynamic (state) variables will be the absolute temperature (the fundamental quantity for thermodynamics), the pressure p , the mass density ρ , the specific (per mass unity) internal energy e and the specific entropy s . The last two state variables will be defined in what follows.

The main aim of thermodynamics is to establish a certain functional dependence among the state (thermodynamic) variables known as *constitutive (behaviour) laws (equations)*. These constitutive equations will contribute to the mathematical “closure” of the equations system describing the deformable continuum motion.

Obviously the deformation of the material systems depends essentially on the temperature when this deformation takes place. That is why, for a complete study, a deformable continuum should be considered as a *thermodynamic system*, i.e., a *closed* material system (no matter enters or leaves the system) which changes energy with its surrounding through work done or heat (added or taken).

By the *thermodynamic state* of a system, at a certain instant, we understand the set of all the values of the state (thermodynamic) variables (parameters) which characterize the system at that moment.

By a *thermodynamic process* we understand a change of the thermodynamic state (i.e., of the values of the state variables) as a consequence of certain operations or actions or, shorter, when a thermodynamic system changes from one state to another one.

A system is called in *thermodynamic equilibrium* if its thermodynamic state is time invariant.

Suppose now that a thermodynamic system has changed from an initial state (1) to a new state (2). By producing changes in either the system or its surrounding, it would be possible to reverse the state from (2) to (1). If this is possible to be done without any modification in both system and surrounding, the process is called *reversible*. On the contrary it is *irreversible*.

The reversible processes characterize the ideal media and they imply infinitesimal changes which have been carried out so slowly that both the system and the surrounding pass successively through a sequence of equilibrium states.

The *internal energy* E_i , associated to a material system, is the complementary value of the kinetic energy E_C , vs. the total energy E , i.e., $E = E_i + E_C$.

Depending only on the state of the system at the considered moment (and not on the way this state has been reached), the internal energy is an objective quantity (while the kinetic energy, due to the presence of \mathbf{v} , is not objective). If we postulate that the internal energy is an absolutely continuous function of mass, there will be a function e , called the *specific internal energy*, such that

$$E_i(P) = \int_P e dm = \int_D \rho e dv.$$

In fact the first law of thermodynamics postulates the possibility to transform the heat (thermal energy) into mechanical energy. More precisely within a thermodynamical process (when the deformable material subsystem passes from a thermodynamical state to a “neighboring” one), the rate of change of the total energy $\frac{dE}{dt}$ is equal to the elemental power $\frac{\delta L}{dt}$ of the direct forces exerted on the system plus the quantity of heat added to or taken out per unit time $\frac{\delta Q}{dt}$, so we have

$$\frac{DE}{Dt} = \frac{\delta L}{dt} + \frac{\delta Q}{dt}.$$

If $\delta Q \equiv 0$, i.e., there is not a heat change with the surrounding, the process (and the motion) are called *adiabatic*. Generally $\delta Q = \delta Q_c + \delta Q_d$, where δQ_c and δQ_d are, respectively, “contact actions” (the conduction heat) and “distance actions” (the radiation heat). By accepting (to introduce the corresponding densities) that $\frac{\delta Q_c}{dt}$ and $\frac{\delta Q_d}{dt}$ are absolutely continuous functions of surface and, respectively, mass, we will have that

$$\frac{\delta Q_c}{dt} = \int_S q(\mathbf{v}, \mathbf{n}, t) da, \quad \frac{\delta Q_d}{dt} = \int_D \rho r_d(\mathbf{r}, t) dv,$$

D being, at the respective moment, the configuration of the subsystem P and S its boundary.

Under these circumstances, for any deformable continuum subsystem P , the first law of thermodynamics can be written

$$\frac{D}{Dt} \int_D \rho \left(e + \frac{1}{2} v^2 \right) dv = \int_S (\mathbf{v} \cdot \mathbf{T} + q) da + \int_D \rho (\mathbf{f} \cdot \mathbf{v} + r_d) dv.$$

On the other side the energy variation principle, stated in the previous section, is

$$\frac{D}{Dt} \int_D \frac{1}{2} \rho v^2 dv + \int_D 2w dv = \int_S \mathbf{v} \cdot \mathbf{T} da + \int_D \rho \mathbf{f} \cdot \mathbf{v} dv$$

such that, using also the transport formula and the continuity equation, the first law of thermodynamics could be written

$$\int_D \rho \dot{e} dv = \int_S q(\mathbf{r}, \mathbf{n}, t) da + \int_D (2w + \rho r_d) dv.$$

By introducing now the *heat flux principle* (Fourier–Stokes) which states that there is a vector $\mathbf{q}(\mathbf{r}, t)$, called *heat density vector*, so that

$$\mathbf{q}(\mathbf{r}, \mathbf{n}, t) = -\mathbf{n} \cdot \mathbf{q}(\mathbf{r}, t)^9,$$

the Gauss divergence theorem leads to

$$\int_D (\rho \dot{e} + \operatorname{div} \mathbf{q} - 2w - \rho r_d) = 0,$$

that is, using the fundamental lemma too,

$$\rho \dot{e} = 2w - \operatorname{div} \mathbf{q} + \rho r_d.$$

Obviously if we did not “split” δQ into the conduction heat and the radiation heat, the last two terms of the above relation would be represented by the unique term $\rho \frac{\delta q}{dt}$, δq being the total heat density per unit of mass.

To conclude, the energy equation together with the first law of thermodynamics could be written both in a *nonconservative form*

$$\rho \frac{D}{Dt} \left(e + \frac{v^2}{2} \right) = \rho \frac{\delta q}{dt} + \operatorname{div}[\mathbf{T}]\mathbf{v} + \rho \mathbf{f} \cdot \mathbf{v},$$

and in a *conservative form* or of *divergence type*¹⁰

$$\frac{\partial}{\partial t} \left[\rho \left(e + \frac{v^2}{2} \right) \right] + \nabla \cdot \left[\rho \left(e + \frac{v^2}{2} \right) \mathbf{v} \right] = \rho \frac{\delta q}{dt} + \operatorname{div}[\mathbf{T}]\mathbf{v} + \rho \mathbf{f} \cdot \mathbf{v},$$

this last form playing a separate role in CFD.

The second law of thermodynamics, known also as the Kelvin–Planck or Clausius principle, is a criterion which points out in what sense a thermodynamic process is irreversible.

It is well known that all the real processes are irreversible, the reversibility being an attribute of only ideal media. While the first law of thermodynamics does not say anything on the reversibility of the postulated transformations, the second law tries to fill up this gap. More

⁹For sake of simplicity we consider only the case of the heat added to P and corresponding “ $-\mathbf{n}$ ” will represent the unit inward normal drawn to S and this is the right unit normal vector we deal with in our case.

The heat flux principle could be got by applying the above form of the first law of thermodynamics to a tetrahedron of Cauchy type (that is a similar tetrahedron with that used in the proof of the Cauchy theorem)

¹⁰The transformation of the left side could be done by using the derivative of a product and the equation of continuity.

precisely, in a simplified form, one postulates that a transformation, a thermodynamical process, takes place in such a way that the *entropy does not decrease or remain the same*.

What is the entropy ? In the case of reversible processes, the specific entropy (per mass unit) s is defined by the differential relation $ds = \frac{\delta q}{T}$, where δq is the total heat per mass unit while T is the absolute temperature — an objective and intensive quantity (i.e., it is not an absolute continuous function of volume) — whose values are strictly positive and which is the fundamental quantity of thermodynamics. But, generally, the entropy S for the material subsystem P will also be a state quantity which is an absolute continuous function of mass (extensive quantity) and it can be expressed, via Radon–Nycodim’s theorem as $S = \int_P s(\mathbf{r}, t) dm$, s being the specific entropy. In the case of an irreversible process this entropy changes as a result of both interaction with surroundings (external action) and inside transformations (internal actions) such that we have $ds = ds_e + ds_i$.

Since $ds_i \geq 0$ (a result coming from kinetics) and $ds_e = \frac{\delta q}{T}$, we have that $ds \geq \frac{\delta q}{T}$ which is the *local form* of the second law, also known as the *Clausius–Duhem inequality*. We remark that the “equality symbols” would belong to the case $ds = ds_e$ and, implicitly, to the reversible (ideal) processes. Obviously for these reversible processes, using also the first law of thermodynamics under the form $\rho \dot{e} = [\mathbf{T}] \cdot [\mathbf{D}] + \rho \frac{\delta q}{dt}$, one obtains the so-called Gibbs equation

$$\rho \dot{e} = [\mathbf{T}] \cdot [\mathbf{D}] + \rho T \dot{s},$$

which is fundamental in the study of ideal continua.

Concerning the general (unlocal) formulation for the second law of thermodynamics, the condition of some real (irreversible) processes, this could be the following:

For any material subsystem P of the deformable continuum M , which is seen in the configuration D of boundary ∂D , there is a state quantity S , called entropy, whose rate of change, when the subsystem is passing from a state to another (neighboring) one, satisfies

$$\dot{S} \geq \frac{\delta Q}{T} = - \int_{\partial D} \frac{\mathbf{q} \cdot \mathbf{n}}{T} da + \int_D \rho \frac{r_d}{T} dv \geq 0.$$

3.1.1 Specific Heats. Enthalpy

The *specific heat* is defined as the amount of heat required to increase by unity the temperature of a mass unit of the considered medium. Correspondingly, the specific heat is

$$C = \frac{\delta q}{dT}.$$

Supposing that the temperature is a function of p and $\frac{1}{\rho} = v$, we have

$$dT = \left(\frac{\partial T}{\partial p} \right)_v dp + \left(\frac{\partial T}{\partial v} \right)_p dv,$$

where the subscript denotes the fixed variable for partial differentiation. Analogously, assuming that the specific internal energy e is also a function of p and v we have

$$de = \left(\frac{\partial e}{\partial p} \right)_v dp + \left(\frac{\partial e}{\partial v} \right)_p dv.$$

Referring to the case of fluids, as the work done by a unit mass “against” the pressure forces is $\delta w = pd\left(\frac{1}{\rho}\right) = pdv$, the first law of thermodynamics can be written

$$de = \delta q - pdv,$$

where δq is the heat added to the unit mass. Because $\frac{1}{T}$ is an integrating factor for δq , in the sense that $ds_e = \frac{\delta q}{T}$, we get $Tds_e = de + pdv$. Obviously, for reversible processes (ideal media) $ds_e = ds$ and the last relation becomes $Tds = de + pdv$, an equation which could be also deduced as a consequence of Gibbs’ equations (for inviscid fluids).

Generally, for any fluids, by using the above expression for de and the first law of thermodynamics, we have that

$$\delta q = \left(\frac{\partial e}{\partial p} \right)_v dp + \left(\frac{\partial e}{\partial v} \right)_p dv + pdv.$$

Hence the specific heat is

$$C = \frac{\delta q}{dT} = \frac{\left(\frac{\partial e}{\partial p} \right)_v dp + \left(\frac{\partial e}{\partial v} \right)_p dv + pdv}{\left(\frac{\partial T}{\partial p} \right)_v dp + \left(\frac{\partial T}{\partial v} \right)_p dv}.$$

From this expression it will be possible to define two “principal” specific heats: one C_p , for $dp = 0$ ($p = \text{constant}$), called the *specific heat at*

constant pressure, and the other C_v , for $dv = 0$ ($v = \text{constant}$), called the *specific heat at constant volume*. Thus

$$C_p = \left(\frac{\delta q}{dT} \right)_{dp=0} = \frac{1}{\left(\frac{\partial T}{\partial v} \right)_p} \left[\left(\frac{\partial e}{\partial v} \right)_p + p \right] = \left(\frac{\partial e}{\partial T} \right)_p + p \left(\frac{\partial v}{\partial T} \right)_p$$

and¹¹

$$C_v = \left(\frac{\delta q}{dT} \right)_{dq=0} = \frac{\left(\frac{\partial e}{\partial p} \right)_v}{\left(\frac{\partial T}{\partial p} \right)_v} = \left(\frac{\partial e}{\partial T} \right)_v.$$

Obviously, for the reversible processes (ideal media) we also have $C_p = \left(\frac{\delta q}{\delta T} \right)_p = T \left(\frac{\partial s}{\partial T} \right)_p$ and $C_v = \left(\frac{\delta q}{\delta T} \right)_v = T \left(\frac{\partial s}{\partial T} \right)_v$.

Concerning the difference $C_p - C_v$, this is equal to $T \left(\frac{\partial p}{\partial T} \right)_v \left(\frac{\partial v}{\partial T} \right)_p$ a result which can be found, for instance, in [33].

Now, let us introduce a new state variable H called *enthalpy* or *total heat*. The enthalpy h per unit mass or the specific enthalpy is defined by $h = e + pv$.

Differentiating this relation with respect to T , while keeping p constant, we obtain

$$\left(\frac{\partial h}{\partial T} \right)_p = \left(\frac{\partial e}{\partial T} \right)_p + p \left(\frac{\partial v}{\partial T} \right)_p = C_p.$$

In terms of h , the above Gibbs' equation could also be written as

$$Tds = dh - vdp,$$

a form which will be important in the sequel.

3.2 Constitutive (Behaviour, "Stresses-Deformations" Relations) Laws

The system of equations for a deformable continuum medium — the translation of the Newtonian mechanics principles into the appropriate language of these media — should be *closed* by some equations of specific structure characterizing the considered continuum and which influence its motion. Such equations of specific structure, consequences

¹¹ We have used here some results of the type $\frac{1}{\left(\frac{\partial T}{\partial v} \right)_p} = \left(\frac{\partial v}{\partial T} \right)_p$ etc. which come from the classical calculus.

of the motion equations of particles within the microscopic theory and which, in our phenomenological approach are given by experience as physical laws, will be designated as *constitutive* or *behaviour laws* or simply “*stresses-deformations*” *relations* (in fact they are functional dependences between the stress tensor and the mechanical and thermodynamical parameters which are associated to the motion, between the quantities which characterize the deformation and the stresses which arise as a reaction to this deformation).

Noll has given a set of *necessary* conditions, in the form of general principles, which should be fulfilled by any constitutive law. By using the necessary conditions, some general dependences between the mechanical and thermodynamical parameters will be “filtered” and thus a screening of real candidates among different “stresses-deformations” relations is performed [95].

In what follows we will present, in short, the most important of such principles (the details could be found in [95]).

The first principle is that of *dependence on “the history” of the material*. According to this principle the stress state at a certain point of the deformable continuum and at a given moment, depends on the whole “history” of the evolution (from the initial to the given considered moment) of the entire material system. In other words, this principle postulates that the stress at a point of continuum and at a certain moment is determined by a sequence of all the configurations the continuum has passed through from the initial moment till the considered moment (included).

A second principle which is in fact a refinement of the previous principle is that of *spatial localization*. According to this principle, to determine the stress state at a certain point and at a certain moment t , *not* the whole history of the *entire* continuum is required but only the *history of a certain neighborhood* of the considered point.

Finally, the most powerful (by its consequences) principle would be that of *objectivity* or *material frame indifference*. According to this principle a constitutive law should be objective and so frame invariant which agrees with the intrinsic character of such a law.

An important consequence of this objectivity principle is the impossibility of the time to appear *explicitly* in such a law.

If in a constitutive law the point where the stress is evaluated does not appear explicitly, the respective medium is called *homogeneous*. The homogeneity is also an intrinsic property of the medium. It can be shown then if there is a reference configuration where the medium is homogeneous that it will keep this quality in any other configuration [150].

A deformable continuum is called *isotropic*, if there are *not* privileged directions or, in other terms, the (“answering”) functional which defines the stress tensor is isotropic or frame rotation invariant.

According to the Cauchy–Eriksen–Rivlin theorem [40], a tensor function $f([\mathbf{A}])$, defined on a set of symmetric tensors of second order from E_3 and whose values are in the same set, is isotropic if and only if it has the form $f([\mathbf{A}]) = \varphi_0[\mathbf{I}] + \varphi_1[\mathbf{A}] + \varphi_2[\mathbf{A}]^2$, where φ_k are isotropic scalar functions of the tensor $[\mathbf{A}]$ which could always be expressed as functions of the principal invariants I_1, I_2, I_3 of the tensors $[\mathbf{A}]$, i.e., $\varphi_k = \varphi_k(I_1, I_2, I_3)$.

As a corollary any linear isotropic tensor function $l([\mathbf{A}])$ in E_3 should be under the form $l([\mathbf{A}]) = c_0 \text{tr}([\mathbf{A}])[\mathbf{I}] + c_1[\mathbf{A}]$, where c_0 and c_1 are constants.

3.3 Inviscid (Ideal) Fluids

The simplest of all the mathematical and physical models associated to a deformable continuum is the *model of the inviscid (ideal) fluid*.

By an *inviscid (ideal) fluid* we understand that deformable continuum which is characterized by the constitutive law $[\mathbf{T}] = -p[\mathbf{I}]$ (or, on components, $\tau_{ij} = -p\delta_{ij}$) where p is a positive scalar depending only on \mathbf{r} and t (and *not* on \mathbf{n}), physically coinciding with the (thermodynamical) *pressure*.

The “hydrostatic” form (characterizing the equilibrium) of the stress tensor $[\mathbf{T}] = -p[\mathbf{I}]$ shows that the stress vector \mathbf{T} is collinear with the outward normal \mathbf{n} drawn to the surface element (and, obviously, of opposite sense) i.e., for an inviscid fluid the tangential stresses (which withstand the sliding of neighboring fluid layers) are negligible.

The same structure of the constitutive law for an inviscid fluid points out that this fluid is always a homogeneous and isotropic medium.

In molecular terms, within an inviscid fluid, the only interactions between molecules are the random collisions. Air, for instance, can be treated as an inviscid fluid (gas).

With regard to the flow equations of an inviscid (ideal) fluid, known as *Euler equations*, these could be got from the motion equations of a deformable continuum (Cauchy equations), i.e., from $\rho a_i = \rho f_i + \tau_{ij,j}$ where we use now the specific structure of the stress tensor $\tau_{ij} = -p\delta_{ij}$; hence

$$\rho a_i = \rho f_i - p_{,i}$$

or, in vector language

$$\rho \mathbf{a} = \rho \mathbf{f} - \text{grad} p,$$

a system which should be completed by the equation of continuity.

Of course the Euler equations could be rewritten in a “conservative” form (by using the continuity equation and the differentiating rule of a product), namely

$$\frac{\partial(\rho v)}{\partial t} + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) = \rho \mathbf{f} - \text{grad} p.$$

If the fluid is incompressible, the Euler equations and the equation of continuity, together with the necessary boundary (slip) conditions (characterizing the ideal media) which now become sufficient conditions too, ensure the coherence of the respective mathematical model, i.e., they will allow the determination of all the unknowns of the problem (the velocity and pressure field). If the fluid is compressible one adds the unknown $\rho(\mathbf{r}, t)$, which leads to a compulsory thermodynamical study of the fluid in order to establish the so-called *equation of state* which closes the associated mathematical model.

The thermodynamical approach to the inviscid fluid means the use of the energy equation (together with the first law of thermodynamics) and of Gibbs’ equation which, being valid for any ideal continuum, synthesizes both laws of thermodynamics.

The energy equation, either under nonconservative form or under conservative form, comes directly from the corresponding forms of an arbitrary deformable continuum, namely from

$$\rho \frac{D}{Dt} \left(e + \frac{v^2}{2} \right) = \rho \frac{\delta q}{dt} - \text{div}(p\mathbf{v}) + \rho \mathbf{f} \cdot \mathbf{v}$$

respectively

$$\frac{\partial}{\partial t} \left\{ \rho \left(e + \frac{v^2}{2} \right) + \nabla \cdot \left[\rho \left(e + \frac{v^2}{2} \right) \mathbf{v} \right] \right\} = \rho \frac{\delta q}{dt} - \text{div}(p\mathbf{v}) + \rho \mathbf{f} \cdot \mathbf{v}.$$

Concerning the Gibbs’ equation, $\rho \dot{e} = [\mathbf{T}] \cdot [\mathbf{D}] + \rho T \dot{s}$, in the case of an inviscid fluid it becomes ($[\mathbf{T}] = -p[\mathbf{I}]$ so that $[\mathbf{T}] \cdot [\mathbf{D}] = -p[\mathbf{I}] \cdot [\mathbf{D}] = -p \text{tr}([\mathbf{D}]) = -p \text{div} \mathbf{v}$)

$$\rho \dot{e} = -p \text{div} \mathbf{v} + \rho T \dot{s}$$

or, by eliminating $\text{div} \mathbf{v}$ from the equation of continuity ($\text{div} \mathbf{v} = -\frac{\dot{\rho}}{\rho}$), we get

$$de = Tds - pd\left(\frac{1}{\rho}\right).$$

This last differential relation could be the departure point in the thermodynamical study of the ideal fluids. If the internal energy e is given as a function of the independent parameters s and $\frac{1}{\rho} = v$, i.e., if we know $e = \hat{e}(s, v)$, then we will immediately have the equations of state $p = -\frac{\partial \hat{e}}{\partial v}(s, v)$ and $T = \frac{\partial \hat{e}}{\partial s}(s, v)$ or, in other words, the function $\hat{e}(s, v)$, determining the thermodynamical state of the fluid, is a thermodynamical potential for this fluid. Obviously, this does not occur if e is given as a function of other parameters when we should consider other appropriate thermodynamical potentials.

If the inviscid fluid is incompressible, from $d\left(\frac{1}{\rho}\right) = dv = 0$ we have that $T = \hat{T}(s)$ or $s = \hat{s}(T)$ and hence $e = \hat{e}(T)$. More, if in the energy equation, written under the form

$$\rho \dot{e} = [\mathbf{T}] \cdot [\mathbf{D}] - \text{div} \mathbf{q} + \rho r_d,$$

we accept the use of the Fourier law $\mathbf{q} = -\chi \text{grad} T$, where χ is the thermal conduction coefficient which is supposed to be positive (which expresses that the heat flux is opposite to the temperature gradient), we get finally

$$\rho \dot{e} = \text{div}(\chi \text{grad} T) + \rho r_d.$$

As $e = \hat{e}(T)$ and r_d (the radiation heat) is given together with the external mass forces, the above equation with appropriate initial and boundary conditions, allows us to determine the temperature T *separately* from the fluid flow which could be made precise by considering only the Euler equations and the equation of continuity.

This dissociation will not be possible, in general, within the compressible case. Even the simplest statics (equilibrium) problems for the fluids testify that.

An important situation for the compressible fluids is that of the perfect fluids (gases), the air being one of them.

By a *perfect gas*, we understand an ideal gas which is characterized by the equation of state (Clapeyron) $p = \rho RT$ (where R is a characteristic constant). For such a perfect gas the relation $C_p - C_v = T \left(\frac{\partial p}{\partial v}\right)_v \left(\frac{\partial v}{\partial T}\right)_p$ becomes $C_p - C_v = R$, even if C_p and C_v are functions of temperature (Joule). Since $(\delta q)_p = C_p dT$, $(\delta q)_v = C_v dT$, the first law of thermodynamics under the form $\delta q = de + pdv$, leads to

$$C_p dT = de + pd \left(\frac{1}{\rho} \right) = dh \quad (\text{as } p = \text{constant})$$

and

$$C_v dt = de.$$

At the same time, from the transcription of Gibbs' equation $Tds = de + pdv$, and from $C_p - C_v = R$, we have

$$Tds = C_v dT - T(C_p - C_v) \frac{d\rho}{\rho}$$

or

$$\frac{Tds}{C_v} = dT - T(\gamma - 1) \frac{d\rho}{\rho},$$

where $\gamma = \frac{C_p}{C_v} > 1$.

From $C_p - C_v = R$ we also get $\gamma = \frac{\frac{C_p}{R}}{\frac{C_p}{R} - 1}$ (Eucken's formula), while the state equation, in γ , becomes $T = \frac{p}{\rho C_v (\gamma - 1)}$.

The relation $\frac{Tds}{C_v} = dT - T(\gamma - 1) \frac{d\rho}{\rho}$ together with the above expression for T , assuming that C_p and C_v are constants, lead, by a direct integration, to

$$T\rho^{1-\gamma} \exp \left(\frac{-s}{C_v} \right) = \text{const}$$

respectively

$$p\rho^{-\gamma} \exp \left(\frac{s}{C_v} \right) = \text{const.}$$

If there is an *adiabatic process* (which means without any heat change with the surrounding), from $\delta q = 0$ we get $ds = 0$, i.e., the entropy s is constant along any trajectory and the respective fluid flow is called *isentropic* (if the value of the entropy constant is the same in the whole fluid, the flow will be called *homentropic*). In this case the perfect gas is characterized by the equation of state $T = K_0 \rho^{\gamma-1}$ and $p = K \rho^\gamma$, where K_0 and K are constants while we also have

$$h = C_p T, \quad e = C_v T.$$

Obviously in the case of an adiabatic process, the equation of state $p = K \rho^\gamma$, together with the Euler equations and the equation of continuity, will be sufficient for determining the unknowns (v_i, p, ρ) (the temperature

T being determined at the same time with ρ). For the same perfect gas, under the circumstances of the constancy of the specific principal heats but in the nonadiabatic case, the first law of thermodynamics (by neglecting the radiation heat) leads to

$$\rho \dot{h} = \dot{p} - \text{div } \mathbf{q}$$

or, by using the Fourier law, we arrive at

$$\rho C_p \dot{T} = \dot{p} + \text{div } (\chi \text{grad} T),$$

an equation which allows the determination of the temperature not separately, but together with Euler's equations, i.e., using the whole system of six equations with six unknowns (v_i, p, ρ, T).

Generally, the fluid characterized by the equations of state under the form $f(p, \rho) = 0$ with f satisfying the requirements of the implicit functions theorem, are called *barotropic*. For these fluids, the determination of the flow comes always to a system of five equations with five unknowns, with given initial and boundary conditions.

3.4 Real Fluids

By definition a deformable continuum is said to be a *real fluid* if it satisfies the following postulates (Stokes):

1) The stress tensor $[\mathbf{T}]$ is a continuous function of the rate-of-strain tensor $[\mathbf{D}]$, while it is independent of all other kinematic parameters (but it may depend on thermodynamical parameters such as ρ and T);

2) The function $[\mathbf{T}]$ of $[\mathbf{D}]$ does not depend on either a space position (point) or a privileged direction (i.e., the medium is homogeneous and isotropic);

3) $[\mathbf{T}]$ is a Galilean invariant;

4) At rest ($[\mathbf{D}] = 0$), $[\mathbf{T}] = -p[\mathbf{I}], p > 0$.

The scalar $p > 0$ designates the *pressure* of the fluid or the *static pressure*. A fundamental postulate states that p is identical with the thermodynamic pressure. We will see later in what circumstances this pressure is an average of three normal stresses.

Generally the structure of the stress tensor should be $[\mathbf{T}] = -p[\mathbf{I}] + [\boldsymbol{\sigma}]$, where the part "at rest" $-p[\mathbf{I}]$ is isotropic while the remaining $[\boldsymbol{\sigma}]$ is an anisotropic part. For the so-called Stokes ("without memory") fluids, $[\boldsymbol{\sigma}] = \Phi(\mathbf{v}, \text{grad } \mathbf{v}, [\mathbf{D}])$, with restriction $[\boldsymbol{\sigma}] \equiv 0$ for the fluid flows of "rigid type" (without deformations), while for the fluids "with memory", $[\boldsymbol{\sigma}]$ depends upon the time derivatives of $[\mathbf{D}]$ too.

The postulate 2) implies, through the medium isotropy, that the function $[\mathbf{T}]$ is also an isotropic function in the sense of the constitutive laws

principles. At the same time, within the frame of Noll's axiomatic system, the postulate 3), which states the inertial frame invariance of $[\mathbf{T}]$, is a consequence of the objectivity principle.

At last, the necessary and sufficient condition for the isotropy of a tensorial dependence (the Cauchy–Eriksen–Rivlin theorem) shows that, in our working space E_3 , the structure of this dependence should be of the type

$$[\mathbf{T}] = [-p(\rho, T) + \varphi_0(\rho, T, I_1, I_2, I_3)][\mathbf{I}] \\ + \varphi_1(\rho, T, I_1, I_2, I_3)[\mathbf{D}] + \varphi_2(\rho, T, I_1, I_2, I_3)[\mathbf{D}]^2,$$

where $\varphi_0, \varphi_1, \varphi_2$ are isotropic scalar functions depending upon the principal invariants I_1, I_2, I_3 of $[\mathbf{D}]$, where $I_1 = \text{tr}[\mathbf{D}] = \text{div } \mathbf{v}$, $2I_2 = (\text{tr}[\mathbf{D}])^2 - \text{tr}[\mathbf{D}]^2$ and $I_3 = \det[\mathbf{D}]$, and with the obvious restriction $\varphi_0(\rho, T, 0, 0, 0) = 0$ (conditions required by the postulate 4)).

This general form for the constitutive law defines the so-called *Reiner–Rivlin fluids*, after the names of the scientists who established it for the first time.

Those real fluids characterized by a *linear* dependence between $[\mathbf{T}]$ and $[\mathbf{D}]$ are called *Newtonian* or *viscous*. By using the corollary which gives the general form of a linear isotropic tensorial function $l([\mathbf{A}])$, observing the hydrostatic form at rest, we necessarily have for these fluids the constitutive law

$$[\mathbf{T}] = [-p(\rho, T) + \lambda(\rho, T) \text{tr}[\mathbf{D}]] [\mathbf{I}] + 2\mu(\rho, T) [\mathbf{D}],$$

where the scalars μ and λ are called, respectively, the first and the second viscosity coefficient. By accepting the Stokes hypothesis $3\lambda + 2\mu = 0$, which reduces to *one* the number of the independent viscosity coefficients and which is rigorously fulfilled by the monoatomic gases (helium, argon, neon, etc.) and approximately fulfilled by other gases (provided that $\text{div } \mathbf{v}$ is not very large) we would have (from $\text{tr}[\mathbf{T}] = -3p + (3\lambda + 2\mu) \text{tr}[\mathbf{D}]$), that $\tau_{11} + \tau_{22} + \tau_{33} = -3p$, i.e., the above mentioned result on the equality of pressure with the negative mean of normal stresses.

Obviously, for a viscous fluid there are also tangential stresses and so there is a resistance to the fluid layers sliding. The viscosity of fluids is basically a molecular phenomenon.

For the incompressible viscous fluid from $\text{tr}[\mathbf{D}] = \text{div } \mathbf{v} = 0$ we get $[\mathbf{T}] = -p[\mathbf{I}] + 2\mu[\mathbf{D}]$.

Sutherland, in the hypothesis that the colliding molecules of a perfect or quasiperfect gas are rigid interacting spheres, got for the viscosity coefficient μ the evaluation $\mu = \alpha T^{1/2} \left(1 + \frac{\beta}{T}\right)^{-1}$, where α and β are constants [153].

Fluids that do not observe a linear dependence between $[\mathbf{T}]$ and $[\mathbf{D}]$ are called *non-Newtonian*. Many of the non-Newtonian fluids are “with memory”, blood being such an example.

In the sequel we will establish the equation for viscous fluid flows without taking into account the possible transport phenomenon with mass diffusion or chemical reactions within the fluid.

Writing the stress tensor under the form $[\mathbf{T}] = -p[\mathbf{I}] + [\boldsymbol{\sigma}]$, the Cauchy equations for a deformable continuum lead to

$$\rho \frac{D\mathbf{v}}{Dt} = \rho \mathbf{f} - \text{grad} p + \text{div}[\boldsymbol{\sigma}]$$

or, in conservative form,

$$\frac{\partial}{\partial t} (\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) = \rho \mathbf{f} - \text{grad} p + \text{div}[\mathbf{G}].$$

We remark that all the left sides of these equations could be written in one of the below forms, each of them being important from a mathematical or physical point of view:

$$\begin{aligned} \rho \mathbf{a} &= \rho \left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \text{grad}) \mathbf{v} \right] = \rho \left[\frac{\partial \mathbf{v}}{\partial t} + (\text{grad} \mathbf{v}) \mathbf{v} \right] \\ &= \rho \left[\frac{\partial \mathbf{v}}{\partial t} + \text{div} (\mathbf{v} \otimes \mathbf{v}) - \mathbf{v} \text{div} \mathbf{v} \right] = \rho \left[\frac{\partial \mathbf{v}}{\partial t} + \text{grad} \left(\frac{1}{2} v^2 \right) + \boldsymbol{\omega} \times \mathbf{v} \right]. \end{aligned}$$

Concerning $\text{div} [\boldsymbol{\sigma}] = \text{div} [\lambda (\text{div} \mathbf{v}) [\mathbf{I}] + 2\mu [\mathbf{D}]]$, which is a vector, by using the formulas $2[\mathbf{D}] = (\text{grad} \mathbf{v}) + (\text{grad} \mathbf{v})^T$, $\text{div} (\text{grad} \mathbf{v})^T = \text{grad} (\text{div} \mathbf{v})$, $2[\boldsymbol{\Omega}] \mathbf{a} = \boldsymbol{\omega} \times \mathbf{a}$ ($[\boldsymbol{\Omega}]$ being the rotation tensor – the skew-symmetric part of $\text{grad} \mathbf{v}$ and \mathbf{a} an arbitrary vector), we get for $\text{div}[\boldsymbol{\sigma}]$ a first form

$$\begin{aligned} \text{div}[\boldsymbol{\sigma}] &= (\lambda + \mu) \text{grad} (\text{div} \mathbf{v}) + \mu \nabla^2 \mathbf{v} + (\text{div} \mathbf{v}) \text{grad} \lambda \\ &\quad + 2 \text{grad} \mathbf{v} (\text{grad} \mu) + (\text{grad} \mu) \times \boldsymbol{\omega}, \end{aligned}$$

where

$$\nabla^2 \mathbf{v} = \text{div} (\text{grad} \mathbf{v}), \quad \text{grad} \mathbf{v} (\text{grad} \mu) = (\text{grad} \mu \cdot \text{grad}) \mathbf{v}.$$

A *second form* is obtained by using the additional formulas

$$\operatorname{div}(\operatorname{grad} \mathbf{v}) = \operatorname{grad}(\operatorname{div} \mathbf{v}) - \operatorname{rot} \boldsymbol{\omega} = \nabla^2 \mathbf{v}, \quad 2 \operatorname{div}[\boldsymbol{\Omega}] = -\operatorname{rot} \boldsymbol{\omega};$$

more precisely, we have

$$\begin{aligned} \operatorname{div}[\boldsymbol{\sigma}] &= \operatorname{grad}[(\lambda + 2\mu)(\operatorname{div} \mathbf{v})] - \operatorname{rot} \mu \boldsymbol{\omega} + 2 \operatorname{grad} \mathbf{v}(\operatorname{grad} \mu) \\ &\quad + 2(\operatorname{grad} \mu) \times \boldsymbol{\omega} - 2(\operatorname{div} \mathbf{v}) \operatorname{grad} \mu. \end{aligned}$$

At last, by introducing some known vectorial-tensorial identities (see Appendix A), one can get a *third form*,

$$\begin{aligned} \operatorname{div}[\boldsymbol{\sigma}] &= \operatorname{grad}[(\lambda + 2\mu)(\operatorname{div} \mathbf{v})] - \operatorname{rot} \mu \boldsymbol{\omega} \\ &\quad + 2 \operatorname{grad}[(\operatorname{grad} \mu) \cdot \mathbf{v}] - 2 \operatorname{div}[(\operatorname{grad} \mu) \otimes \mathbf{v}]. \end{aligned}$$

With regard to the energy equation, by using the nonconservative, respectively the conservative form of this equation for an arbitrary deformable continuum, in the case of the viscous fluid we get

$$\rho \frac{D}{Dt} \left(e + \frac{v^2}{2} \right) = \rho \frac{\delta q}{dt} - \operatorname{div} p \mathbf{v} + \operatorname{div}([\boldsymbol{\sigma}] \mathbf{v}) + \rho \mathbf{f} \cdot \mathbf{v}$$

(the nonconservative form), respectively

$$\frac{\partial}{\partial t} \left[\rho \left(e + \frac{v^2}{2} \right) \right] + \nabla \cdot \left[\rho \left(e + \frac{v^2}{2} \right) \mathbf{v} \right] = \rho \frac{\delta q}{dt} - \operatorname{div} p \mathbf{v} + \operatorname{div}([\boldsymbol{\sigma}] \mathbf{v}) + \rho \mathbf{f} \cdot \mathbf{v}$$

(the conservative form), where, obviously,

$$[\boldsymbol{\sigma}] \mathbf{v} = \lambda(\operatorname{div} \mathbf{v}) \mathbf{v} + \mu \left[(\operatorname{grad} \mathbf{v}) + (\operatorname{grad} \mathbf{v})^T \right]^T \mathbf{v}.$$

If we are interested in the mathematical nature of these equations we remark that, firstly, the equation of continuity is a partial differential equation of first order which could be written, in Lagrangian coordinates, $J\rho(\mathbf{R}, t) = \rho_0$, such that $J\rho = \text{constant}$ is a solution of this equation which also defines the trajectories (obviously real). As these trajectories are characteristic curves too, the equation of continuity is then of hyperbolic type.

Concerning the equation of flow, if from the first form of $div[\sigma]$ we take out the second derivative terms (the “dominant” terms), they could be grouped into

$$\mu \nabla^2 \mathbf{v} + (\lambda + \mu) grad(div \mathbf{v}).$$

According to the classification of the second order partial differential equations, these equations are elliptic if the eigenvalues μ and $\lambda + 2\mu$ of the associated quadratic form are positive. Consequently, in the steady case, if $\mu > 0$ and $\lambda + 2\mu > 0$ the flow equations are of elliptic type. The same property belongs to the energy equation if, by accepting for the conduction heat the Fourier law, the thermal conduction coefficient is positive. In the unsteady case the previous equations become parabolic.

Globally speaking, the whole system of equations would be elliptic-hyperbolic in the steady case and parabolic-hyperbolic in the unsteady case. If $\mu = 0$, then the elliptic and parabolic properties will be lost.

Concerning the initial and boundary conditions, the first ones specify the flow parameters at $t = t_0$, being thus compulsory in the unsteady case. As regards the boundary conditions, they imply some information about the flow parameters on the boundary of the fluid domain and they are always compulsory for determining the solution of the involved partial differential equation in both steady and unsteady cases.

For a viscous fluid which “passes” along the surface of a rigid body, the fluid particles “wet” the body surface, i.e., they adhere. This molecular phenomenon has been proved for all the continuous flows as long as the Knudsen number (K_n) $< 0,01$.¹²

Due to this adherence the relative velocity between the fluid and the surface of the body is zero or, in other terms, if \mathbf{V}_S is the absolute velocity of the body surface and \mathbf{v} the absolute velocity of the fluid, we should have $\mathbf{v}_{surface} = \mathbf{V}_S$. If $\mathbf{V}_S = 0$, that means the body surface is at rest, then $v_t = 0$ and also $v_n = 0$, \mathbf{t} being a unit tangent vector on the surface and \mathbf{n} is the unit normal vector drawn to the surface.

These conditions are called the *adherence* or *non-slip* conditions, in opposition with the *slip* conditions $v_n = 0$ and $v_t \neq 0$ which characterize the inviscid (ideal) fluid.

Obviously the presence of a supplementary condition ($v_t = 0$) for the viscous fluids equations should not surprise because these are partial differential equations of second order while the ideal fluids equations are of first order.

¹²This number is an adimensional parameter defined by $K_n = \frac{l}{L}$, where l is the mean free path and L a reference length.

We will see that if the viscosity coefficients tend to zero, the solution of a viscous fluid problem does not converge to the solution of the *same* problem considered for an inviscid fluid. More precisely, we will establish that this convergence is non-uniform in an immediate vicinity of the body surface (where the condition $v_t = 0$ is also lost) where another approximation (than that given by the model of inviscid fluid) should be considered.

Concerning the boundary conditions they should be completed, in the case of unbounded domains, with a given behaviour at infinity (far distance) for the flow parameters.

All these features analyzed above are associated with the physical nature of the fluid flow. Within the CFD we must take care to use the most appropriate and accurate numerical implementation of the boundary conditions, a problem of great interest in CFD. We will return to this subject later in this book.

3.5 Shock Waves

In a fluid, besides the surfaces (curves) loci of weak discontinuities there could also occur some *strong* discontinuities surfaces (curves) or *shock waves* where the unknowns themselves have such discontinuities in passing from one side to the other side of the surface (curve). To determine the relations which connect the limiting values of the unknowns from each side of the shock wave (the shock relations), we should use again the general principles but under the integral form which accepts lower regularity requirements on these unknowns. Once these relations are established, we will see that if we know the state of the fluid in front of the wave (the state “0”) and the discontinuities displacement velocity d , it will be always possible to determine the state of the fluid “behind” the shock wave (the state “1”). We will deal only with the case of perfect gases where the internal specific energy is $e = \frac{p}{\rho} \left(\frac{1}{\gamma-1} \right)$ and the total specific energy is $\frac{1}{2}\rho v^2 + \rho e$, the fluid being considered in adiabatic (isothermic) evolution. This entails total energy conservation, a requirement which prevails in the equation of state in the form $p = k\rho^\gamma$.¹³

Now we introduce the concept of “weak” solution which allows the consideration of unknowns with discontinuities. Let us take, for instance,

¹³It is shown that the entropy increase, required by the second law of thermodynamics, associated with a shock raise, does not agree with an equation of state in the form $p = k\rho^\gamma$ where k is constant.

a nonlinear equation written in conservative form, i.e., in a domain D of the plane (x, t) , namely

$$u_t + (f(u))_x = 0$$

or

$$\operatorname{div} \mathbf{F} = 0$$

where $\mathbf{F} = (f(u), u)$ and “*div*” is the space-time divergence operator. If Φ is a smooth function with compact support in the plane (x, t) , then the above differential equation leads to the fulfilment, for any Φ , of the “orthogonality” relation $\int_D \Phi \operatorname{div} \mathbf{F} dx dt = 0$ which comes, by integrating by parts, to $\int_D \operatorname{grad} \Phi \cdot \mathbf{F} dx dt = 0$.

If u is a smooth function the last relation is *equivalent* with the given differential equation; but if it is not smooth enough, the last equality keeps its sense while the differential equation *does not*.

We will say that u is a *weak solution* of the differential equation if it satisfies $\int_D \operatorname{grad} \Phi \cdot \mathbf{F} dx dt = 0$, for *any* smooth function Φ with compact support. Obviously, if we want to join also the initial conditions $u(x, 0) = u_0$, then, integrating on D_t (a subdomain of D from the half-plane $t \geq 0$) we get

$$\int_{D_t} \operatorname{grad} \Phi \cdot \mathbf{F} dx dt + \int_{D \cap O_x} \Phi(x, 0) u_0(x) dx = 0$$

and if Φ has its support far from the real axis, the last term would disappear again.

So we have both a *differential* and a *weak* form for the considered equation. We will also have an *integral* form if we integrate the initial equation along an interval $[a, b]$ of the real axis, precisely $\frac{d}{dt} \int_a^b u dx = f(u)|_a^b$.

Of course we should ask if a weak solution satisfies necessarily the integral form of the equation? Provided that *the same* quantities, which showing up in the conservative form of the equation are kept for the integral form too, the answer is affirmative. That is why the weak solution will be basically the target of our searches.

Let us now investigate the properties of the weak solutions of the conservation law $u_t + (f(u))_x = 0$ in the neighborhood of a *jump* discontinuity (i.e., of first order, the only ones with physical sense). Let u be a weak solution along the smooth curve Σ in the plane (x, t) . Let Φ

be a smooth function vanishing in the closed outside of a domain S , the curve Σ dividing the domain S into the disjoint subdomains S_1 and S_2 ($S = S_1 \cup S_2$). Then

$$0 = \int_S \text{grad}\Phi \cdot \mathbf{F} dxdt = \int_{S_1} \text{grad}\Phi \cdot \mathbf{F} dxdt + \int_{S_2} \text{grad}\Phi \cdot \mathbf{F} dxdt.$$

Since u is a regular function in both S_1 and S_2 , if \mathbf{n} is the unit normal vector oriented from S_1 to S_2 , then by applying Gauss' divergence formula and the validity of the relation $\text{div}\mathbf{F} = 0$ in S_1 and in S_2 we are led to

$$\int_{\Sigma} \Phi (\mathbf{F}_1 - \mathbf{F}_2) \cdot \mathbf{n} ds = 0,$$

where \mathbf{F}_1 and \mathbf{F}_2 are the \mathbf{F} values for u taking the limiting values from S_1 , respectively S_2 .

As the above relation takes place for *any* Φ , we will have $[\mathbf{F} \cdot \mathbf{n}] = 0$ on Σ where $[\mathbf{F} \cdot \mathbf{n}] = \mathbf{F}_1 \cdot \mathbf{n} - \mathbf{F}_2 \cdot \mathbf{n}$ denotes the "jump" of $\mathbf{F} \cdot \mathbf{n}$ across Σ .

Suppose that Σ is given by the parametric equation $x = x(t)$, so that the displacement velocity of discontinuity is $d = \frac{dx}{dt}$. Further $\mathbf{n} = \frac{(1, -d)}{\sqrt{1+d^2}}$ and \mathbf{F} being $(f(u), u)$, the above relation becomes

$$-d[u] + [f(u)] = 0 \quad \text{pe } \Sigma,$$

where again $[]$ designates the jump of the quantity which is inside the parentheses, when the point (x, t) is passing across Σ (from S_1 to S_2).

A function u satisfying the differential equation whenever it is possible (in our case in S_1 and S_2) and the above jump relation across the discontinuity surface Σ , will satisfy both the integral and the weak form of the equation.

Obviously, all the above comments could be extended to the *conservative laws systems*. Let us consider, as a conservative system, the system of equations for an isentropic gas in a one-dimensional flow, precisely

$$\rho_t + m_x = 0,$$

$$m_t + \left(\frac{m^2}{\rho} + p \right)_x = 0,$$

where $m = \rho v$ (the specific momentum), system which is completed by the state equation

$$p = k\rho^\gamma.$$

But if instead of the equation of state $p = k\rho^\gamma$, we consider the energy equation

$$e_t + \left[(e + p) \frac{m}{\rho} \right]_x = 0,$$

with $e = \frac{1}{2}\rho v^2 + \frac{p}{\gamma-1}$, then some physical reasons show that the acceptance of the energy conservation is a much more realistic condition than $p = k\rho^\gamma$, k , in general, depending on entropy and so it cannot be constant.

In fact the above system together with $p = k\rho^\gamma$ *does not* have the same weak solution as the same system but is completed with the energy conservation.

There are special subjects as, for instance, the wave theory in hydrodynamics, where the results obtained by considering the equation of state $p = k\rho^\gamma$ are close to reality. But, generally speaking, the shock phenomena should be treated with the system completed with the above energy equation instead of the equation of state.

From the jump relation $[\mathbf{F} \cdot \mathbf{n}] = 0$, across the discontinuity surface Σ which moves with velocity d , we get, for any of the equations of the above system, the jump relations

$$d[\rho] = [m],$$

$$d[m] = \left[\frac{m^2}{\rho} + p \right],$$

$$d[e] = [(e + p)v],$$

called the *Rankine–Hugoniot jump relations*.

If it takes a coordinate system whose displacement with uniform velocity would be, at a moment $t = 0$, equal with the displacement velocity of a discontinuity located at the origin of this system, then within this new frame of coordinates, the previous relations will be rewritten

$$\rho_0 v_0 = \rho_1 v_1,$$

$$\rho_0 v_0^2 + p_0 = \rho_1 v_1^2 + p_1,$$

$$(e_0 + p_0) v_0 = (e_1 + p_1) v_1,$$

where the subscripts identify the state “0” before the jump and the state “1” after the jump. If $m = \rho_0 v_0 = \rho_1 v_1 = 0$, the respective discontinuity is of *contact* type because $v_0 = v_1 = 0$ show that these discontinuities move with the fluid.

If $m \neq 0$ the discontinuity will be called a *shock wave* or, shorter, a *shock*. As $v_0 \neq 0$, $v_1 \neq 0$, the fluid is passing through shock or, equivalently, the shock is moving through fluid.

That part of the gas (fluid) which does not cross the shock is called the shock *front* (the state “0”) while the part after the shock is called the *back* of the shock (the state “1”).

From the Rankine–Hugoniot relations we could get simple algebraic relations which allow the determination of the parameters after shock (state “1”) by using their values before shock (state “0”).

If $c_0 = \gamma \frac{p_0}{\rho_0}$ and $c_1 = \gamma \frac{p_1}{\rho_1}$ are the sound speed in front and, respectively, behind the shock, then denoting by $M_0 = \frac{d-v_0^n}{c_0}$ and $M_1 = \frac{d-v_1^n}{c_1}$ (v_0^n and v_1^n being the projections of the fluid velocity on the shock normal, at the origin of the system) and by $\tau_0 = \frac{1}{\rho_0}$ and $\tau_1 = \frac{1}{\rho_1}$, we easily get the relations

$$\frac{\tau_1 - \tau_0}{\tau_0} = \frac{2}{\gamma + 1} \left(\frac{1}{M_0^2} - 1 \right),$$

$$\frac{p_1 - p_0}{p_0} = \frac{2\gamma}{\gamma + 1} (M_0^2 - 1),$$

which determine τ_1 and p_1 with the data before the shock.

Analogously, we have

$$1 - M_1^2 = \frac{M_0^2 - 1}{1 + \frac{2\gamma}{\gamma+1} (M_0^2 - 1)}$$

and from the perfect gases law $p = \rho RT$ we obtain for the “new” temperature T_1 the evaluation

$$T_1 = T_0 \frac{\tau_1 p_1}{\tau_0 p_0} = 1 - \frac{2(\gamma - 1)}{(\gamma + 1)} \frac{(\gamma M_0^2 + 1) (M_0^2 - 1)}{M_0^2},$$

relation which, together with the above ones, solves completely the proposed problem.

In what follows we will see what type of conditions should be imposed to ensure the uniqueness of the (weak) physically correct solution.

It is easy to check that through every point of a shock in the (x, t) plane one can draw two characteristics, one of each side of the shock,

i.e., the shock “separates” the characteristics. These characteristics are oriented (both of them) towards the “past”, i.e., to the initial data line $t = t_0$ or towards the “future” i.e., towards larger t .

A shock is said to obey the *entropy condition* if the two characteristics which cross at each point of it are oriented backwards to the initial line $t = t_0$. A shock which does not observe the entropy condition is called a *rarefaction shock*. In gas dynamics the rarefaction shocks are excluded because if such shock exists, the (weak) solutions of the problem will not be unique and, more, such a solution does not depend continuously on the initial data (the characteristics cannot be “traced back” to the initial line) and the basic thermodynamic principles are violated.

We shall allow only shocks which do obey the entropy condition. This restriction will make the (weak) solution of the problem unique.

A shock is called *compressive* if the pressure behind the shock is greater than the pressure in front of the shock.

One shows that for a fluid with an equation of state under the form $p = k\rho v^\gamma$ (or, more generally, whose total energy is conserved while the specific energy is given by $e = \frac{1}{2}\rho v^2 + \frac{p}{\gamma-1}$), the fulfilment of the entropy condition holds if and only if the shock is compressive.

It has been proved that, for a perfect gas, the so-called *Weyl hypotheses* are satisfied, which means

$$\frac{\partial p}{\partial \tau} < 0, \quad \frac{\partial^2 p}{\partial \tau^2} > 0, \quad \frac{\partial p}{\partial s} > 0.$$

Then, besides the fact that the knowledge of the values of the flow parameters before the shock together with the shock displacement velocity allows the determination of the flow parameters behind the shock, the following properties across the shock take place:

1) There is an entropy increase which is of order 3 in $\tau_0 - \tau_1$ or in $p_1 - p_0$;

2) The pressure and the specific mass increase such that the shock is compressive ($p_1 > p_0$ and $\tau_1 < \tau_0$);

3) The normal component of the fluid velocity vs. the shock front is supersonic before the shock, becoming subsonic after shock. Further, the fluid flow before the shock will obviously be supersonic while after shock it will be subsonic, the shock waves arising only within the supersonic flows.

One can show that the Weyl hypotheses are satisfied by other gases too.

3.6 The Unique Form of the Fluid Equations

In the sequel we will analyze the conservative form of all the equations associated with fluid flows — the equations of continuity, of momentum tensor and of energy within a unique frame. Then we will show which are the most appropriate forms for CFD. We notice, first, that all the mentioned equations (even on axes projection if necessary) could be framed in the same generic form

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} + \frac{\partial \mathbf{H}}{\partial z} = \mathbf{J}$$

where \mathbf{U} , \mathbf{F} , \mathbf{G} , \mathbf{H} and \mathbf{J} are column vectors given by

$$\mathbf{U} = \begin{cases} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho v_3 \\ \rho \left(e + \frac{v^2}{2} \right) \end{cases}, \quad \mathbf{J} = \begin{cases} 0 \\ \rho f_1 \\ \rho f_2 \\ \rho f_3 \\ \rho (v_1 f_1 + v_2 f_2 + v_3 f_3) + \rho \frac{\delta q}{dt} \end{cases},$$

$$\mathbf{F} = \begin{cases} \rho \\ \rho v_1^2 + p - \sigma_{11} \\ \rho v_2 v_1 - \sigma_{12} \\ \rho v_3 v_1 - \sigma_{13} \\ \rho \left(e + \frac{v^2}{2} \right) v_1 + p v_1 - v_1 \sigma_{11} - v_2 \sigma_{12} - v_3 \sigma_{13} \end{cases},$$

$$\mathbf{G} = \begin{cases} \rho v_2 \\ \rho v_1 v_2 - \sigma_{21} \\ \rho v_2^2 + p - \sigma_{22} \\ \rho v_3 v_2 - \sigma_{23} \\ \rho \left(e + \frac{v^2}{2} \right) v_2 + p v_2 - v_1 \sigma_{21} - v_2 \sigma_{22} - v_3 \sigma_{23} \end{cases},$$

$$\mathbf{H} = \begin{cases} \rho v_3 \\ \rho v_1 v_3 - \sigma_{31} \\ \rho v_2 v_3 - \sigma_{32} \\ \rho v_3 v_3 + p - \sigma_{33} \\ \rho \left(e + \frac{v^2}{2} \right) v_3 + p v_3 - v_1 \sigma_{31} - v_2 \sigma_{32} - v_3 \sigma_{33} \end{cases},$$

where σ_{ij} are the components of the tensor $[\boldsymbol{\sigma}]$, f_i of the vector \mathbf{f} and v_i of the vector \mathbf{v} .

In the above equations the column vectors \mathbf{F} , \mathbf{G} and \mathbf{H} are called the flux terms while \mathbf{J} is a “source” term (which will be zero if the external forces are negligible). For an unsteady problem \mathbf{U} is called the solution vector because its elements are dependent variables which

can be numerically evaluated by considering, usually, some time steps. Therefore using this approach one calculates numerically the elements of \mathbf{U} instead of genuine variables v_1, v_2, v_3 and e . Of course once the numerical values for the \mathbf{U} components are determined, the numerical values for the genuine parameters are immediately obtained by $\rho = \rho$,

$v_i = \frac{\rho v_i}{\rho}$, $e = \frac{\rho \left(e + \frac{v^2}{2} \right)}{\rho} - \frac{\sum v_i^2}{2}$. In the case of the inviscid fluids we will follow the same procedure with the simplification $\sigma_{ij} = 0$.

In the case of the steady flow, we have $\frac{\partial \mathbf{U}}{\partial t} = 0$. That is why for such problems one frequently uses numerical techniques of marching type. For instance if the solution of the problem is obtained via a marching procedure in the direction of the x axis, then our equation could be written in the form $\frac{\partial \mathbf{F}}{\partial x} = \mathbf{J} - \frac{\partial \mathbf{G}}{\partial y} + \frac{\partial \mathbf{H}}{\partial z}$.

Here \mathbf{F} becomes the “solution” vector while the dependent variables are ρv_1 , $\rho v_1^2 + p$, $\rho v_1 v_2$, $\rho v_1 v_3$ and $\left[\rho v_1 \left(e + \frac{v^2}{2} \right) + p v_1 \right]$. From these variables it would be possible to get again the genuine variables even if this time the calculations are more complicated than in the previous case.

Let us notice now that the generic form considered for our equations contains only the first order derivatives with respect to x_i and t and all these derivatives are on the left side, which makes it a *strong* conservative form. This is in opposition with the previous forms of our equations (for instance the energy equation) where the spatial coordinates derivatives could occur on the right side too. That is why these last equations are considered to be in a *weak* conservative form.

The strong conservative form is the most used in CFD. To understand “why”, it would be sufficient to make an analysis of the fluid flows which involve some “shock waves”. We will see later, that such flows imply discontinuities in variable p , ρ , u_i , T etc. If for determining of such flows we would use, for instance, the so-called “shock capturing” method, the strong conservative form leads to such numerical results that the corresponding fluid is smooth and stable, while the other forms of these equations lead to unrealistic oscillations, to an incorrect location of the discontinuities (the shock) and to unstable solutions. The main reason for this situation consists in the remark that whereas the “genuine” variables are discontinuous, the dependent variables like ρv and $p + \rho v^2$ are continuous across the shock wave (Rankine–Hugoniot relations).

Chapter 2

DYNAMICS OF INVISCID FLUIDS

The inviscid (ideal) fluids are hypothetical fluids in which the viscosity is neglected and consequently there is no opposition while the fluid layers slide “one on another”. Although such fluids don’t occur in nature, their study offers useful information in the regions far enough from the solid surfaces embedded in fluids. At the same time the neglect of viscosity (i.e., all the coefficients of viscosity are zero) simplifies considerably the flow equations (Euler) which allows a deep approach via the classical calculus. Nowadays the interest has been renewed in inviscid fluid flows because up-to-date computers are capable of solving their equations, without any other simplifications for problems of great practical interest. It is also interesting to note that for $R' = \infty$ (the inviscid fluid case) we have accomplished the conditions for a “perfect continuum”, the Knudsen number K_n being zero [153].

The target of this chapter is to set up the main results coming from the Euler flow equations which allows a global understanding of flow phenomena in both the incompressible and compressible case. Obviously, due to the high complexity of the proposed aim, we will select only the most important results within the context of numerical and computational methods.

1. Vorticity and Circulation for Inviscid Fluids. The Bernoulli Theorems

Suppose that in the equations of vorticity under the hypothesis that the external forces derive from a potential, which means in

$$\frac{D\boldsymbol{\omega}}{Dt} = (\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v} - \boldsymbol{\omega} (\text{div} \mathbf{v}) + \text{rot} \left(-\frac{\text{grad} p}{\rho} + \frac{1}{\rho} \text{div}[\boldsymbol{\sigma}] \right)$$

we set $[\boldsymbol{\sigma}] = \mathbf{0}$, then we get

$$\frac{D\boldsymbol{\omega}}{Dt} = (\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v} - \boldsymbol{\omega} (\text{div} \mathbf{v}) + \text{rot} \left(-\frac{\text{grad} p}{\rho} \right).$$

For a barotropic fluid (obviously for an incompressible fluid too) because $\text{grad} \int \frac{dp}{\rho} = \frac{\text{grad} p}{\rho}$ and taking into account the equation of continuity, it turns out that $\text{div} \mathbf{v} = -\frac{1}{\rho} \frac{D\rho}{Dt}$, so that we obtain

$$\frac{D}{Dt} \left(\frac{\boldsymbol{\omega}}{\rho} \right) = \left(\frac{\boldsymbol{\omega}}{\rho} \cdot \text{grad} \right) \mathbf{v} = (\text{grad} \mathbf{v}) \frac{\boldsymbol{\omega}}{\rho}.$$

Similarly, from

$$\frac{D\Gamma}{Dt} = \int_C \left(-\frac{\text{grad} p}{\rho} + \frac{1}{\rho} \text{div}[\boldsymbol{\sigma}] \right) \cdot \delta \mathbf{r},$$

we get $\frac{D\Gamma}{Dt} = -\int \frac{\text{grad} p}{\rho} \cdot \delta \mathbf{r}$ such that, for a barotropic fluid, we finally have $\frac{D\Gamma}{Dt} = 0$. This result, also known as *the Thompson (Lord Kelvin) theorem*, states that the circulation along a simple closed curve, observed during its motion, is *constant* whenever the fluid is inviscid (ideal), barotropic (or incompressible) and the mass (external) forces are potential.¹ Correspondingly, in the above conditions, the strength of a vortex tube is a constant too (Helmholtz).

In the case of the ideal incompressible or barotropic compressible fluid flows, the vorticity (rotation) equation (obtained by taking the *curl* of each term of the Euler equation) could be written as $\frac{\partial \boldsymbol{\omega}}{\partial t} + \text{rot}(\boldsymbol{\omega} \times \mathbf{v}) = 0$. On the other hand, if we consider the flux of rotation (vorticity) across a fluid surface Σ , that is $\Phi = \iint_{\Sigma} \boldsymbol{\omega} \cdot \mathbf{n} d\sigma$, as $\text{div} \boldsymbol{\omega} = 0$ and the formula

[153],

$$\frac{d\Phi}{dt} = \iint_{\Sigma} \left[\frac{\partial \boldsymbol{\omega}}{\partial t} + \text{rot}(\boldsymbol{\omega} \times \mathbf{v}) \right] \cdot \mathbf{n} d\sigma$$

holds, we can state the following theorem:

¹ The Thompson theorem requires, basically, the existence of a uniform potential of accelerations. Some recent results, which have also taken into consideration the case of nonuniform potential of accelerations, should be mentioned [122].

THEOREM 2.1. *The rotation (vorticity) flux across a certain part of a fluid surface (which is watched during its motion) is constant.*

As direct consequences of this theorem we have the following results which can be proved by “reductio ad absurdum”:

- A fluid surface Σ , which at a certain instant t_0 is a rotation (vorticity) surface will preserve this quality all the time, i.e., it will be a rotation (vorticity) surface during the motion. A similar result could also be formulated for the vorticity (rotation) lines, these lines being defined as the intersection of two vorticity (rotation) surfaces;

- If, at a certain moment, the fluid flow is irrotational (potential), then this quality will be kept at any later moment.

This last result, known as *the Lagrange theorem* and which is valid in the above mentioned hypotheses, could be obtained either by reductio ad absurdum (supposing that the flux of rotation across a certain surface, with $\boldsymbol{\omega} \neq 0$, would be different from zero which leads obviously to a contradiction) or by remarking that the equation $\frac{D}{Dt} \left(\frac{\boldsymbol{\omega}}{\rho} \right) = (\text{grad } \mathbf{v}) \frac{\boldsymbol{\omega}}{\rho}$ has the solution (in Lagrangian coordinates) $\frac{\boldsymbol{\omega}}{\rho} = \frac{\partial \mathbf{r}}{\partial X^\alpha} \frac{\boldsymbol{\omega}_0^\alpha}{\rho_0}$, where $\boldsymbol{\omega}_0$ ($\boldsymbol{\omega}_0^\alpha$) is the vorticity vector at the moment t_0 and $\rho = \rho_0$ is the mass density at the same moment.

If the fluid flow is irrotational, then there will be a velocity potential Φ such that $\mathbf{v} = \text{grad } \Phi$. As $\frac{\partial \mathbf{v}}{\partial t} = \text{grad} \frac{\partial \Phi}{\partial t}$, from Euler’s equation in Helmholtz form, in the same hypotheses of a barotropic fluid and of the conservative character of the external forces, we also get

$$\text{grad} \left(\frac{\partial \Phi}{\partial t} + \frac{v^2}{2} + \int \frac{dp}{\rho} - U \right) = 0.$$

In other words, in an irrotational flow of an inviscid barotropic fluid with external forces coming from a potential U , we have $\frac{\partial \Phi}{\partial t} + \frac{v^2}{2} + \int \frac{dp}{\rho} - U = C(t)$, where $C(t)$ is a function depending only on time (in the steady case this function becomes a constant, which does not change its value in the whole fluid domain). This result, known as *the second Bernoulli theorem (integral)* could be also extended in the case of a rotational fluid flow. Precisely, by considering the inner product of both sides of Euler’s equation with \mathbf{v} , we will have that $\rho \frac{\partial}{\partial t} \frac{v^2}{2} + \rho \mathbf{v} \cdot \text{grad} K = 0$, where $K = \frac{v^2}{2} + \int \frac{dp}{\rho} - U$.

If the flow is steady, then we will have at once $\mathbf{v} \cdot \text{grad} K = \frac{DK}{Dt} = 0$, i.e., the quantity $K = \frac{v^2}{2} + \int \frac{dp}{\rho} - U$ is constant at any path line, the value of this constant being different when we change the trajectory. This last result is known as *the first Bernoulli theorem (integral)*.

Now we remark that the above quantity K also satisfies, in the steady state case, the equation $\boldsymbol{\omega} \times \mathbf{v} = -\text{grad}K$ and, correspondingly, $\mathbf{v} \cdot \text{grad}K = 0$ which could be obtained from the Euler equation in the Helmholtz form, with the same previous assumptions. Consider the energy equation for an inviscid fluid with no heat change with its surrounding ($\delta q \equiv 0$) and with a time-free potential of the external forces, that is

$$\rho \frac{D}{Dt} \left(e + \frac{v^2}{2} \right) = -\text{div}(\rho \mathbf{v}) + \rho \text{grad}U \cdot \mathbf{v} = -\rho \frac{D}{Dt} \left(\frac{p}{\rho} \right) + \rho \frac{D}{Dt} U$$

or $\frac{dH}{dt} = 0$, where $H = \frac{1}{2}v^2 + e + \frac{p}{\rho} - U$.

The energy equation shows that $H = \text{constant}$ on each streamline. From the expression of H we get, by taking the grad operator and using the equality $\text{grad} \int \frac{dp}{\rho} + p \text{grad} \left(\frac{1}{\rho} \right) = \text{grad} \left(\frac{p}{\rho} \right)$, that

$$\text{grad} H = \text{grad} \left(\frac{1}{2}v^2 + \int \frac{dp}{\rho} - U \right) + \text{grad} e + p \text{grad} v,$$

where $v = \frac{1}{\rho}$ is the specific volume.

At the same time the first law of thermodynamics written under the “gradient” form, i.e., $T \text{grad} s = \text{grad} e + p \text{grad} v$, allows us to write that $\text{grad} H = T \text{grad} s + \text{grad} B$ or

$$\text{grad} H = T \text{grad} s - \boldsymbol{\omega} \times \mathbf{v}.$$

The last equality is known as *the Crocco–Vazsonyi equation* and it shows that H is constant in the whole domain of the flow provided that $s = \text{constant}$ and $\boldsymbol{\omega} = 0$. In other words, for the isentropic steady potential fluid flows H is constant together with K .

In the absence of the external forces $H = h_0$, where h_0 is the enthalpy at the zero velocity (stagnation) points. In this case the Crocco – Vazsonyi equation can be written in the simplified form as $\text{grad} h_0 = T \text{grad} s - \boldsymbol{\omega} \times \mathbf{v}$.

Generally, the values of the constants taken by K and H along a certain streamline, in the steady case, are different. But in the case of isentropic flows ($s = \text{constant}$), the constants for K and H will be the same.

It has been shown that the modification of these constants while the streamlines are changing (which does not occur in the case of irrotational flows) is a direct consequence of the existence of the rotational feature of the whole fluid flow [153].

2. Some Simple Existence and Uniqueness Results

In what follows we will present, successively, some existence and uniqueness results for the solutions of the Euler system (equations). A special accent is put on the uniqueness results because, in fluid dynamics, there is a large variety of methods, not necessarily direct (i.e., they could also be inverse, semi-inverse, etc.), which enable us to construct a solution fulfilling the given requirements and which, if a uniqueness result already exists, will be *the right solution we were looking for*.

At the same time we will limit our considerations to the “strong” solutions, i.e., the solutions associated to the continuous flows, while the other solutions (weak, etc.) will be considered within a more general frame, in the next chapter.

We will start by focussing on additional requirements concerning the associated boundary conditions. The slip-conditions on a rigid wall — which are necessary conditions for any deformable continuum and which, in the particular case of the inviscid fluids are proved to be also sufficient for the mathematical coherence of the joined model — take the known form $\mathbf{v} \cdot \mathbf{n} = 0$ or, when the wall is moving, $\mathbf{v}_r \cdot \mathbf{n} = 0$ (\mathbf{v}_r being the relative velocity of the fluid versus the wall).

If our fluid is in contact with another ideal fluid, the contact surface (interface) is obviously a material surface whose shape is not “a priori” known. But we know that across such an interface the stress should be continuous. As in the case of the ideal fluid the stress comes to the pressure, we will have that across this contact surface of (unknown) equation $F = 0$, there are both $\dot{F} = 0$ (the Euler–Lagrange criterion for material surfaces) and $p_1 = p_2$ (p_1 and p_2 being the limit values of the pressure at the same point of the interface, a point which is “approached” from the fluid (1), and from the fluid (2) respectively). The existence of two conditions, *the kinematic condition* ($\dot{F} = 0$) and *the dynamic condition* ($p_1 = p_2|_{F=0}$) does not lead to an over-determined problem because this time, we should not determine only the solution of the respective equation but also the shape of the boundary $F = 0$, the boundary which carries the last data. In other words, in this case, we deal with an inverse problem.

If the flow is not adiabatic we will have to know either the temperature $T(\mathbf{r}, t)$ or the vector \mathbf{q} on the boundary of the flow domain.

If the flow is adiabatic, from the energy equation we will have $\dot{s} = 0$ and, if the fluid is also perfect $s = c_v \ln \frac{p}{\rho^\gamma} + s_0$, the Euler system will have five equations with five (scalar) unknowns \mathbf{v} , p , ρ . If, additionally,

the flow is homentropic then (as we have already seen) the fifth equation will be $p = K\rho^\gamma$.

Concerning the initial conditions for the Euler equations, they arise from the evolution character of these equations. Such initial conditions imply that we know p , ρ , T and \mathbf{v} at an “initial” moment so that these conditions, together with the Euler equations, set up a Cauchy problem. From the classical Cauchy–Kovalevski theorem we can conclude that this Cauchy problem for the Euler system (with $\mathbf{f} \equiv 0$), the equation of continuity, the constancy of entropy on each path line ($\dot{s} = 0$, which means in adiabatic evolution) and the state equation $p = p(s, \rho)$, together with the initial conditions $\mathbf{v}|_{t=0} = \mathbf{v}^0(\mathbf{r})$, $\rho|_{t=0} = \rho^0(\mathbf{r})$, $s|_{t=0} = s^0(\mathbf{r})$, $\mathbf{r} \in \mathcal{R}$, where $\inf_{\mathbf{r} \in \mathcal{R}} \rho^0(\mathbf{r}) = \rho^0 > 0$, is a well-posed problem and for any initial data and analytical state equations, i.e., there is a unique *analytical* solution defined on the domain $\mathcal{V} = \{\mathbf{r} \in \mathbb{R}^3, |t| < T(\mathbf{r})\} \subset \mathbb{R}^4$, where $T(\mathbf{r})$, for any \mathbf{r} , is a function depending continuously on initial data in the metrics of analytical spaces.

Of course the above mentioned result is a locally time existence and uniqueness theorem which is valid only for continuous functions (data and solution).

Generally, there are not global (for all time) existence and uniqueness results, excepting the two-dimensional case due to the vorticity conservation $(\frac{d\omega}{dt} = 0)^2$. Nevertheless the practical applications require certain sharp global uniqueness conditions for the Cauchy problem or more generally for the Cauchy mixed problem (with also boundary conditions, at any time t) associated with the Euler system.

Before presenting such uniqueness results we remark that the “non-uniqueness” of the Euler system solution would be linked to the “suddenness” of the approximation of a viscous and non-adiabatic fluid by an ideal fluid in adiabatic evolution. R. Zeytonnian³ has shown that the loss of the boundary conditions associated with the mentioned approximation, in the circumstances of the presence of some bodies of “profile type”, could be completed by the introduction of some Joukowski type conditions (to which we will return) while in the case of some bodies of “non-profile type”, the model should be corrected by introducing a vortices separation (vortex sheets).

Let now $U = (v_1, v_2, v_3, \rho, s)^T$ be a solution of the Euler system for $t \geq 0$, a solution which is defined in a bounded domain $\Omega \subset \mathbb{R}^4$. We accept that the boundary of this domain is composed of a three-dimensional spatial domain ω_0 , enclosed in the hyperplane $t = 0$, and by a sectionally

²See R. Zeytonnian, *Mécanique Fondamentale des Fluides*, t.1, pp. 154 – 158 [160].

³See R. Zeytonnian, *Mécanique Fondamentale des Fluides*, t.1, p. 126 [160].

smooth hypersurface Γ (for $t > 0$) which has a common border with the domain ω_0 . Let also $\xi = (\xi_0, \xi_1, \xi_2, \xi_3)$ be the outward unit normal to Γ . It is proved that the uniqueness of U in Ω is intimately linked with the hyperbolicity of the Euler equations which requires the fulfilment of the following complimentary hypothesis: at each point of the hypersurface Γ the inequality

$$\xi_0 + v_1 \xi_1 + v_2 \xi_2 + v_3 \xi_3 \geq a (\xi_1^2 + \xi_2^2 + \xi_3^2)^{\frac{1}{2}} \tag{2.1}$$

should be satisfied.

More precisely, one states that ([160]) if the solution U of the Euler system exists in the class $C^1(\Omega)$ and this solution satisfies the condition (2.1), while $\inf_{\omega_0} \rho^0(\mathbf{r}) > 0$, then for any other solution $U' \in C^1(\Omega)$ of the Euler system, one could find a constant $k_0 > 0$ such that $\delta U = U' - U$ fulfils $\|\delta U; t\| \leq k_0 \|\delta U; t = 0\|^4$ for $t > 0$. Consequently if the equality $U' = U$ holds on ω_0 (that means in the hyperplane $t = 0$), then it will be satisfied at any point-moment $(\mathbf{r}, t) \in \Omega(\omega_0)$. Obviously $\Omega(\omega_0)$, called the determination domain for the solution of the Cauchy problem with the initial data on ω_0 , is the union of all the domains which back on ω_0 and on whose boundary the inequality (2.1) is satisfied.

It has been also proved that if $\Gamma(\omega_0)$ is a smooth boundary (of C^1 class) of the determination domain $\Omega(\omega_0)$, then this hypersurface will be a characteristic surface of the Euler system, the inequality sign of (2.1) being replaced by that of equality.

We now remark that in the conditions of an Euler system in adiabatic evolution with a state equation $p = p(\rho, s)$ of C^2 class, assuming that the domain $D(t)$ of the fluid flow has the boundary $\Sigma(t)$, which is composed of both rigid and “free” parts, and v_p is the propagation velocity of the surface Σ [33] then, if

- (i) $\mathbf{v}(\mathbf{r}, t), \rho(\mathbf{r}, t), s(\mathbf{r}, t)$ are functions of class C^1 on $[0, T] \times D$,
- (ii) the initial conditions $\mathbf{v}(\mathbf{r}, 0), \rho(\mathbf{r}, 0), s(\mathbf{r}, 0)$ are given together with
- (iii) the boundary conditions $\mathbf{v}_r \cdot \mathbf{n} = 0$ on $[0, T] \times \Sigma$ and, similarly, \mathbf{v}, ρ, s in the regions where $v_p < 0$,

then the Euler system (even with $\mathbf{f} \neq 0$), in adiabatic evolution, with the state equation $p = p(\rho, s)$, has a unique solution⁵.

The uniqueness is still kept even in the case when there are not boundary conditions at the points of Σ where $v_p \geq c$, c being the speed of sound.

⁴For the definition of the norm we deal with, we should first consider all the cuts $\omega(t)$ of Ω by the hyperplane $t = \text{constant} > 0$. Then by introducing the vectorial function $\mathbf{v} = \{v_i\}$ on Ω , its norm corresponding to the cut $\omega(t)$ will be defined by $\|\mathbf{v}; t\| = \iiint_{\omega(t)} (\sum v_i^2) d\omega$.

⁵J.Serrin [135].

In the case of the incompressible inviscid isochrone ($\frac{d\rho}{dt} = 0$) or barotropic compressible fluid flows, Dario Graffi has given a uniqueness result which requires [57]:

(i) the functions $\mathbf{v}(\mathbf{r}, t)$, $\rho(\mathbf{r}, t)$ and $p(\mathbf{r}, t)$ are continuously differentiable with bounded first derivative on $[0, T] \times D$,

(ii) the initial conditions $\mathbf{v}(\mathbf{r}, 0)$, $\rho(\mathbf{r}, 0)$, the boundary conditions $\mathbf{v} \cdot \mathbf{n}$ and the external mass forces \mathbf{f} are given, respectively, on Σ and $[0, T] \times \Sigma$,

(iii) the state equation (in the barotropic evolution) is of the C^2 class.

We remark that these results keep their validity if D becomes unbounded — the most frequent case of fluid mechanics — under the restriction of a certain asymptotic behaviour at far distances (infinity) for the magnitude of velocity, pressure and mass density, namely of the type

$$v = v_\infty + O\left(r^{-(\frac{3}{2}+\varepsilon)}\right), p = p_\infty + O\left(r^{-(\frac{3}{2}+\varepsilon)}\right), \rho = \rho_\infty + O\left(r^{-(\frac{3}{2}+\varepsilon)}\right),$$

where ε is a positive small parameter.

We conclude this section with a particular existence and uniqueness result which implies an important consequence about the nonexistence of the Euler system solution for the incompressible, irrotational and steady flows.

More precisely, if D is a simply connected and bounded region, whose boundary ∂D moves with the velocity \mathbf{V} , it can easily shown that [19]:

(i) there is a unique incompressible, potential, steady flow in D , if and only if $\int_{\partial D} \mathbf{V} \cdot \mathbf{n} ds = 0$,

(ii) this flow minimizes the kinetic energy $E_{cin} = \frac{1}{2} \int_D \rho v^2 dv$ over all the vectors \mathbf{u} with zero divergence and satisfying $\mathbf{u} \cdot \mathbf{n}|_{\partial D} = \mathbf{V} \cdot \mathbf{n}|_{\partial D}$.

We remark that this simple result, through (ii), associates to the problem of solution determining a minimum problem for a functional, that is a variation principle. Such principles will be very useful in numerical approaches to the fluid dynamics equations and we will return to them them later in this book.

At the same time if our domain D is bounded and with fixed boundary ∂D ($\mathbf{V} \equiv 0$), only the trivial solution $\mathbf{v} \equiv 0$ (the rest) corresponds to a potential incompressible steady flow. Obviously in the case of the unbounded domains this result will be not true provided that the boundary conditions on ∂D should be completed with the behaviour at infinity.

The same result (the impossibility of an effective flow) happens even if the domain is the outside of a fixed body or a bodies system, the fluid flow being supposed incompressible with uniform potential (without circulation) and at rest at infinity.

3. Irrotational Flows of Incompressible Inviscid Fluids. The Plane Case

The Lagrange theorem, stated in the first section of this chapter, establishes the conservation of the irrotational character of certain fluid flows. An important application of this theorem is the case when the fluid starts its flow from an initial rest state (where, obviously, $\boldsymbol{\omega} \equiv 0$).

If a fluid flow is irrotational, then from the condition $\text{rot } \mathbf{v} = 0$ we will deduce the existence of a scalar function $\Phi(x_1, x_2, x_3, t)$, defined to within an additive function of time, such that $\mathbf{v} = \text{grad } \Phi$. Obviously, the determination of this function, called *the velocity potential*, is synonymous with that of the velocity field. But from the equation of continuity we also get $0 = \text{div } \mathbf{v} = \text{div}(\text{grad } \Phi) = \Delta\Phi$, while the slip condition on a fixed wall (Σ), immersed in the fluid, becomes

$$0 = \mathbf{v} \cdot \mathbf{n}|_{\Sigma} = \text{grad } \Phi \cdot \mathbf{n}|_{\Sigma} = \left. \frac{d\Phi}{dn} \right|_{\Sigma},$$

that is the determination of Φ comes to the solving of a boundary value problem of Neumann type joined to the Laplace operator.

Obviously, if the domain flow is “unbounded” we need some behaviour conditions at far distances (infinity) which, in the hypothesis of a fluid stream “attacking” with the velocity \mathbf{v}_{∞} an obstacle whose boundary is (Σ), implies that

$$\lim_{x_1^2 + x_2^2 + x_3^2 \rightarrow \infty} \text{grad } \Phi = \mathbf{v}_{\infty}.$$

So that in this particular case the flow determining comes either to a Neumann problem for the Laplace operator (the same problem arises in the tridimensional case too), that means $\Delta\Phi = 0$ in the fluid domain D with $\left. \frac{d\Phi}{dn} \right|_{\partial D} = 0$, or to a Dirichlet problem for the same Laplace operator (which is specific only in the 2-dimensional case) when $\Delta\psi = 0$ in D with $\psi|_{\partial D} = \text{constant}$.

In the conditions of an unbounded domain (the case of a flow past a bounded body being included too), the above two problems should be completed by information about the velocity (that is about $\text{grad } \Phi$ and $\text{grad } \psi$ respectively) at far distances (infinity).

Now we will show that in a potential flow past one or more body(ies), the maximum value for the velocity is taken on the body(ies) boundary. If M is an arbitrary point in the fluid which is also considered the origin of a system of axes, the Ox axis being oriented as the velocity at M , then we have $v^2(M) = \Phi_x^2(M)$, while for any other point P , we have $v^2(P) = (\Phi_x^2 + \Phi_y^2 + \Phi_z^2)(P)$.

If the function Φ_x is harmonic and consequently it does not have an extremum inside the domain, then there will always be some points P

so that $\Phi_x(P) > \Phi_x(M)$, which means $v^2(P) > v^2(M)$. In other words the unique possibility for the velocity to get a maximum value is only on the boundary. Concerning the minimum value of the velocity this could be reached inside the domain, namely in the so-called *stagnation points* (with zero velocity). If the fluid flow is steady and the external forces can be neglected, from the second Bernoulli theorem (integral) it comes that, at a such stagnation point, the pressure has a maximum while at boundary points of maximum velocity, the pressure should have a minimum.

Let us now consider the case of an incompressible irrotational plane (2-dimensional) fluid flow.

Let Oxy be the plane where we study the considered fluid flow, u and v being the velocity vector components on Ox and Oy respectively, and q the magnitude of this vector. The fluid being incompressible, the equation of continuity can be written $\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$, such that $u dx - v dy$ is, for every fixed t , an exact total differential in x and y . Consequently, there is a function $\psi(x, y, t)$, defined to within an additive function of time by the equality $u dy - v dx = d\psi$, where t is seen as a parameter and not as an independent variable.

This function $\psi(x, y, t)$ is *the stream function* of the flow since the curves $\psi = \text{constant}$, at any fixed moment t , define the streamlines of the flow that has been shown. On the other side, the flow being irrotational, we also have $\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} = 0$ which proves the existence of a second function $\Phi(x, y, t)$, *the velocity potential*, defined also to within an additive function of time, such that $u dx + v dy = d\Phi$ where again t is considered a parameter and not an independent variable. Hence

$$u = \frac{\partial \Phi}{\partial x} = \frac{\partial \psi}{\partial y}, \quad v = \frac{\partial \Phi}{\partial y} = -\frac{\partial \psi}{\partial x}$$

or, under vectorial form

$$\mathbf{v} = \text{grad}\Phi = -\mathbf{k} \times \text{grad}\psi,$$

\mathbf{k} being the unit vector of the axis directly perpendicular on the plane Oxy .

But these equalities show that the two functions Φ and ψ satisfy the classical Cauchy-Riemann system and, consequently, the function $f = \Phi + i\psi$ is a monogenic (analytic) function of the complex variable $z = x + iy$ which could depend, eventually, on the parameter t . This function is called *the complex potential* of the flow and it is obviously defined to within an additive function of time. The real and imaginary part of $f(z)$, which means the velocity potential and the stream function of the flow, are two conjugate harmonic functions; the equipotential lines

$\Phi = \text{constant}$ and the streamlines $\psi = \text{constant}$ form, at any point of the fluid flow, an orthogonal network, the inner product $\text{grad } \psi \cdot \text{grad } \Phi$ being zero. At the same time we also have

$$\frac{df}{dz} = \frac{\partial \Phi}{\partial x} + i \frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial y} - i \frac{\partial \Phi}{\partial y} = u - iv.$$

The function $\frac{df}{dz} = u - iv$ is also an analytic function of z , called *the complex velocity* of the flow and which will be denoted by ζ ; the modulus and the argument of ζ define, respectively, the magnitude q of the velocity and the angle ω , with changed sign, made by the velocity vector with the axis Ox , as

$$\zeta = \frac{df}{dz} = u - iv = qe^{-i\omega}.$$

We conclude that the kinematic description, the whole pattern of the considered flow, could be entirely determined by knowing only the analytic function $f(z; t)$, the complex potential of this flow at the considered moment t .

In the previous considerations we have seen that, to any incompressible potential plane fluid flow it is possible to associate a complex potential. It is important to find out if, conversely, any analytic function of z can be seen as a complex potential, i.e., it determines an incompressible irrotational plane flow of an inviscid (ideal) fluid. To answer this question we recall that, from the physical point of view, it is necessary to choose the function f such that its derivative, the complex velocity, is not only an analytic function but also a *uniform* function in the considered domain (\mathcal{D}), so that, at any point of (\mathcal{D}), $\zeta = \frac{df}{dz}$ takes only one value.

Once accomplished this requirement, due to the analyticity of the function at any point of (\mathcal{D}), the conjugate harmonic functions $u(x, y)$ and $-v(x, y)$ (the real and the imaginary part of ζ) satisfy the Cauchy–Riemann system, that is $\frac{\partial u}{\partial x} = -\frac{\partial v}{\partial y}$, $\frac{\partial u}{\partial y} = \frac{\partial v}{\partial x}$; but such a fluid flow should be an incompressible irrotational plane flow of an inviscid fluid. On the other hand, if the domain (\mathcal{D}) is simply connected, we will also deduce that $f(z)$ is analytic and uniform too, which means a *holomorphic* function in (\mathcal{D}). Really, z_0 being the affix of a point of (\mathcal{D}), we have $f(z) = f(z_0) + \int_{z_0}^z \zeta dz$, the integral being taken along an arbitrary arc connecting the points M_0 and M (or z_0 and z). The Cauchy–Goursat theorem proves, ζ being uniform and (\mathcal{D}) simply connected, that the above expression for $f(z)$ does not depend on the chosen arc and consequently $f(z)$ is uniform. It will not be the same if the domain (\mathcal{D}) is

multiply connected. Let (\mathcal{D}) , for example, be the domain sketched in Figure (2.1) where (L_1) and (L_2) are two arcs joining M_0 and M oriented as it is shown; by calculating the integral $\int_{z_0}^z \zeta dz$ along L_1 and then along L_2 , we will get distinct values whose difference is equal to the integral, of function ζ , calculated along the closed contour $L = L_1^- \cup L_2$. On the

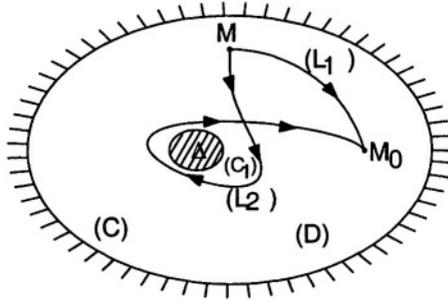


Figure 2.1. The case of a multiply connected domain

other hand it is known that the difference is equal to $m(\Gamma + iD)$, where m is a positive, negative or null integer⁶ while $\Gamma + iD$ is the number given by

$$\Gamma + iD = \int_{(C)} \zeta dz = \int_{(C)} [u dx + v dy + i(udy - vdx)],$$

(C) being a closed contour of (\mathcal{D}) , encircling once, in the direct sense, the domain (Δ) of boundary (C_1) . We remark that $\Gamma = \int_{(C)} \mathbf{v} \cdot d\mathbf{x}$ is

the circulation of the velocity vector when we contour once, in a direct sense, the curve (C) and $D = \int_{(C)} \mathbf{v} \cdot \mathbf{n} ds$ is the flux across (C) , as we

have already made precise.

But then the function $\frac{\Gamma + iD}{2\pi i} \log(z - a)$, where a is the affix of an inside point A of (Δ) , has exactly the same nonuniformity properties as $f(z)$, which means, by deplating along the same (L) the difference between

⁶The modulus of the integer m is the number which expresses how many times the respective contour encircles the simply connected domain (Δ) of boundary (C_1) ; m is negative if the contour is encircled, $|m|$ times, in an inverse sense and it is positive if the encircling is in a direct sense (in the case of Figure 2.1, $m = -1$).

the initial and the final value is again $m(\Gamma + iD)$. Consequently the function $f(z) - \frac{\Gamma+iD}{2\pi i} \log(z - a)$ is uniform, that is holomorphic in (\mathcal{D}) .

We conclude that a function $f(z)$, in the case of a doubly connected domain, could be considered a complex potential if it admits the representation $\frac{\Gamma+iD}{2\pi i} \log(z - a)$ plus a holomorphic function of z .

More generally, the following result holds:

Let $(\Delta_1), (\Delta_2), \dots, (\Delta_p)$ be the connected components of the complement of a bounded domain (\mathcal{D}) , and let $A_q(a_q)$ be a set of internal points of (Δ_q) , respectively $(q = \overline{1, p})$. An analytic function $f(z)$ can be considered a complex potential of a fluid flow in (\mathcal{D}) , if and only if there are a set of real numbers Γ_q and D_q $(q = \overline{1, p})$ such that

$$f(z) - \sum_{q=1}^p \frac{\Gamma_q + iD_q}{2\pi i} \log(z - a_q)$$

is a holomorphic function in (\mathcal{D}) .

Case of steady flows. If the flow is steady u and v will be free of t (they do not depend explicitly on time) and consequently we may suppose that Φ, ψ and $f(z)$ have the same property.

Concerning the effective determination of the complex potential for a certain plane flow, it could be done taking into account the boundary conditions. In the particular case when the fluid past a fixed wall, this wall, due to the slip condition $\mathbf{v} \cdot \mathbf{n} = 0 = d\psi$, is a streamline of our flow and consequently, along this curve, $\psi = \text{Im } f(z)$ is constant. Conversely, if a plane fluid flow is known (given), we could always suppose that a streamline is a “solid wall”, because the slip condition is automatically fulfilled $\left(\frac{d\psi}{ds} = \frac{d\Phi}{dn} = 0\right)$; shortly, we could say that it is possible to *solidify (materialize)* the streamlines of a given flow (under the above assumption).

Finally, supposing that $f(z)$ and implicitly the velocity field are determined, it will always be possible to calculate the pressure at any point of the fluid flow by using the second Bernoulli theorem which can be written as $\mathcal{K} = \frac{q^2}{2} + \frac{p}{\rho} - U = \text{constant}$. To assess this constant it is sufficient to have both the magnitude of the velocity q_1 and the pressure p_1 at a point M_1 belonging to the flow domain. Additionally, if $\mathbf{f} = 0$, we also have $\frac{p-p_1}{\frac{1}{2}\rho q_1^2} = 1 - \frac{q^2}{q_1^2}$. Each of the two sides of the previous equality is non-dimensional. The first one, denoted by C_p , is called *the pressure coefficient*.

Starting from some analytical functions $f(z)$ satisfying the uniformity properties stated above, it could always build up corresponding fluid flows. For instance a linear function $f(z) = az + b$, a and b being

constants, will lead to a *uniform* (constant velocity) flow while the logarithmic functions $f(z) = \frac{D}{2\pi} \log z$ and $f(z) = \frac{\Gamma}{2\pi i} \log z$, defined on the whole plane without its origin (D and Γ being real constants) correspond respectively to a *source* (*sink*) — according to the sign of flow rate D — and to a point *vortex* of circulation Γ , all of them being located at the origin. For practical applications one considers also the so-called *doublet* (*dipole*) of axis Ox and strength (moment) K , located at the origin, whose complex potential is $f(z) = -\frac{K}{2\pi z}$.

Of course all these singular flows could be shifted to another location z_0 of the plane (and even with an axis making an angle α with Ox) by considering the change of coordinates

$$z = z_0 + Ze^{i\alpha}.$$

Properties of the above elementary flows as well as a set of additional examples of such simple flows one finds, for instance, in Caius Iacob's book "Introduction mathématique à la mécanique de fluides", chapter VII, page 407 [69].

We now remark that any linear combination of the complex potentials $f_i(z)$ is still a complex potential in the common definition domain where the analytic functions $f_i(z)$ satisfy the uniformity requirements stated above. Consequently, starting with some given fluid flows, it is always possible to set up, by *superposition*, new flows, that means to consider linear combinations of the respective complex potentials.

For instance by superposition of a uniform flow parallel to the Ox axis, of complex potential $V_0 z$, and of a doublet placed at the origin of complex potential $V_0 \frac{R^2}{z}$ (V_0 and R being positive real constants), one gets the complex potential of the fluid flow past a circular disk (cylinder) of radius R *without circulation*. If we superpose on the previous flow a point vortex located at the origin, which leads to the complex potential

$$f(z) = V_0 \left(z + \frac{R^2}{z} \right) + \frac{\Gamma}{2\pi i} \log z,$$

we obtain the fluid flow past the same disk of radius R but this time *with circulation* Γ .

Detailed considerations on the steady, plane, potential, incompressible flows past a circular obstacle can be found, for instance, in the same [69] or in [52].

4. Conformal Mapping and its Applications within Plane Hydrodynamics

In the previous section we mentioned the technique to build up fluid flows by considering elementary analytic functions. But it will be im-

portant and very useful to have at our disposal more general construction methods for the fluid flows. The conformal mapping will be such a method for determining a fluid flow satisfying some “a priori” given requirements.

Generally, a *conformal transformation* of a domain (d) from the plane (z), onto a domain (D) from the plane (Z), is a holomorphic function $h : (d) \rightarrow (D)$ which fulfils the condition $h'(z) \neq 0$ (the angles preserving condition). If the conformal mapping is also univalent (injective) this will be a *conformal mapping* of the domain (d) onto the domain (D). Obviously the holomorphicity is preserved by a conformal mapping. The same thing happens with the connection order of the domain (d). We know that the determination of the conformal mapping (on a canonical domain) is synonymous with that of the Green function associated to the Laplace operator and to the involved domain, that is with the possibility to solve a boundary value problem of Dirichlet type for the same operator and domain [69].

Concerning the existence of conformal mapping, in the case of a simply-connected domain, a classical result known as Riemann–Carathéodory’s theorem states that:

For a given simply-connected domain (d) from the plane (z) and whose boundary contains more than a point, it is always possible to map it conformally, in a unique manner, onto the circular disk $|Z| < 1$ from the plane (Z), such that to a certain point $z_0 \in (d)$ there corresponds an internal given point Z_0 from $|Z| < 1$ and to a certain direction passing through z_0 there corresponds a given direction passing through Z_0 .

We remark that the uniqueness of the conformal mapping holds to within three arbitrary parameters, so that we deal, basically, with a class of functions which defines the considered conformal mapping.

Unfortunately the proof of the existence in this theorem is far from being a constructive one such that, in practical problems, we are faced with the effective determination of the conformal mapping. There are few cases when these conformal mappings are explicitly (analytically) found. That is why the approximative procedures (one of them being sketched in a next section) are of the greatest interest.

Finally, the above result could also be extended to the doubly-connected domains (see, for instance, Y. Komatu [75]) and even to the general multiply-connected domains but, in this last case, it is extremely difficult to determine and work with the involved functions. As a consequence the conformal mapping method is not practically used in the case of domains with a higher order of connection.

Returning to the simply-connected case, the following result is of remarkable interest in different applications:

THEOREM 2.2. *If (d) is a simply-connected domain from the plane (z) , bounded by a simply closed curve c , and if $Z = h(z)$, a holomorphic function in (d) , has the additional property that when z is displaced along the contour c in a certain sense, its image Z describes a simply closed curve C —delimiting a domain (D) from the plane (Z) , in such a way that the correspondence between c and C is a bijection, then the correspondence between (d) and (D) will also be a bijection and, consequently, the function $Z = h(z)$ will be a conformal mapping of (d) onto (D) .*

Let now $F(Z)$ be the complex potential of a given fluid flow defined in a domain (D) of the plane (Z) ; we suppose as known the function $Z = h(z)$ and its inverse $z = H(Z)$ which establish a conformal mapping between the domain (D) of the plane (Z) and a domain (d) of the plane (z) . Then the function $f(z) = F(h(z))$, with the same regularity properties as $F(Z)$, will be the complex potential of a new fluid flow defined in (d) and called the *associated (transformed) flow* of the given fluid flow by the above mentioned conformal mapping.

Really $f(z)$ could be considered as a complex potential because

$$\frac{df}{dz} = \frac{dF}{dZ} \frac{dZ}{dz} = \frac{dF}{dZ} \frac{dh}{dz}$$

and so $f'(z)$ will be a uniform function in (d) together with $F'(Z)$ in (D) , as well as $h'(z)$ is also uniform together with $h(z)$.

We also remark that in two homologous points z and Z of the considered conformal mapping, we have $f(z) = F(Z)$. But then the values of the velocity potential and of the stream function are equal at such homologous points; consequently, the streamlines and the equipotential lines of the two flows are also homologous within the considered conformal mapping. More, the circulations along two homologous arcs and the rates of the flow across two homologous arcs are equal. Particularly, if a fluid flow defined by $F(Z)$ has a singularity at $Z_0 \in D$ (source, point vortex, etc.), the associated flow will have at the point z_0 , the homologous of Z_0 , a singularity of the same nature and even strength. Of course, at two homologous points the fluid velocities are not (in general) the same, which comes out from the above equalities for the complex velocities.

Concerning the kinetic energy this will be preserved too, as from the relation between the surface elements $dA = |Z'|^2 da$ it results that $\rho v^2 da = \rho V^2 dA$, v and V being the velocities magnitude in the associated flows of the same fluid density ρ .

4.1 Helmholtz Instability

Now we will study the stability of an inviscid, incompressible, parallel fluid flow, containing a velocity discontinuity, following [22]. Precisely, we will suppose that, above the Ox axis, the fluid moves with a uniform velocity U in the positive sense and, below, it moves with a uniform velocity of equal magnitude but in the opposite sense. In this case, the Ox axis represents a discontinuity surface for the velocity and it is the site of a vortex sheet of uniform circulation $2U$ per unit of width. We remember that the circulation is

$$\Gamma = \int V \cdot ds$$

where V is the magnitude of the velocity of the fluid and ds is the arc element along a closed curve encircling the vortex.

Such a vortex sheet is unstable i.e., if a displacement happens the sheet will go away and will not return to its initial position. This could be shown by analytical studies, considering small sinusoidal perturbations. Here we will numerically analyze the time evolution of such perturbations.

We divide the vortex sheet into segments of equal length λ on Ox and each segment will be divided into m equispaced discrete vortices. As the total circulation per unit length is $2U$, each discrete vortex has the circulation $2U\lambda/m$. We will suppose that at the initial moment these vortices are displaced from their initial positions $y_k = 0$ to the positions

$$y_k = a \sin \left(\frac{2\pi x_k}{\lambda} \right), k = \dots - 2, -1, 0, 1, \dots \quad (2.2)$$

Let us consider the row of vortices containing the vortices $k, k \pm m, k \pm 2m, \dots$. The complex potential generated by this row is

$$w_k(z) = \sum_{n=-\infty}^{\infty} i \frac{2U\lambda}{2\pi m} \log(z - z_k - n\lambda) = i \frac{U\lambda}{m\pi} \log \left[\sin \frac{\pi(z - z_k)}{\lambda} \right].$$

Thus the complex potential generated by all the m rows which compose the sheet is

$$w(z) = \sum_{k=1}^m w_k(z) = \sum_{k=1}^m i \frac{U\lambda}{m\pi} \log \left[\sin \frac{\pi(z - z_k)}{\lambda} \right].$$

Replacing this potential in the relation

$$\frac{dw}{dz} = u - iv,$$

by differentiating and separating into the real and imaginary parts we obtain the components u and v of the velocity at the point (x, y) . So, for the vortex j we have

$$\frac{dx_j}{dt} = \frac{U}{m} \sum_{k \neq j}^m \frac{\sinh \left[\frac{2\pi(y_j - y_k)}{\lambda} \right]}{\cosh \left[\frac{2\pi(y_j - y_k)}{\lambda} \right] - \cos \left[\frac{2\pi(x_j - x_k)}{\lambda} \right]} \quad (2.3)$$

and

$$\frac{dy_j}{dt} = -\frac{U}{m} \sum_{k \neq j}^m \frac{\left[\frac{2\pi(x_j - x_k)}{\lambda} \right]}{\cosh \left[\frac{2\pi(y_j - y_k)}{\lambda} \right] - \cos \left[\frac{2\pi(x_j - x_k)}{\lambda} \right]}. \quad (2.4)$$

By introducing the dimensionless variables

$$X = \frac{x}{\lambda}, Y = \frac{y}{\lambda}, A = \frac{a}{\lambda}, T = \frac{tU}{\lambda}$$

the relationships (2.2), (2.3) and (2.4) become

$$Y_j = A \sin(2\pi X_j), T = 0, \quad (2.5)$$

$$\frac{dX_j}{dT} = \frac{1}{m} \sum_{k \neq j}^m \frac{\sinh [2\pi (Y_j - Y_k)]}{\cosh [2\pi (Y_j - Y_k)] - \cos [2\pi (X_j - X_k)]}, \quad (2.6)$$

$$\frac{dY_j}{dT} = -\frac{1}{m} \sum_{k \neq j}^m \frac{\sin [2\pi (X_j - X_k)]}{\cosh [2\pi (Y_j - Y_k)] - \cos [2\pi (X_j - X_k)]}. \quad (2.7)$$

Due to the symmetry and periodicity of the involved functions, the computation is needed only for $j = 2, \dots, m/2$ within a half of the wavelength. The greatest part of this computation involves the above Cauchy problem numerically solving.

The computer result is an animation which shows the evolution of the perturbation in time (see also Figure 2.2).

An enlarged picture of the interest zone, obtained by cubical interpolation of X and Y , is shown in Figure 2.3.

The MATLAB code is

```
global m; m=40;A=0.05;
x=0:1/m:1;y=A*sin(2*pi*x); u0=[x;y];
[t,u]=ode45(@edrol,[0,0.3],u0);
p=plot(x,y,'EraseMode','none');axis([0 1 -0.3 0.3]);
for j=1:length(t) set(p,'color','w');
set(p,'Xdata',u(j,1:m+1),'Ydata',...
u(j,m+2:2*m+2),'color','k');
```

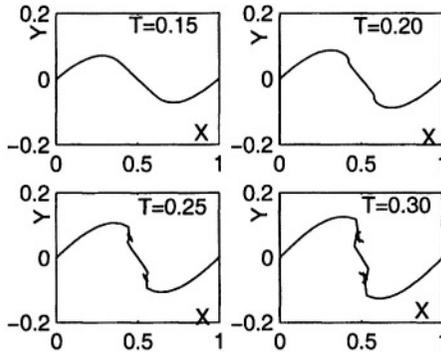


Figure 2.2. Evolution of a vortex sheet after perturbation

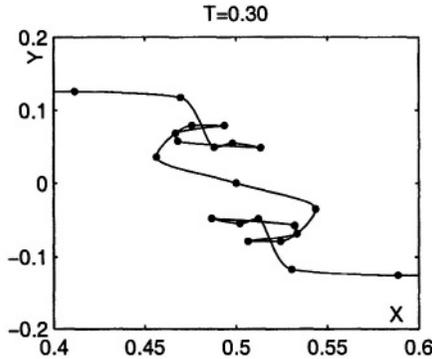


Figure 2.3. Evolution of a vortex sheet after perturbation, $T = 0.30$

```
drawnow;end;
```

The differential system is described by the function M-file `edrol.m`

```
function yprime=edrol(x,y);
global m; disp(x); yprime=zeros(2*m+2,1);
for j=1:m for k=1:m if k~=j
yprime(j)=yprime(j)+1/m*sinh(2*pi*(y(m+1+j)...
-y(m+1+k)))/(cosh(2*pi*(y(m+1+j)-y(m+1+k)))...
-cos(2*pi*(y(j)-y(k))));
yprime(m+1+j)=yprime(m+1+j)-1/m*sin(2*pi*(y(j)...
-y(k)))/(cosh(2*pi*(y(m+1+j)-y(m+1+k)))...
-cos(2*pi*(y(j)-y(k))));
yprime(m+1)=yprime(1); yprime(2*m+2)=yprime(m+2);
end; end; end;
```

5. Principles of the (Wing) Profiles Theory

5.1 Flow Past a (Wing) Profile for an Incidence and a Circulation “a priori” Given

Let (c) be a contour — the right section, in the working plane, of an arbitrary cylinder; in aerohydrodynamics such a cylinder could be seen as an airfoil or a wing of a very large (“infinite”) span (to ensure the plane feature of the flow) and the respective right section (c) is called *wing profile* or shorter *profile*⁷.

The main problem of the theory of profiles is to study the steady flow of a fluid past a profile (obstacle), a flow which behaves at infinity (that means for $|z|$ very large) as a uniform flow of complex velocity

$$V_0 e^{-i\alpha} = V_0 \cos \alpha - iV_0 \sin \alpha.$$

By *incidence* of the profile with respect to Ox , we will understand the angle α made by the velocity vector at far field (infinity) with the x - axis. Besides the incidence of the profile let us also establish precisely (“a priori”) the *circulation* Γ of the flow around the profile.

The determination of the complex potential comes then to the search for an analytic function $f(z)$ such that:

- 1) $f(z) - \frac{i\Gamma}{2\pi} \log z$ is an analytic and uniform function in (d) ;
- 2) its imaginary part is constant along (c) ;
- 3) $\lim_{|z| \rightarrow \infty} \frac{df}{dz} = V_0 e^{-i\alpha}$.

Let (D) be the domain of the plane (Z) defined by $|Z| > R$ and let $z = H(Z)$ or $Z = h(z)$ be the canonical conformal mapping⁸ which maps (D) onto the domain (d) , the exterior of the given profile (c) .

The complex potential $F(Z)$ of the associated (transformed) flow will satisfy the properties 1), 2) and 3) provided that f and z are replaced by F and Z , while (d) and (c) are replaced, respectively, by (D) and (C) . More precisely, the fulfilment of the conditions 1) and 2) comes from the already studied parallelism between $f(z)$ and $F(Z)$, while the condition

⁷With regard to the geometry of profiles, some additional considerations can be found, for instance, in the Caius Iacob book “Introduction mathématique à la mécanique des fluides”, pp. 652-654 [69]. In this book, starting with p. 435, some special classes of profiles are envisaged too.

⁸We recall the following basic theorem: “There is a unique conformal mapping, called canonical, of the domain (d) – the outside of the closed contour (c) – onto the outside of a circular circumference (C) of radius R , centered at the origin, a mapping which in $V(\infty)$ admits a development in the form $z = Z + \sum_{n=0}^{\infty} \frac{a_n}{z^n}$. The radius R of the circumference (C) is an “a priori” unknown length which depends only on the given contour (c) .”

3) is a direct consequence of the equality $\lim_{|Z| \rightarrow \infty} \frac{dz}{dZ} = 1$ which is always valid for a canonical conformal mapping.

But we have already established a function $F(Z)$ answering these questions; hence, the function $f(z)$ that we seek is given by $f(z) = F(h(z))$ where, of course,

$$F(Z) = V_0(Ze^{-i\alpha} + \frac{R^2 e^{i\alpha}}{Z}) - \frac{i\Gamma}{2\pi} \log Z.$$

It is shown that the thus determined function $f(z)$ is, up to an additive constant without importance, the *unique*⁹ function satisfying the conditions 1), 2) and 3). The fundamental problem of the theory of profiles is thus reduced to the problem of determination of the canonical conformal mapping of the domain (d) — the exterior of the profile — onto the outside of the circular disk.

If the fluid flow past a circular disk has some singularities (sources, point vortices, doublets, etc.) an important result which allows the determination of the corresponding complex potential is the “circle (Milne–Thompson) theorem” which states the following:

The function $f(z)$ which is analytic in D — the exterior of the circumference $|z| = R$ — except at finite number of singular points $E \subset D$, whose principal parts with respect to these singularities is $f_0(z)$ and which is continuous on $\overline{D} \setminus E$, will satisfy the requirement $\text{Im } f(z)|_{|z|=R} =$

0 only if $f(z) = f_0(z) + \overline{f_0}\left(\frac{R^2}{\overline{z}}\right) + a$, a being a real constant.

Some remarkable extensions of the circle (Milne–Thompson) theorem are given by Caius Iacob [69].

The Blasius formulae [52] allow us to evaluate directly the global efforts exerted on the profile by the fluid flow. We will limit ourselves to the determination of the general resultant of these efforts, which comes to the “complex force” \mathcal{F} given by the formula (Blasius–Chaplygin) [52]

$$\mathcal{F} = \frac{i\rho}{2} \int_{(c)} \left(\frac{df}{dz}\right)^2 dz, \quad (c) \text{ being considered in a direct sense.}$$

To calculate this integral we remark that it is possible to continuously deform the integration contour (c) into a circular circumference of an arbitrarily large radius, centered at the origin, $\frac{df}{dz}$ being analytic and uniform in the whole outside of (c), that means in (d); on the other

⁹This result is a consequence of the uniqueness of the solution of the external Dirichlet problem for a disk with supplementary condition of a given non-zero circulation. See, for instance, Paul Germain, “Mécanique des milieux continus”, pag. 325, Ed. Masson, 1962 [52].

hand, for $|z|$ large enough, using $z = Z + \sum_{n=0}^{\infty} \frac{a_n}{Z^n}$ and $F(Z) = V_0(Ze^{-i\alpha} + \frac{R^2 e^{i\alpha}}{Z}) - \frac{i\Gamma}{2\pi} \log Z$ we also have

$$\begin{aligned} \frac{df}{dz} &= \frac{dF}{dZ} \frac{dZ}{dz} = \left\{ V_0 e^{-i\alpha} - \frac{i\Gamma}{2\pi} \frac{1}{Z} - \frac{R^2 e^{i\alpha}}{Z^2} \right\} \cdot \left\{ 1 - \sum_{n=1}^{\infty} \frac{a_n}{Z^{n+1}} \right\}^{-1} \\ &= V_0 e^{-i\alpha} - \frac{i\Gamma}{2\pi} \frac{1}{Z} + \dots = V_0 e^{-i\alpha} - \frac{i\Gamma}{2\pi} \frac{1}{z} + \dots, \end{aligned}$$

the unwritten terms being infinitesimally small of second order in z^{-1} and Z^{-1} . Hence

$$\left(\frac{df}{dz} \right)^2 = V_0^2 e^{-2i\alpha} - \frac{i\Gamma V_0 e^{-i\alpha}}{\pi} \frac{1}{z} + \dots$$

such that

$$\mathcal{F} = \frac{i\rho}{2} (2i\pi) \left(-\frac{i\Gamma V_0 e^{-i\alpha}}{\pi} \right) = i\rho\Gamma V_0 e^{-i\alpha}.$$

So, we can see that the general resultant is acting on a direction which is perpendicular to the attack (far field) velocity, its algebraic magnitude being $-\rho\Gamma V_0$. This result is known as the Kutta–Joukowski theorem and, according to it the resultant component on the velocity direction — the so-called *drag* —, is zero, which represents *D’Alembert’s paradox*, while the normal component vs. the velocity direction, the so-called *lift*, would be zero if the flow is without circulation.

D’Alembert’s paradox also holds for three-dimensional potential flows. This “weakness” of the mathematical model could be explained not only by accepting the inviscid character of fluid and, implicitly, the slip-condition on rigid walls but also by assuming the potential (irrotational) character of the entire fluid flow, behind the obstacle too. However experience shows that, behind the obstacles, there are vortices separations. That is why we will consider, in the next sections, the case of the *almost (nearly) potential flows* — that is with vortices separation — and when D’Alembert’s paradox does not show up.

5.2 Profiles with Sharp Trailing Edge. Joukowski Hypothesis

Many aerodynamics profiles have “behind” an angular point, the plane trace of the sharp edge of the wing with infinite span. Let z_F be the

affix of this *sharp trailing edge* of (c) and $Z_F = R e^{i\beta}$ be the affix of its homologous from (C) (by the canonical conformal mapping considered before). The function $z = H(Z)$, in the neighborhood of $Z = Z_F$ behaves as¹⁰

$$z - z_F = A (Z - Z_F)^p + \dots,$$

the omitted terms in this expansion being of order higher than p in $Z - Z_F$. According to the above expansion if a direction, passing through Z_F , is rotated with an angle α , then the homologous direction passing through z_F , will rotate with the angle $p\alpha$. If we denote by $\delta\pi$ ($0 \leq \delta < 1$), the angle of the semitangents drawn to (c) , at z_F (that is the “jump” of a semitangent direction passing through z_F is $2\pi - \delta\pi$, see Figure 2.4 A), one could see that the exponent p in the above expansion should necessarily be $2 - \delta$, the “jump” of the homologous direction from the plane Z , thus being π (see Figure 2.4 B).

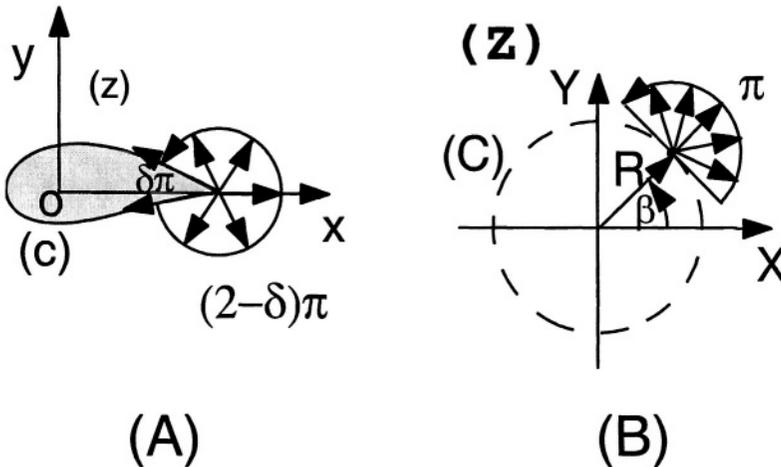


Figure 2.4. Profile with sharp trailing edge

Consequently, in the vicinity of Z_F , $\left(\frac{dz}{dZ}\right)_F = A(2-\delta)(Z - Z_F)^{1-\delta} + \dots$ and this derivative vanishes at $Z = Z_F$. But then, from $\frac{df}{dz} = \frac{dF}{dZ} \frac{dZ}{dz}$, one could see that the complex velocity in the neighborhood of the sharp

¹⁰See, for instance, C. Iacob, “ Introduction mathématique à la mécanique des fluides”, p. 645 [69].

trailing edge of the profile of the affix z_F , has, in general, an unbounded modulus. This situation does not arise when $Z = Z_F$ is a zero velocity (stagnation point) for the envisaged flow; really, $Z = Z_F$ being a simple zero for $\frac{dF}{dZ}$ and

$$\left(\frac{dZ}{dz}\right)_F = \frac{1}{A(2-\delta)(Z-Z_F)^{1-\delta}} + \dots,$$

$\frac{df}{dz}$ will be zero at $z = z_F$ if $0 < \delta < 1$ or, bounded, if $\delta = 0$ (this last case corresponds to the presence, at the trailing edge, of a *cuspidal point* of (c)).

To avoid the existence of infinite velocities in the neighborhood of the sharp trailing edge (which does not have any physical support), one states the following hypothesis, called also *the Joukovski–Kutta hypothesis (condition)*: “The circulation which, for a given incidence, should be considered for the flow around a profile with sharp trailing edge, is that which leads to a finite velocity at the trailing edge”.

To determine the effective value of this circulation it would be sufficient to write that $Z_F = R e^{i\beta}$ is a stagnation (zero velocity) point for the transformed (associated) flow around the disk (C).

From the expression of the complex velocity on the circular boundary in the fluid flow past the disk [69], that is

$$\zeta = 2ie^{-i\theta} V_0 [\sin(\theta - \alpha) - \sin \gamma],$$

we could see that this implies $\gamma = \beta - \alpha$ and hence

$$\Gamma = 4\pi V_0 R \sin \gamma = 4\pi V_0 R \sin(\beta - \alpha).$$

So that, taking into account the Joukovski hypothesis, *there is only one flow* past a profile when the incidence is “a priori” given. The angle β defines the so-called *zero lift* direction because, if $\beta = \alpha$, $\Gamma = 0$ and the lift will be also zero by the above evaluation for Γ .

5.3 Theory of Joukovski Type Profiles

Let us consider the transformation $z = \frac{1}{2} \left(Z + \frac{1}{Z} \right)$ whose derivative is $\frac{dz}{dZ} = \frac{1}{2} \left(1 - \frac{1}{Z^2} \right)$. This transformation defines a conformal mapping between the planes (z) and (Z) except the singular points $Z = \pm 1$ where the conformal character is lost.

It is shown that if $Z = r e^{i\sigma}$ ($r \neq 1$), its image in the plane (z) will be the ellipse

$$x = \frac{1}{2} \left(r + \frac{1}{r} \right) \cos \sigma, \quad y = \frac{1}{2} \left(r - \frac{1}{r} \right) \sin \sigma$$

whose focuses are located at the points $A(1,0)$ and $A'(-1,0)$. In the case when $r = 1$ the image in the plane (Z) will be the segment $[-1,1]$ run in both senses (on the “upper border” and then, in the opposite sense, on the “lower border”). Obviously, in this case, the considered transformation would map both the outside and inside of the unit disk $|Z| \leq 1$, onto the whole plane (z) with a cut along the segment $[-1,1]$ (in accord with the existence of two inverse transformations $Z = z \pm \sqrt{z^2 - 1}$, where, to fix the ideas, the positive determination of the root at $z = x > 1$ is considered).

If Γ is a circumference passing by A and A' , its image will be only a circular arc joining A and A' and crossing the center C of Γ , an arc which is run in both senses. Let’s now consider a circumference Γ_1 passing only through the singular point A (and *not* through A'). Its image will be a closed curve with a sharp cuspidal point at A where the tangent is the same with that to the arc ACA' which is also “the skeleton” of this contour.

This image contour is called *the Joukovski (wing) profile*, and the initial considered transformation is of *Joukovski or Kutta–Joukovski type*.

Obviously to a fluid flow around Γ_1 , of $\frac{V_0}{2}$ velocity at far field, it could associate a fluid flow of V_0 velocity at infinity, past the considered Joukovski profile, the incidences in both flows being the same.

The Joukovski profiles are technically hard to make and more, they are not very realistic for practical purposes. That is why their importance is mainly theoretical.

The above Joukovski type transformation could be generalized by considering

$$z = \frac{1}{2} \left(Z + \frac{R^2}{Z} \right)$$

or even $z = Z + \frac{R^2}{Z}$, the last transformation having the advantage of equal velocities at far field in the associated flows. We remark that the last form could be rewritten as

$$\frac{z - 2R}{z + 2R} = \left(\frac{Z - R}{Z + R} \right)^2,$$

and it transforms the outside of $|Z| = R$ onto the whole plane (z) with a cut along the segment $[-2R, 2R]$. A direct generalization would be

$$\frac{z - kR}{z + kR} = \left(\frac{Z - R}{Z + R} \right)^k, \quad 1 < k < 2,$$

which points out that

$$z - kR = (Z - R)^k \varphi(Z), \quad \varphi(R) \neq 0,$$

a form which avoids the sharp cuspidal point and which, in the vicinity of infinity, has the expansion

$$z = Z + \frac{k(k-1)}{2} \cdot \frac{R^2}{Z} + \dots$$

In this case the image of a circumference Γ passing through $-R$ and R will be the union of two circular arcs, symmetrical versus Ox and passing through $-kR$ and kR .

Finally, if one considers the image of a circumference Γ_1 , passing only through $Z = R$ and centered on the Ox axis, this image will be tangent to the previous symmetrical contour at kR where it has also a sharp point with the angle of semitangents equal to $2k\pi$. Such an image is known as a *Karman–Trefftz profile*. An application on a dirigible balloon of Karman–Trefftz type is given in chapter 6, 3.3.

Writing the Joukovski type transformation under the form

$$\frac{dz}{dZ} = \left(1 - \frac{R}{Z}\right) \left(1 + \frac{R}{Z}\right),$$

von Mises has considered the generalization

$$\frac{dz}{dZ} = \left(1 - \frac{R}{Z}\right)^{k-1} \left(1 - \frac{\mu_1}{Z}\right) \dots \left(1 - \frac{\mu_n}{Z}\right), \quad 1 < k < 2, \mu_j \neq R.$$

Again a circumference passing through $Z = R$ is transformed onto a (*wing profile of von Mises type*), with a sharp point at a certain z_0 and where the jump of each semitangent is $k\pi$.

We remark that if the Joukovski type profiles depend on two parameters (like the coordinates of the Γ_1 center), the Karman – Trefftz type profiles depend on three parameters (with the additional k) while the von Mises type profiles depend on $n + 1$ parameters.

E. Carafoli has introduced the transformations of the type $z = Z + \frac{R^2}{Z} + \frac{a}{(Z-b)^p}$ with p a positive integer (the order of the pole b). For small a one obtains quasi-Joukovski profiles.

Caius Jacob has considered a class of profiles defined by the conformal mappings expressed in terms of rational functions [70].

Recently, I. Taposu has emphasized a special class of profiles (“dolphin profiles”) whose use in practice could improve the classical concepts of aerodynamics [139].

In different laboratories around the world one deals with classes of profiles (Naca, Göttingen, ONERA, RAE, Tzagy, etc.) which are given, in general, “by points” and, seldom, by their analytical form.

5.4 Example

In the sequel we will illustrate a particular transformation (mapping), namely the Joukowski transformation (see section 2.5.3), $z = z' + \frac{a^2}{z'}$. By this transformation the complex potential of a uniform flow becomes $f(z') = U \left(z' + \frac{a^2}{z'} \right)$ i.e., the potential for a uniform flow past a circular cylinder of radius a , U being the magnitude of the velocity at far field.

This transformation, $z = z' + \frac{b^2}{z'}$ where $b^2 < a^2$, allows the conformal transformation of a circle of radius a centered at $P(x'_P, y'_P)$ from the second quadrant $Ox'y'$ onto a so-called Joukowski airfoil (profile) in the Oxy plane.

Let us now consider a uniform flow of velocity U in the positive Ox direction past the above Joukowski airfoil. In particular, its sharp trailing edge at $x = 2b$, is the image of the point Q at $z' = b$ where Ox' is crossed by the above circle.

The magnitude V of the velocity in the Oxy plane is related to the magnitude V' of the velocity in the $Ox'y'$ plane by the relation

$$\left| \frac{df}{dz} \right| = \frac{\left| \frac{df}{dz'} \right|}{\left| \frac{dz}{dz'} \right|},$$

i.e.,

$$V = \frac{V'}{\left| 1 - \left(\frac{b}{z'} \right)^2 \right|}. \quad (2.8)$$

We remark that if the velocity $V' \neq 0$ at Q where $z' = b$, then the velocity V at the sharp trailing edge $z = 2b$ becomes infinite, which is a contradiction with the *Joukowski-Kutta condition*. Thus, we must impose that the point Q on the circle be a stagnation point; this goal may be reached if we create a clockwise circulation Γ on the circle, and this circulation is then conserved by the conformal mapping. The magnitude of this circulation is $\Gamma = 4\pi a U \sin \theta = 4\pi y'_P U$ and the flow past the circle is then constructed by adding to the uniform stream a doublet and a point vortex, so that we get the complex potential of the resultant flow

$$f = U \left[z' - z'_P + \frac{a^2}{z' - z'_P} + i2y'_P \log \left(\frac{z' - z'_P}{a} \right) \right].$$

Here the constant term $-i2y'_P \log a$ has been added but the values of the stream function Ψ on the circle do not change after this superposition.

The variables a, b, x'_P, y'_P are related by the relationship $a^2 = y'^2_P + (b - x'_P)^2$ and they control the shape of the airfoil. For instance, a and b determine the thickness and the chord length while the ordinate of P the “camber” of the airfoil.

For our example we will take $U = 1\text{m/s}$, $a = 1\text{m}$, $b = 0.8\text{m}$, $y'_P = 1.199\text{m}$. Using the formula for the uniform motion with circulation past a circle in the $Ox'y'$ plane, we generate the airfoil profile as a level curve ($\Psi = 0$) in the Oxy plane. Other level curves $\Psi = \text{Const}$ give other streamlines around the airfoil, see Figure 2.5.

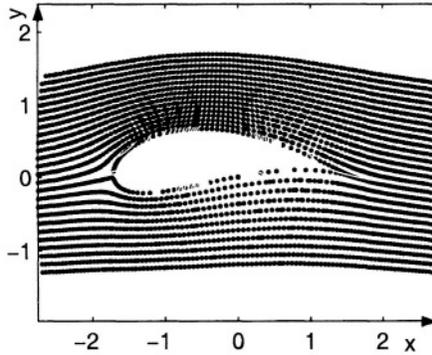


Figure 2.5. Uniform flow past a Joukovski airfoil

The pressure on the surface could be calculated using the velocities, from the formula (2.8)

$$V = U \left| \frac{1 - \left(\frac{a}{z' - z'_P} \right)^2 + i \frac{2y'_P}{z' - z'_P}}{1 - \left(\frac{b}{z'} \right)^2} \right|,$$

and then the dimensionless pressure difference (the pressure coefficient) at every point can be calculated according to Bernoulli’s relation by

$$c_p = \frac{p - P}{\frac{1}{2}\rho U^2} = 1 - \left(\frac{V}{U} \right)^2. \tag{2.9}$$

It is shown in Figure 2.6.

The MATLAB program is

```
a=1;b=0.8;U=1;yp1=0.189;
xp1=b-sqrt(a^2-yp1^2);zp1=xp1+i*yp1;
x=-2.5:0.05:2.5;y=-2.5:0.05:2.5;
```

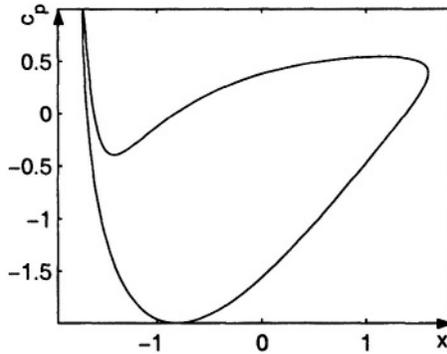


Figure 2.6. The pressure distribution around the airfoil

```
[X,Y]=meshgrid(x,y);Z=X+i*Y;Z2=Z-zp1;
PSI=U*imag(Z2+a^2./Z2+i*2*yp1*log(Z2/a));
c=contour(X,Y,PSI,[0 0]);axis('equal');
z=c(1,:)+i*c(2,:);
for j=1:length(c) if abs(z(j)-zp1)<a z(j)=0;end;end;
f=z+b^2./z;
for j=1:length(c) if abs(f(j))>3 f(j)=0;end;end;
plot(f,'r.');
```

```
axis('equal');hold on;
c=contour(X,Y,PSI,[-1:0.1:-0.1 0.1:0.1:1.5],'f');
axis('equal');
z=c(1,:)+i*c(2,:);
for j=1:length(c) if abs(z(j)-zp1)<a z(j)=0;end;end;
f=z+b^2./z;
for j=1:length(c) if abs(f(j))>3 f(j)=0;end;end;
plot(f,'k.');
```

```
axis('equal');hold off;pause;
fi=linspace(0,2*pi,200); z2=a*exp(i*fi);
z1=z2+zp1; z=z1+b^2./z1;
V=U*abs((1-(a./z2).^2+i*2*yp1./z2)./(1-(b./z1).^2));
plot(real(z),1-(V/U).^2);axis('equal');
```

5.5 An Iterative Method for Numerical Generation of Conformal Mapping

In the sequel, we will present a method for the approximate construction of conformal mappings for arbitrary shaped obstacles [87].

It is known that a function $z = H(Z)$, which maps conformally the outside of a profile (c) from the plane (z) onto the outside of a disk (C), of radius R , from the plane (Z), can be represented as a series

$$z = Z + p_0 + iq_0 + \sum_{n=1}^{\infty} (p_n + iq_n) \left(\frac{R}{Z} \right)^n .$$

The main problem is the effective calculation of the coefficients $p_0, q_0, \dots, p_n, q_n$. To do that, we will consider the previous development at the point $Z = R e^{i\theta}$, $0 \leq \theta \leq 2\pi$, of the circumference (C) and then we will separate the real and the imaginary parts, thus obtaining

$$x(\theta) = p_0 + (R + p_1) \cos \theta + q_1 \sin \theta + \sum_{n=2}^{\infty} (p_n \cos n\theta + q_n \sin n\theta) ,$$

$$y(\theta) = q_0 + (R - p_1) \sin \theta + q_1 \cos \theta + \sum_{n=2}^{\infty} (q_n \cos n\theta - p_n \sin n\theta) .$$

Although the coordinates (x, y) of the points of the contour (c) are known, either in a tabular or in a functional form, the functions $x(\theta)$ and $y(\theta)$ are still unknown. That is why an iterative method to calculate $x(\theta)$ and $y(\theta)$ must use the coefficients $p_0, q_0, \dots, p_n, q_n$.

First, due to the orthogonality conditions for the trigonometric functions, we have

$$p_0 = \frac{1}{2\pi} \int_0^{2\pi} x(\theta) d\theta ,$$

$$R + p_1 = \frac{1}{\pi} \int_0^{2\pi} x(\theta) \cos \theta d\theta , \quad R - p_1 = \frac{1}{\pi} \int_0^{2\pi} y(\theta) \sin \theta d\theta ,$$

$$p_n = -\frac{1}{\pi} \int_0^{2\pi} y(\theta) \sin n\theta d\theta , \quad n > 1 , \quad q_n = \frac{1}{\pi} \int_0^{2\pi} y(\theta) \cos n\theta d\theta , \quad n \geq 0$$

and, from here, we could write that

$$R = \frac{1}{2\pi} \int_0^{2\pi} [x(\theta) \cos \theta + y(\theta) \sin \theta] d\theta ,$$

$$p_1 = \frac{1}{2\pi} \int_0^{2\pi} [x(\theta) \cos \theta - y(\theta) \sin \theta] d\theta .$$

Then we choose for $x(\theta)$ its “initial” (of order zero) approximation $x^0(\theta) = \alpha + \beta \cos \theta$ where α and β are arbitrary. From the expression of p_0 and $p_1 + R$, we have $p_0^{(0)} = \alpha$, $p_1^{(0)} + R^{(0)} = \beta$.

To the above abscissa $x^{(0)}$ it is possible to join the corresponding ordinate $y^{(0)}$, either from the given tabular or from the functional form, and then we can also obtain the coefficients $R^{(1)} - p_1^{(1)}, p_n^{(1)}, q_n^{(1)}$ which will be calculated via the mentioned integral relations. Using these coefficients new abscissas and then new ordinates are calculated and so the process is continued. For instance, within the iteration of m order (m -th iteration) we have

$$x^{(m-1)}(\theta) = \alpha + \beta \cos \theta + q_1^{(m-1)} \sin \theta + \sum_{n=2}^{\infty} \left(p_n^{(m-1)} \cos n\theta + q_n^{(m-1)} \sin n\theta \right),$$

$$p_1^{(m)} + R^{(m)} = \beta, \quad R^{(m)} - p_1^{(m)} = \frac{1}{\pi} \int_0^{2\pi} y^{(m)}(\theta) \sin \theta d\theta,$$

from where

$$R^{(m)} = \frac{\beta}{2} + \frac{1}{2\pi} \int_0^{2\pi} y^{(m)}(\theta) \sin \theta d\theta.$$

The iterative method sketched above is easy to use on a computer. The only additional required subprograms are connected to the interpolation such that in each “sweep” new values of the ordinates, respectively abscissas, become available. The method converges quite fast.

6. Panel Methods for Incompressible Flow of Inviscid Fluid

The panel methods in both source and vortex variants, are numerical methods to approach the incompressible inviscid fluid flow, and which, since the late 1960s, have become standard tools in the aerospace industry. Even if in the literature the panel methods occur within “computational aeronautics”, we will consider them as a method of CFD.

In this section we will “sketch” the panel method, separately in the source variant and then in the vortex variant, by considering only the “first order” approximation.

6.1 The Source Panel Method for Non-Lifting Flows Over Arbitrary Two-Dimensional Bodies

Let us consider a given body (profile) of arbitrary shape in an incompressible inviscid fluid flow with free-stream velocity \mathbf{V}_∞ . Let a contin-

uous distribution of sources be along the contour (surface) of the body and let $\lambda(s)$ be the source strength, per unit length, of this distribution where s is the natural parameter (the distance measured along this contour in the edge view). Obviously an infinitesimal portion ds of the boundary (source sheet) can be treated as a distinct source of strength λds . The effect induced by such a source at a point $P(x, y)$, located a distance r from ds , is a fluid flow with an infinitesimally small velocity potential $d\phi$ given by

$$d\phi = \frac{\lambda ds}{2\pi} \ln r.$$

The total velocity potential at the point P , induced by all the sources from a to b , is obtained by summing up the above infinitesimal potentials, which means

$$\Phi(x, y) = \int_a^b \frac{\lambda ds}{2\pi} \ln r.$$

Obviously, the fluid velocity induced by the source distribution (sheet) will be superposed, at any point P , on the free-stream (attack) velocity. The problem we intend to solve (numerically) is that of the determination of such a source distribution $\lambda(s)$ which “observes” the surface (boundary) of the body (profile), i.e., the combined action of the uniform flow and the source sheet makes the profile boundary a streamline of the flow.

To reach this target, let us approximate the profile boundary by a set of straight panels (segments), the source strength λ per unit length being constant over a panel but possibly varying from one to another panel.

Thus, if there is a total of n panels and $\lambda_1, \lambda_2, \dots, \lambda_j, \dots, \lambda_n$ are the constant source strengths over each panel respectively, these “a priori” unknown λ_j will be determined by imposing the slip-condition on the profile boundary. This boundary condition is imposed numerically by defining the midpoint of each panel to be the *control point* where the normal component of the fluid velocity should be zero.

In what follows, for sake of simplicity, we will choose the control points to be the midpoints of each panel (segment).

Let us denote by r_{pj} the distance from any point (x_j, y_j) on the j -th panel to the arbitrary point $P(x, y)$. The velocity potential induced at P due to the j -th panel of constant source strength λ_j is

$$\Phi_j = \frac{\lambda_j}{2\pi} \int_j \ln r_{pj} ds_j.$$

Obviously, the potential at P due to all the panels is the sum

$$\Phi(P) = \sum_{j=1}^n \Phi_j = \sum_{j=1}^n \frac{\lambda_j}{2\pi} \int_j \ln r_{pj} ds_j.$$

Suppose now that P is the control point, that is the midpoint of the i -th panel. Then we have

$$\Phi(x_i, y_i) = \sum_{j=1}^n \frac{\lambda_j}{2\pi} \int_j \ln r_{ij} ds_j,$$

while the normal component of the velocity at (x_i, y_i) is

$$v_n = \frac{d}{dn_i} \Phi(x_i, y_i),$$

$\mathbf{n}(n_i)$ being the outward unit normal vector to the i -th panel. Because for $j = i$, $r_{ij} = 0$ at the control point and, when the derivative is carried out, r_{ij} appears in the denominator (thus creating a singular point), it would be useful to evaluate directly the contribution of the i -th panel to this derivative calculated at (x_i, y_i) . Since it is about a source which acts only on a half-circumference (the other half-circumference does not interfere due to the rigid wall), its strength will be $\frac{\lambda_i}{2}$ and this is the looked for contribution to the normal component of the velocity. Hence

$$v_n = \frac{\lambda_i}{2} + \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\lambda_j}{2\pi} \int_j \frac{d}{dn_i} (\ln r_{ij}) ds_j.$$

Taking into account that the normal component of the free-stream velocity \mathbf{V}_∞ at the same point (x_j, y_j) is $v_{\infty, n} = \mathbf{V}_\infty \cdot \mathbf{n}_i = v_\infty \cos \beta_i$, β_i being the angle between \mathbf{V}_∞ and \mathbf{n}_i , the slip-condition will be $v_{\infty, n} + v_n = 0$, which means

$$\frac{\lambda_i}{2} + \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\lambda_j}{2\pi} \int_j \frac{d}{dn_i} (\ln r_{ij}) ds_j + v_\infty \cos \beta_i = 0.$$

Applying this approach to all the panels, the above equalities with $i = 1, 2, \dots, n$, represent a linear algebraic system with n unknowns $\lambda_1, \lambda_2, \dots, \lambda_n$, which can be solved by conventional numerical methods.

Certainly this approximation could be made more accurate by increasing the number of panels and, if necessary, by considering panels of different length (for instance, in the case of a profile shape, one gets a good accuracy by considering 50 to 100 panels which are either smaller in the leading edge region of a rapid surface curvature or longer over the quasi-flat portions of the profile).

Obviously, following the same way, we can also obtain the tangential components of the velocity at the same point (x_j, y_j) , precisely

$$v_i = v_{\infty,t} + v_t = v_{\infty} \sin \beta_i + \sum_{j=1}^n \frac{\lambda_j}{2\pi} \int_j \frac{d}{ds} (\ln r_{ij}) ds_j.$$

Hence, the pressure at the same control point is calculated by the Bernoulli theorem while the pressure coefficients are $C_{p,i} = 1 - \left(\frac{v_i}{v_{\infty}}\right)^2$.

Before ending this section it is important to give a procedure for testing the accuracy of the above method. If S_j is the length of the j -th panel of source strength λ_j (per unit length), then the strength of the entire panel will be, obviously, $S_j \lambda_j$. But the mass conservation, in the hypothesis of a closed contour, allows us to write $\sum_{j=1}^n S_j \lambda_j = 0$ which provides an independent criterion to test the obtained results.

6.2 **The Vortex Panel Method for Lifting Flows Over Arbitrary Two-Dimensional Bodies**

Consider now a continuous distribution of vortices (vortex sheet) over the surface (contour) of a body (profile) in an incompressible flow with free-stream velocity \mathbf{V}_{∞} . Let $\gamma = \gamma(s)$ be the strength (circulation) of the vortex sheet, per unit length along s . Thus the strength of an infinitesimal portion ds of the boundary (vortex sheet) is γds and this small section could be treated as a distinct vortex of strength γds . Introducing again the point $P(x, y)$ in the flow, located at distance r from ds , the infinitesimal portion ds of the boundary (vortex sheet) of strength γds induces an infinitesimal velocity potential at P , namely

$$d\Phi = -\frac{\gamma ds}{2\pi} \theta$$

and, correspondingly, the entire distribution of vortices from $s = a$ and $s = b$ will generate a velocity potential

$$\Phi = -\frac{1}{2\pi} \int_a^b \theta \gamma ds.$$

Analogously, the circulation around the vortex sheet from $s = a$ to $s = b$ is the sum of the strength of the elemental vortices, that is $\Gamma = \int_a^b \gamma ds$. Another property of this vortices distribution is that the tangential component of the fluid velocity experiences a discontinuity across the sheet in the sense that, for every s , $\gamma = u_1 - u_2$, u_1 and u_2 being the tangential velocities “above” and “below” the sheet respectively.

This last relation is used to demonstrate that, for flow past a wing profile, the value of γ is zero at the trailing edge, which means $\gamma_F = 0$. In fact this relation is one form of the Joukowski condition which fixes the values of the circulation around the profile with a sharp trailing edge, the lift force L being related to this circulation through the Kutta–Joukowski theorem, that is $L = \rho_\infty v_\infty \Gamma$. The goal of this method is to find $\gamma(s)$ such that the body (profile) surface (boundary) becomes a streamline of the flow. At the same time we wish to calculate the amount of circulation and, implicitly, the lift on the body.

As in the case of sources, we will approximate the vortex sheet by a series of n panels (segments) of constant strength (per unit length) which form a polygonal contour “inscribed” in the profile contour. Let us denote by $\gamma_1, \gamma_2, \dots, \gamma_j, \dots, \gamma_n$ the constant vortex strength over each panel respectively. Our aim is to determine these unknown strengths such that both the slip-condition along the profile boundary and the Joukowski condition are satisfied. Again the midpoints of the panels are the control points at which the normal component of the (total) fluid velocity is zero.

Let $P(x, y)$ be a point located a distance r_{pj} from any point of the j -th panel, the radius r_{pj} making an angle θ_{pj} to the Ox axis. The velocity potential induced at P due to *all* the panels is

$$\Phi(P) = \sum_{j=1}^n \Phi_j = - \sum_{j=1}^n \frac{\gamma_j}{2\pi} \int_j \theta_{pj} ds_j,$$

where $\theta_{pj} = \arctg \frac{y-y_j}{x-x_j}$.

If P is the control point of the i -th panel, then

$$\Phi(x_i, y_i) = - \sum_{j=1}^n \frac{\gamma_j}{2\pi} \int_j \theta_{ij} ds_j, \quad \theta_{ij} = \arctg \frac{y_i - y_j}{x_i - x_j}.$$

Hence the normal component of the total fluid flow at the point (x_j, y_j) is

$$v_\infty \cos \beta_i - \sum_{j=1}^n \frac{\gamma_j}{2\pi} \int_j \frac{\partial \theta_{ij}}{\partial n_i} ds_j$$

which, vanishing for every i (the slip-condition), will generate a linear algebraic system to determine the unknowns $\gamma_1, \gamma_2, \dots, \gamma_j, \dots, \gamma_n$. But this time, in contrast with the source panel method, the system should be completed with the Joukovski condition $\gamma_F = 0$. In fact, the fulfilment of this last condition could be performed by considering two small panels (panels i and $i-1$), in the neighborhood of the sharp trailing edge, such that the control points i and $i-1$ are close enough to the trailing edge, and imposing that $\gamma_i = -\gamma_{i-1}$. This leads to the “a priori” fulfilment of the Joukovski condition. At the same time, to avoid the approach of an over-determined system of n unknowns with $n+1$ equations we will ignore the slip-condition at *one* of the control points and so we get again a system of n linear algebraic equations with n unknowns, which can be solved by conventional techniques.

Obviously, the obtained solution, besides the slip-condition, will satisfy the Joukovski condition too. More, the tangential velocities to the boundary are equal to γ which could be seen clearly supposing that, at every point *inside* the body (on the “lower” part of the vortex sheet too) the velocity $u_2 = 0$. Hence, the velocity outside the vortex sheet is $\gamma = u_1 - u_2 = u_1 - 0 = u_1$ so that the local velocities tangential to the surface (boundary) are equal to the local values of γ .

Concerning the circulation, if S_j is the length of the j -th panel, then the circulation due to the j -th panel is $\gamma_j S_j$ and the total circulation is

$$\Gamma = \sum_{j=1}^n \gamma_j S_j \text{ and, correspondingly, the lift } L \text{ is } \rho_\infty V_\infty \sum_{j=1}^n \gamma_j S_j.$$

Finally, we remark that the accuracy problems have encouraged the development of some higher-order panel techniques. Thus a “second-order” panel method assumes a linear variation of γ over a given panel such that, once the values of γ are matched at the edges to its neighbors, the values of γ at the boundary points become the unknowns to be solved. Yet the slip-condition, in terms of the normal velocity at the control points, is still applied.

There is also a trend to develop panel techniques using a combination of source panels and vortex panels (source panels to accurately represent “the thickness” of the profile while vortex panels to effectively provide the circulation). At the same time, there are many discussions on the control point to be ignored for “closing” the algebraic system in the case

of the vortex panels. References can be found, for instance, in the book of Chow [22].

6.3 Example

Let us consider, for instance, a source panel of length $2L$, lying symmetrically on the Oy axis [22]. Assume that on it, sources of the strength λ per unit length are distributed. The velocity potential induced at every point (x, y) by the source contained in the infinitesimal panel element dy' at $(0, y')$ is $\frac{\lambda dy'}{2\pi} \ln [x^2 + (y - y')^2]^{\frac{1}{2}}$ (this expression is obtained by taking the real part of the source complex potential).

The potential induced by the entire panel is

$$\Phi(x, y) = \frac{\lambda}{4\pi} \int_{-L}^L \ln [x^2 + (y - y')^2] dy'$$

and the velocity components can be obtained by derivation with respect to x , respectively y ,

$$u(x, y) = \frac{\lambda}{2\pi} \left[\operatorname{arctg} \left(\frac{y+L}{x} \right) - \operatorname{arctg} \left(\frac{y-L}{x} \right) \right]$$

$$v(x, y) = \frac{\lambda}{4\pi} \ln \frac{x^2 + (y + L)^2}{x^2 + (y - L)^2}.$$

Considering a point (x, y) such that $x > 0$ and $y \in (-L, L)$, if $x \rightarrow 0$ from the right of the panel we obtain the limit $u(+0, y) = \frac{\lambda}{2}$. On the other hand, by a similar approach from the left, we obtain the limit $u(-0, y) = -\frac{\lambda}{2}$. Thus the panel generates a flow having an outward normal velocity of magnitude $\frac{\lambda}{2}$. The tangential velocity v is the same on both sides of the panel and it is zero at the midpoint and infinite at the edges of the panel.

If such a panel with sources of strength $\lambda = 2U$ is placed normal to a uniform flow of speed U , the induced normal velocity cancels the oncoming flow on the left side and thus the resultant flow is tangent to the surface. So, the panel becomes coincident with one of the streamlines of the flow.

If the panel makes an angle θ with the uniform stream, the generated flow cancels the normal induced flow if its strength is $\lambda = 2U \sin \theta$.

Let now m be the number of the panels. On each panel are distributed uniform sources of strength $\lambda_1, \dots, \lambda_m$ (strength per unit length) respectively. The velocity potential of the resultant flow at every point (x_i, y_i) from the flow field, generated by the sources from the j -th panel is, as above, $\frac{\lambda_j}{2\pi} \int_J \ln r_{ij} ds_j$ where J is the panel and $\lambda_j ds_j$ is the strength of

the source from the element ds_j located at (x_j, y_j) on that panel. Here $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ is the distance from the control point (x_i, y_i) to an arbitrary point (x_j, y_j) on the j -th panel.

The velocity potential for the flow obtained by superposition of the given uniform flow and the m source panels is then

$$\Phi(x_i, y_i) = Ux_i + \sum_{j=1}^m \frac{\lambda_j}{2\pi} \int_J \ln r_{ij} ds_j.$$

Let now (x_i, y_i) be the control point on the i -th panel, where the outward normal n_i makes an angle β_i with the uniform stream. At this point on the surface of the body, the above slip condition becomes

$$\frac{\lambda_i}{2} + \sum_{j \neq i}^m \frac{\lambda_j}{2\pi} I_{ij} = -U \cos \beta_i, \quad i = 1, \dots, m \tag{2.10}$$

where

$$I_{ij} = \int_J \frac{d}{dn_i} \ln r_{ij} ds_j .$$

The calculations become easier if we express the integrals I_{ij} in terms of the geometrical elements of the panels, see Figure 2.7.

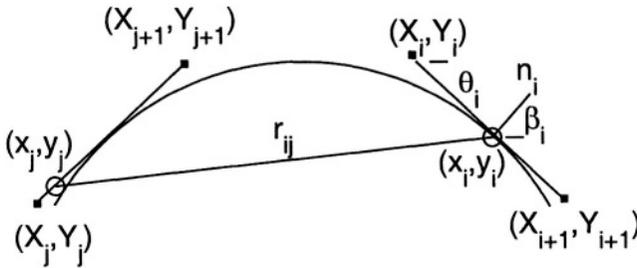


Figure 2.7. Evaluation of the integrals I_{ij}

The length of each panel is

$$S_j = \sqrt{(X_{j+1} - X_j)^2 + (Y_{j+1} - Y_j)^2}.$$

The angle θ_i at (X_i, Y_i) between the panel and the Ox axis is related with the similar angle of the normal n_i at the control point (x_i, y_i) by

the relation

$$\beta_i = \theta_i + \frac{\pi}{2}$$

from where

$$\sin \beta_i = \cos \theta_i, \cos \beta_i = -\sin \theta_i .$$

After derivation with respect to the normal we get

$$I_{ij} = \int_0^{S_j} \frac{(x_i - x_j) \cos \beta_i + (y_i - y_j) \sin \beta_i}{(x_i - x_j)^2 + (y_i - y_j)^2} ds_j$$

where

$$x_j = X_j + s_j \cos \theta_j, y_j = Y_j + s_j \sin \theta_j .$$

By replacement, the integral becomes

$$I_{ij} = \int_0^{S_j} \frac{Cs_j + D}{s_j^2 + 2As_j + B} ds_j$$

where

$$\begin{aligned} A &= -(x_i - X_j) \cos \theta_j - (y_i - Y_j) \sin \theta_j , \\ B &= (x_i - X_j)^2 + (y_i - Y_j)^2 , \\ C &= \sin(\theta_i - \theta_j), \\ D &= -(x_i - X_j) \sin \theta_i + (y_i - Y_j) \cos \theta_i . \end{aligned}$$

But the denominator of the integrand is of the form

$$(s_j + A)^2 + B - A^2 = (s_j + A)^2 + E^2 > 0$$

where

$$E = (x_i - X_j) \sin \theta_j - (y_i - Y_j) \cos \theta_j$$

thus, consequently,

$$\begin{aligned} I_{ij} &= \frac{1}{2} \sin(\theta_i - \theta_j) \ln \left[1 + \frac{S_j^2 + 2AS_j}{B} \right] \\ &\quad - \cos(\theta_i - \theta_j) \left[\arctg \left(\frac{S_j + A}{E} \right) - \arctg \left(\frac{A}{E} \right) \right] . \end{aligned} \tag{2.11}$$

By using the system (2.10), with the introduction of the dimensionless (undimensional) variables $\lambda'_j = \frac{\lambda_j}{2\pi U}$, we get

$$\sum_{j=1}^m I_{ij} \lambda'_j = \sin \theta_i, i = 1, \dots, m$$

where I_{ij} are given by (2.11), excepting $I_{ii} = \pi$ for every i .

We remark that for a body of a complicated shape the calculation of the normals to the panels at control points is not always easily performed. We can modify the above algorithm by choosing the boundary points (X_i, Y_i) to be on the surface of the body and the control points (x_i, y_i) to be the midpoints of the panels. The panel orientation is given by

$$\theta_i = \text{Arctg} \left(\frac{Y_{i+1} - Y_i}{X_{i+1} - X_i} \right), i = 1, \dots, m$$

where *Arctg* takes its values on $[-\pi, \pi]$. This technique is easier to apply but it is not as accurate as the previous method. Now the control points are located near the surface of the body and they will approach the surface if the number of panels increases.

Other remark is that the panels could be of different sizes. It is useful to take small panels in a part of the body of large curvature, in order to increase the accuracy of the method.

After the calculation of the dimensionless strengths λ'_j , the velocity potential $\Phi(x_i, y_i)$ may be written. The velocities at the control points are tangent to the panels and thus at these points

$$V(x_i, y_i) = \frac{d}{dt_i} \Phi(x_i, y_i)$$

where t_i is a tangent vector to the surface of the i -th panel.

Taking the derivative of Φ with respect to n_i we also obtain

$$\frac{V(x_i, y_i)}{U} = \cos \theta_i + \sum_{j=1}^m I'_{ij} \lambda'_j .$$

Here I'_{ij} is given by

$$I'_{ij} = -\frac{1}{2} \cos(\theta_i - \theta_j) \ln \left[1 + \frac{S_j^2 + 2AS_j}{B} \right] - \sin(\theta_i - \theta_j) \left[\text{arctg} \left(\frac{S_j + A}{E} \right) - \text{arctg} \left(\frac{A}{E} \right) \right]$$

for $i \neq j$ and $I'_{ii} = 0$ for every i .

Finally, the pressure on the surface of the body could be described by the pressure coefficient (2.9)

$$c_p = \frac{p - P}{\frac{1}{2}\rho U^2} = 1 - \left(\frac{V}{U} \right)^2 .$$

We will illustrate this method with the following problem. Let us consider two circular cylinders of radius 1 *m*, placed in a uniform flow of

velocity 1m/s . The centers of the cylinders are separated by a distance of $d = 2.5\text{m}$, in a direction perpendicular on the flow. Considering n panels on each cylinder, let us calculate for every $2n$ control points the values of the velocity and the pressure coefficient.

We choose the simplified variant, with the boundary points on the surface of the cylinders and the control points are the midpoints of the panels. The variables P of the program will contain all the characteristics of every panel.

The results are presented in Figure 2.8.

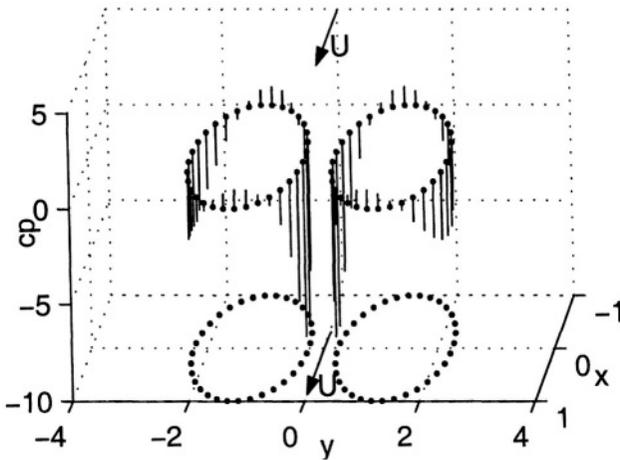


Figure 2.8. The pressure coefficient on the surface of the cylinders

The MATLAB program is

```
n=32;r=1;d=2.5;U=1;
P=zeros(2*n,8);I=zeros(2*n);Ip=zeros(2*n);
for i=1:n ui=pi-(i-1)*2*pi/n;
P(i,1)=r*cos(ui);P(n+i,1)=P(i,1);
P(i,2)=r*sin(ui)+d/2;P(n+i,2)=P(i,2)-d;
P(i,3)=r*cos(ui-2*pi/n);P(n+i,3)=P(i,3);
P(i,4)=r*sin(ui-2*pi/n)+d/2;P(n+i,4)=P(i,4)-d;
end;
for i=1:2*n
P(i,5)=(P(i,1)+P(i,3))/2;
P(i,6)=(P(i,2)+P(i,4))/2;
P(i,7)=atan2(P(i,4)-P(i,2),P(i,3)-P(i,1));
P(i,8)=sqrt((P(i,3)-P(i,1))^2+(P(i,4)-P(i,2))^2);
end;
```

```

for i=1:2*n for j=1:2*n
if j ~=i
A=-(P(i,5)-P(j,1))*cos(P(j,7))-...
(P(i,6)-P(j,2))*sin(P(j,7));
B=(P(i,5)-P(j,1))^2+(P(i,6)-P(j,2))^2;
E=(P(i,5)-P(j,1))*sin(P(j,7))-...
(P(i,6)-P(j,2))*cos(P(j,7));
I(i,j)=1/2*sin(P(i,7)-P(j,7))*...
log(1+(P(j,8)^2+2*A*P(j,8))/B)-...
cos(P(i,7)-P(j,7))*(atan((P(j,8)+A)/E)-atan(A/E));
Ip(i,j)=-1/2*cos(P(i,7)-P(j,7))*...
log(1+(P(j,8)^2+2*A*P(j,8))/B)-...
sin(P(i,7)-P(j,7))*(atan((P(j,8)+A)/E)-atan(A/E));
else I(i,i)=pi;Ip(i,i)=0;
end;
end; end;
Lp=I\ sin(P(:,7));
VPU=cos(P(:,7))+Ip*Lp;V=VPU*U;
cp=1-VPU.^2;
for i=1:2*n disp([i cp(i) V(i)]);end;
for i=1:2*n
plot3([P(i,5),P(i,5)+eps],[P(i,6),P(i,6)+eps],[0,cp(i)]);
set(gca,'view',[95,20]);
xlabel('x');ylabel('y');zlabel('cp');hold on;
end;
plot3([P(:,5);P(1,5)],[P(:,6);P(1,6)],zeros(2*n+1,1),'.');
plot3([P(:,5);P(1,5)],[P(:,6);P(1,6)],...
-10*ones(2*n+1,1),'.');
grid;
hold off;

```

We remark the low-pressure region between the two cylinders.

7. Almost Potential Fluid Flow

By *almost (slightly) potential flows*, we understand the flows in which the vorticity is concentrated in some thin layers of fluid, being zero outside these thin layers, and there is a mechanism for producing vorticities near boundaries.

For such models the Kutta–Joukowski theorem does not apply and the drag may be different from zero, which means one can avoid the D’Alembert paradox.

There are many situations in nature or in engineering where the viscous flows can be considered, in an acceptable approximation, as

“nearly potential”. Such situations occur in particular when it considers “streamlined” bodies, that is bodies so shaped as to reduce their drag.

Now we shall analyze the model of incompressible inviscid fluid flow due to the presence of N (point) vortices, located at the points $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ in the plane and of strength $\Gamma_1, \Gamma_2, \dots, \Gamma_N$, respectively. The stream function joined to the j -th vortex, ignoring the other vortices for a moment, is given by

$$\Psi_j(\mathbf{r}) = -\frac{\Gamma_j}{2\pi} \ln |\mathbf{r} - \mathbf{r}_j|.$$

The vorticity associated to the same vortex will be given by

$$\omega_j = -\Delta\psi_j = \Gamma_j \delta(\mathbf{r} - \mathbf{r}_j),$$

where δ is the Dirac function while the corresponding velocity field (ignoring again the influence of the other vortices) is

$$\mathbf{v}_j = (\partial_y\psi_j, -\partial_x\psi_j) = \left(-\frac{\Gamma_j}{2\pi} \frac{y - y_j}{r^2}, \frac{\Gamma_j}{2\pi} \frac{x - x_j}{r^2} \right)$$

with $r = |\mathbf{r} - \mathbf{r}_j|$.

Obviously, due to the interaction of vortices, the points where the vortices are centered (located) start to move. More precisely, taking into account the superposed interaction of all the vortices, $\mathbf{r}_j(x_j, y_j)$ move according to the differential equations

$$\frac{dx_j}{dt} = -\frac{1}{2\pi} \sum_{i \neq j} \frac{\Gamma_i (y_j - y_i)}{r_{ij}^2}, \quad \frac{dy_j}{dt} = \frac{1}{2\pi} \sum_{i \neq j} \frac{\Gamma_i (x_j - x_i)}{r_{ij}^2},$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$.

Then, if we retake the previous way in a reverse sense, we conclude that:

Let a system of constants $\Gamma_1, \dots, \Gamma_N$ and a system of points (initial positions) $\mathbf{r}_1(x_1, y_1), \dots, \mathbf{r}_N(x_N, y_N)$ be in the plane. Suppose we allow these points to move according to the above equations whose solutions could be written in the form $x_j = x_j(t)$ and $y_j = y_j(t)$. Define then $v_j =$

$$\left(-\frac{\Gamma_j}{2\pi} \frac{y - y_j}{r^2}, \frac{\Gamma_j}{2\pi} \frac{x - x_j}{r^2} \right) \text{ and let } \mathbf{v}(\mathbf{r}, t) = \sum_{j=1}^N \mathbf{v}_j(\mathbf{r}, t). \text{ This last equality}$$

provides a solution of Euler’s equations, a solution which preserves the circulation. Really, if C is a contour encircling k vortices $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_k$

then $\Gamma_C = \sum_{i=1}^k \Gamma_i$ and Γ_C is flow invariant (constant).

Of course the relationship between these solutions and the other solutions of the Euler system is not very obvious but it could be established rigorously under some carefully chosen hypotheses.

Now we remark that the above system forms also a Hamilton system. Really, by defining $H = -\frac{1}{4\pi} \sum_{i \neq j} \Gamma_i \Gamma_j \ln |\mathbf{r}_i - \mathbf{r}_j|$, the system is equivalent with

$$\Gamma_j \frac{dx_j}{dt} = \frac{\partial H}{\partial y_j}, \Gamma_j \frac{dy_j}{dt} = -\frac{\partial H}{\partial x_j}, j = \overline{1, N} .$$

Introduce the new variables

$$x'_i = \sqrt{|\Gamma_i|} x_i, y'_i = \sqrt{|\Gamma_i|} \operatorname{sgn}(\Gamma_i) y_i, i = \overline{1, N} ;$$

we get a real Hamilton system

$$\frac{dx'_i}{dt} = \frac{\partial H}{\partial y'_i}, \frac{dy'_i}{dt} = -\frac{\partial H}{\partial x'_i}, i = \overline{1, N}$$

and, as in classical mechanics we have

$$\frac{dH}{dt} = \sum_{i=1}^N \frac{\partial H}{\partial x'_i} \frac{dx'_i}{dt} + \sum_{i=1}^N \frac{\partial H}{\partial y'_i} \frac{dy'_i}{dt} = 0,$$

i.e., H is a constant in time along a path line.

A consequence of this property is that if all the vortices have the same sign for their strength, then they cannot collide during the motion. In other terms, if $|\mathbf{r}_i - \mathbf{r}_j| \neq 0, i \neq j$, at $t = 0$, then this result remains valid for all time since if $|\mathbf{r}_i - \mathbf{r}_j| \rightarrow 0, H$ will become infinite.

We remark that the Euler equations themselves form a Hamiltonian system (see, for instance, [2]) such that the Hamiltonian nature of the vortex model (approximation) should not surprise. What might be of great interest is to establish whether or not this system is completely integrable in the sense of Hamiltonian systems. There are some reasons to suppose the existence of a certain Lie group that generates the equations (in some sense) [19].

Let us generalize the previous case and imagine the N vortices moving in a domain D with boundary ∂D . Following the same way as before we must modify the flow of the j -th vortex (its velocity \mathbf{v}_j) so that $\mathbf{v} \cdot \mathbf{n}|_{\partial D} = 0$. This could be done by adding a potential flow of velocity \mathbf{u}_j such that $\mathbf{v}_j \cdot \mathbf{n} = -\mathbf{u}_j \cdot \mathbf{n}$. In other words, we choose a stream function ψ_j associated with the j -th vortex, which satisfies

$$\left\{ \begin{array}{l} \Delta\psi_j = -\omega_j = -\Gamma_j\delta(\mathbf{r} - \mathbf{r}_j) \\ \left. \frac{d\psi_j}{dn} \right|_{\partial D} = 0 \end{array} \right.,$$

that is, equivalently, to choose $\psi_j(\mathbf{r}) = -\Gamma_j G_N(\mathbf{r}, \mathbf{r}_j)$, where $G_N(\mathbf{r}, \mathbf{r}_j)$ is the Green's function for the Neumann problem associated with the Laplace operator (Laplacian) in the domain D .

Retaking again the Euler system in the form

$$\Delta\psi = -\omega, u = \partial_y\psi, v = -\partial_x\psi, \frac{D\omega}{Dt} = 0,$$

we can write

$$\psi = -\frac{1}{2\pi} \int \omega(\mathbf{r}') \ln|\mathbf{r} - \mathbf{r}'| d\mathbf{r}',$$

and then we set $u = \partial_y\psi$, $v = -\partial_x\psi$. But these equations seem to be just the equations established for a system of point vortices, the integral representation for ψ being replaced by the formula $\psi = \sum_{j=1}^N \psi_j(\mathbf{r})$, valid in the conditions of a point vortices system analogously as a Riemann integral is approximated by a Riemann sum. This suggests that an inviscid incompressible flow can be approximated by the flow induced by a discrete system of vortices, The convergence of solutions of the discrete vortex equations to solutions of Euler's equations as $N \rightarrow \infty$ is studied in [38] and in [61].

Vortex systems provide both a useful tool in the study of general properties of Euler's equations and a good starting point for setting up effective algorithms for solving these equations in specific situations.

8. Thin Profile Theory

The theory of a wing with an infinite span (i.e., the theory of profiles) requires knowledge of the conformal mapping of the profile outside, from the physical plane (z) onto the outside of a disk from the plane (Z). However, for an arbitrary (wing) profile, it is difficult to get effectively this mapping; that is why, many times, one prefers the reverse procedure, that is to construct (wing) profiles as images of some circumferences through given conformal mappings. The Joukovski, Karman-Trefftz, von Mises, etc. profiles belong to this category [69].

In the particular case of the thin profiles with weak curvature, the problem of a flow past such a profile can be directly solved in a quite simple approximative manner. More precisely, this time it will not be necessary to determine the above mentioned conformal mapping but only the solving, in the physical plane, of a boundary value problem of

Hilbert type that reduces, in an acceptable approximation, to a Dirichlet problem for the Laplace equation.

8.1 **Mathematical Formulation of the Problem**

Suppose that our (wing) profile is formed¹¹ by the arcs C_1 and C_2 of equations

$$y = g_j(x) = \varepsilon h_j(x), \quad a \leq x \leq b, \quad j = 1, 2,$$

where $\varepsilon > 0$ is a very small positive parameter; we admit that the functions $h_1(x)$ and $h_2(x)$ are continuous and derivable in $[a, b]$ and $h_j(a) = h_j(b)$, $j = 1, 2$. Suppose also that $h_2(x) \geq h_1(x)$, $a \leq x \leq b$.

This profile is placed in a uniform fluid free-stream of complex velocity $W_\infty = V_\infty e^{-i\alpha}$, both the magnitude of the physical (attack) velocity at far field V_∞ and its angle of incidence α , sufficiently small, being independent of time. In what follows we will look for the complex potential of the fluid flow under the form

$$f(z) = W_\infty z + F(z)$$

or, focussing on the velocity field determination, we set

$$\frac{df}{dz} = W_\infty + \frac{dF}{dz} = u - iv$$

with

$$\frac{dF}{dz} = U - iV.$$

The unknown function $F(z)$, the *corrective* complex potential, induced by the presence of the thin profile, is a holomorphic function in the vicinity of any point at finite field, with a logarithmic singularity at infinity. On the contrary, the derivative of this function, $\frac{dF}{dz}$, is holomorphic in the entire outside of the profile, vanishing at infinity, that is $\left(\frac{dF}{dz}\right)_\infty = 0$. More, the above equality (for the velocity field) generates the representation

$$u = V_\infty \cos \alpha + U, \quad v = V_\infty \sin \alpha + V,$$

U and V playing the roles of some perturbation (corrective) velocities due to the presence in the free-stream of the thin profile.

¹¹Obviously it is about the cross-section of the profile in the plane xOy .

Just the regularity of the function $\frac{dF}{dz}$ in the whole outside of the considered profile leads to the idea of determining of *this* function instead of the corrective potential $F(z)$. To reach this purpose we need first to formulate the boundary conditions of the problem in terms of the functions U and V .

Since the unit normal vector to the contour C_j , of equation $g_j(x) - y = 0$, is $\mathbf{n} [g'_j(x), -1]$, the slip-condition along the walls C_j can be written

$$\mathbf{v} \cdot \mathbf{n} = u g'_j(x) - v = 0, \quad j = 1, 2.$$

Taking into account the above relationship between (u, v) and (U, V) we have finally the condition

$$V = -V_\infty \sin \alpha + g'_j(x) (V_\infty \cos \alpha + U) \quad \text{on } C_j, \quad j = 1, 2,$$

such that the velocity field determination comes to the solving of a Hilbert boundary value problem associated to the Laplace equation.

It is obvious that, additionally, we should observe the Joukovski condition to ensure the boundness of the velocity at sharp trailing edge (that is, at $x = b$).

So far we have not formulated, in the mathematical model associated to the problem, any simplifying hypothesis. Now we assume that $|U|$ is small enough to be neglected in the presence of $V_\infty \cos \alpha$ which agrees with the fact that the considered profile is thin and the incidence itself α is small. On the other hand we may assimilate the profile with the segment AB of the real axis and designating by C' this segment, by C'_2 its side corresponding to $y = +0$ and by C'_1 that corresponding to $y = -0$, the above boundary (slip-) condition could be approximated by

$$V = -V_\infty \sin \alpha + V_\infty \cos \alpha \cdot g'_j(x) \equiv l_j(x) \quad \text{pe } C'_j, \quad j = 1, 2.$$

Thus we are led, in view of the determination of the harmonic function $V(x, y)$, to a Dirichlet problem for the entire plane Oxy with a cut along the segment C' of the real axis.

8.2 Solution Determination

The solving of a Dirichlet problem joined to the Laplace operator for the whole plane with a cut along the segment AB of the real axis, to

which the problem of the fluid flow past a profile is reduced, is a classical issue in the literature¹².

The solution of this problem, applied to the function $i \frac{dF}{dz} = V + iU$ whose real part is known on the boundary AB , leads to

$$U - iV = \frac{1}{2\pi} \int_a^b \frac{l_2(\xi) - l_1(\xi)}{z - \xi} d\xi - \frac{i}{2\pi} \sqrt{P(z)} \int_a^b \frac{l_2(\xi) + l_1(\xi)}{(z - \xi) \sqrt{|P(\xi)|}} d\xi + k,$$

where k is a real constant, $P(z) = (z - a)(z - b)$ while the chosen determination for $\sqrt{P(z)}$ equals to $+\sqrt{P(x)}$ at $z = x > b$.

Unfortunately, this bounded solution of the proposed Dirichlet problem does not satisfy yet the condition expressing the rest of fluid at far distances i.e., $\left(\frac{dF}{dz}\right)_\infty = 0$. To satisfy this condition too we will add

to the previous solution a term of the type $i\lambda \sqrt{\frac{z-b}{z-a}}$, where λ is a real constant (not chosen yet) and the determination of the squared root is the same as the previous one (i.e., it is positive at $z = x > b$)¹³. Since in the neighborhood of infinity we have

$$\begin{aligned} \sqrt{P(z)} &= z \left\{ 1 - \frac{a+b}{2z} + \frac{1}{z^2} (\dots) \right\}, \quad \frac{1}{z-\xi} = \frac{1}{z} \left(1 + \frac{\xi}{z} + \frac{\xi^2}{z^2} + \dots \right), \\ i\lambda \sqrt{\frac{z-b}{z-a}} &= i\lambda \left\{ 1 + \frac{a-b}{2z} + \frac{1}{z^2} (\dots) \right\}, \end{aligned}$$

we could write

$$\begin{aligned} U - iV + i\lambda \sqrt{\frac{z-b}{z-a}} &= -\frac{i}{2\pi} \int_a^b \frac{l_2(\xi) + l_1(\xi)}{\sqrt{|P(\xi)|}} d\xi + k \\ &- \frac{i}{2\pi z} \int_a^b \frac{l_2(\xi) + l_1(\xi)}{\sqrt{|P(\xi)|}} \left(\xi - \frac{a+b}{2} \right) d\xi + \frac{1}{2\pi z} \int_a^b [l_2(\xi) - l_1(\xi)] d\xi \\ &+ \frac{1}{z^2} (\dots) + i\lambda \left\{ 1 + \frac{a-b}{2z} + \frac{1}{z^2} (\dots) \right\}. \end{aligned}$$

¹² A direct and elegant manner for solving this problem, even in the more general case of a boundary formed by n distinct segments on Ox , can be found, starting from page 201, in the book of C. Iacob [69].

¹³ Really, by adding to $U - iV$ a term in the form $i \frac{\lambda z + \mu}{\sqrt{P(z)}}$, where $\lambda, \mu \in \mathbb{R}$, the values of V on AB will not be modified.

Then, to ensure that at far distances ($|z| \rightarrow \infty$) the solution of our problem tends to zero, it is sufficient to choose the real constants λ and k so that

$$\lambda = \frac{1}{2\pi} \int_a^b \frac{l_2(\xi) - l_1(\xi)}{\sqrt{|P(\xi)|}} d\xi \quad \text{and } k = 0.$$

Finally we have for the complex velocity the representation

$$\frac{dF}{dz} = \frac{1}{2\pi} \int_a^b \frac{l_2(\xi) - l_1(\xi)}{z - \xi} d\xi - \frac{i}{2\pi} \sqrt{\frac{z-b}{z-a}} \int_a^b \frac{l_2(\xi) + l_1(\xi)}{(z - \xi)} \sqrt{\frac{\xi - a}{b - \xi}} d\xi,$$

a formula given by L. I. Sedov, but obtained via other technique [134].

On the other hand, as a complex potential $f(z)$, at far field, has an expansion under the form

$$f(z) = w_\infty z + \frac{\Gamma}{2\pi i} \log z + a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \dots$$

and implicitly the complex velocity is

$$\frac{df}{dz} = w_\infty + \frac{\Gamma}{2\pi i} \frac{1}{z} - \frac{a_1}{z^2} + \frac{1}{z^3} (\dots),$$

we get for the circulation Γ , necessarily¹⁴, the value

$$\Gamma = \int_a^b [l_2(\xi) - l_1(\xi)] \sqrt{\frac{\xi - a}{b - \xi}} d\xi.$$

This value corresponds to that obtained by the Joukowski condition (rule), the fluid velocity being, obviously, bounded at the sharp trailing edge. Supported by it we could also calculate the general resultant of the fluid pressures on the thin profile, namely we have¹⁵

$$R_x - iR_y = i\rho V_\infty e^{-i\alpha} \int_a^b [l_2(\xi) - l_1(\xi)] \sqrt{\frac{\xi - a}{b - \xi}} d\xi.$$

Details on the theory of a thin (wing) profile and even some extensions such as the case of the thin airfoil with jet, can be found in the book of C. Iacob [69]. The thin profile with jet in the presence of the ground has been studied in [113].

¹⁴In virtue of the uniqueness of such a series development.

¹⁵By applying directly the Blasius–Chaplygin formulas.

9. **Unsteady Irrotational Flows Generated by the Motion of a Body in an Inviscid Incompressible Fluid**

In what follows we will formulate the mathematical problem for determination of the fluid flow induced by a general displacement (motion) in the fluid mass of a rigid body, this fluid flow being unsteady (in general). Before considering separately either the 2-dimensional (plane) or the 3-dimensional case, we remark that the problem of a uniform displacement of a body with the velocity $-\mathbf{v}_\infty$ in a fluid at rest, is completely equivalent with the problem of a uniform free-stream of velocity \mathbf{v}_∞ past the same body but supposed fixed. This fact comes out at once, if one considers also, besides the fixed system of axes, a mobile reference frame rigidly linked to the body and we express the position vector (radius) of the same point within these two systems, namely $\mathbf{r} = \mathbf{r}' + \mathbf{r}_0$; then, by derivation, one deduces a similar relation between the velocity vectors expressed in the two systems, that is $\mathbf{v}' = \mathbf{v} + \mathbf{v}_\infty$. Hence, the rest state at infinity versus the fixed system ($\mathbf{v} = 0$), will be the state of a uniform motion with the velocity \mathbf{v}_∞ within the mobile system where the body could be seen fixed (being rigidly linked to it).

9.1 **The 2-Dimensional (Plane) Case**

In general, when we deal with the case of unsteady plane flows we need first to introduce a *fixed* system of axes OXY . With respect to this system, at any instant t , the flow will be determined by its complex potential $F(z, t)$, defined up to an additive function of time. The uniform derivative of this complex potential will provide the components U and V on the axes OX and OY .

The function $F(Z, t)$ in the domain (\mathcal{D}), where it is defined at any moment t , is either a uniform function (which means a holomorphic function of Z) or the sum of a holomorphic function and some logarithmic terms, the critical points of these last ones being interior to the connected components (Δ_q) of the complement of (\mathcal{D}). “A priori”, the coefficients $\frac{\Gamma_q + iD_q}{2\pi i}$ of these logarithmic terms can depend on time but, under our assumption, Γ_q are necessary constant. If this does not happen, the circulation along a fluid contour encircling (Δ_q), a contour which is followed during the motion, will not be constant, in contradiction with the Thompson theorem.

The determination of F should be done by using both the initial conditions (a specific feature for the unsteady flows) and the boundary conditions attached to the problem.

In particular, along a wall the normal component of the *relative* velocity of the fluid (versus the wall) should vanish. Concerning the pressure, it can be calculated by the Bernoulli theorem which, in this case, states that

$$p + \frac{\rho q^2}{2} - \rho U_m + \rho \frac{\partial \Phi}{\partial t} = C(t),$$

where the “constant” $C(t)$, depending on time, will be determined with the initial conditions.

An important case is when there is only *one mobile body (obstacle)* in the mass of the fluid, which allows a simple formulation of the initial and boundary conditions (on the body surface). More precisely, by considering a mobile reference frame (system of coordinates) Axy , rigidly linked to the obstacle (body), and by using the linear expression of Z as function of z (with the coefficients depending on time, in fact a change of variables, the flow being watched within the fixed frame OXY), we get first $f(z, t) \equiv F(Z, t)$ which represents the complex potential of the flow expressed in the variables z and t .

Hence for the components u and v of the velocity vector, we have $u - iv = \frac{df}{dz}$ (here u and v are the components of the absolute fluid velocity versus the fixed system OXY , these components being expressed in the variables x and y).

Let us now denote by $\alpha(t)$ and $\beta(t)$ the components on Ax and Ay respectively, of the vector \mathbf{v}_A , the velocity of the point A belonging to the body, and by $\Omega(t)$ the magnitude of the body rotation; the contour (surface) of the obstacle being then defined by the time free parametric equations $x = x(s)$, $y = y(s)$, the velocity \mathbf{v}_P of a point $P(s)$, belonging to this contour, is $\mathbf{v}_P = \mathbf{v}_A + \Omega \mathbf{k} \times \mathbf{AP}$ whose components are $\alpha - \Omega y$, $\beta + \Omega x$.

Then, the normal component of the relative velocity at the point P , belonging to the obstacle contour, is $(u - \alpha + \Omega y) \frac{dy}{ds} - (v - \beta - \Omega x) \frac{dx}{ds}$ such that the slip-condition can be written, for any fixed t , in the form

$$\frac{d\psi}{ds} = \alpha \frac{dy}{ds} - \beta \frac{dx}{ds} - \Omega \frac{d}{ds} \left(\frac{x^2 + y^2}{2} \right).$$

This last expression determines, to within an additive function of time, the value of ψ along the contour, precisely

$$\psi|_P = \alpha(t)y - \beta(t)x - \frac{\Omega(t)}{2} (x^2 + y^2).$$

9.2 The Determination of the Fluid Flow Induced by the Motion of an Obstacle in the Fluid. The Case of the Circular Cylinder

Let us consider an obstacle, bounded by the contour (C), which is moving in the fluid mass supposed at rest at infinity. We know that the circulation along the contour (C) is necessarily constant; in the sequel, we limit ourselves to the case when this constant is zero.

Our aim, using the above notation, is to determine at any instant t , a function $f(z)$ holomorphic outside (C), whose derivative $\frac{df}{dz}$ is zero at far distances and whose imaginary part along (C), fulfils the condition

$$\psi = \alpha(t)y - \beta(t)x - \frac{\Omega(t)}{2}(x^2 + y^2).$$

Suppose now, for sake of simplicity, that we solve first, the following particular cases of the initially proposed problem, which are distinct by the values characterizing the obstacle rototranslation:

- 1) $\alpha = 1$, $\beta = 0$, $\Omega = 0$;
- 2) $\alpha = 0$, $\beta = 1$, $\Omega = 0$;
- 3) $\alpha = 0$, $\beta = 0$, $\Omega = 1$.

In all these cases we may assume that the corresponding complex potential f is independent of time (the attached domains having a fixed in time shape); denote by $f^{(1)}(z)$, $f^{(2)}(z)$, $f^{(3)}(z)$, the complex potentials which correspond to these three cases respectively.

It is obvious that, in general, α , β , Ω being supposed arbitrary continuous functions of time, the function

$$f(z, t) = \alpha(t) f^{(1)}(z) + \beta(t) f^{(2)}(z) + \Omega(t) f^{(3)}(z)$$

represents a solution of the initial proposed problem¹⁶. One could prove that the flow thus determined is unique, according to the uniqueness of the respective Dirichlet problem. Concerning the effective determination of the functions $f^{(i)}(z)$, in the first two cases (when the displacement of the obstacle is a uniform translation of unit velocity) the fluid flow watched from Axy , can be identified with a steady flow of the type already studied in the section devoted to the theory of profiles. The third case is that of a uniform rotation. This case, as the previous two, can be explicitly solved if we know the canonical conformal mapping of the outside of (C) onto the exterior of a circular circumference.

¹⁶The solution of the respective Dirichlet problem being a linear functional of the boundary data.

Let us consider the simple case when (C) is a circular disk centered at A . First we remark that, in this case, the function $f^{(3)}(z)$ is constant and consequently we could eliminate the free of z term $\Omega(t) f^{(3)}(z)$. This result is obvious because the rotation of the disk with respect to its center does not influence the ideal fluid flow. The case when $\alpha = 1$ and $\beta = 0$ corresponds to the situation when (C) is performing a uniform translation along the Ox axis; with respect to (C) (the system Axy), the flow is steady with a velocity at infinity parallel to the Ox axis and whose algebraic magnitude, versus the same axis, is -1 ; then the complex potential associated to this relative flow is $-\left(z + \frac{R^2}{z}\right)$, R being the radius of (C) and consequently the absolute flow watched from the fixed system OXY , has as complex potential

$$f^{(1)}(z) = z - \left(z + \frac{R^2}{z}\right) = -\frac{R^2}{z},$$

which corresponds to a doublet located at the origin A of the plane z , and whose axis is collinear with the velocity. From here, we could deduce, at once, that in the case when the circular cylinder translates with arbitrary components (α, β) , the corresponding complex potential is

$$f(z) = -(\alpha + i\beta) \frac{R^2}{z}.$$

An important generalization of the above situation is the situation when the displacement of the obstacle in the fluid mass takes place in the presence of an unlimited wall (as it is the case of a profile moving in the proximity of the ground, that is the “ground effect” problem). At the same time a great interest arises from the fluid flow induced by a general rototranslation of a system of n arbitrary obstacles in the mass of the fluid. We will come again to this problem after the next section, by pointing out a new general method for approaching the plane hydrodynamics problem [111].

9.3 The 3-Dimensional Case

Consider now the three-dimensional flow induced by the motion of a rigid spatial body (obstacle) in the mass of fluid at rest at far field, i.e., it is about a generalization of the previous study made in the plane case. Let then $OX_1X_2X_3$ be the three-rectangular fixed system and the velocity potential of the absolute fluid flow $\Phi(X_1, X_2, X_3, t)$ be, at any moment, a harmonic function of X_i whose gradient (velocity) is zero at infinity. Introducing also the mobile system $Ax_1x_2x_3$ – rigidly linked to

the obstacle – but watching the absolute flow (that is versus the fixed system $OX_1X_2X_3$) we set again

$$\varphi(x_1, x_2, x_3, t) \equiv \Phi(X_1, X_2, X_3, t).$$

To determine this function φ , the velocity potential of the absolute flow but expressed in the variables of the mobile system $Ax_1x_2x_3$ (a function which is also harmonic and with zero gradient at infinity), we should write the slip-condition on the surface (Σ) of the obstacle. Let then \mathbf{v}_A and $\boldsymbol{\Omega}$ be the velocity of the point A , belonging to the obstacle, and, respectively, the obstacle rotation; these are known vectorial functions of time. At a point P of the contour (Σ) , if \mathbf{n} is the unit outward normal drawn to (Σ) at P , we have for the function φ the condition

$$\mathbf{U} \cdot \mathbf{n} = \frac{d\varphi}{dn} = \mathbf{v}_p \cdot \mathbf{n} = (\mathbf{v}_A + \boldsymbol{\Omega} \mathbf{k} \times \mathbf{AP}) \cdot \mathbf{n},$$

i.e., the projection of the relative velocity $\mathbf{U} - \mathbf{v}_p$ on \mathbf{n} is zero.

We denote now by V_1, V_2, V_3 the components of \mathbf{v}_A on the Ax_1, Ax_2, Ax_3 axes and by V_4, V_5, V_6 those of $\boldsymbol{\Omega}$ on the same axes; let also n_1, n_2, n_3 be the components of \mathbf{n} while n_4, n_5, n_6 are those of $\mathbf{AP} \times \mathbf{n}$ on the same axes of the reference frame $Ax_1x_2x_3$. With this notation, the above condition is

$$\frac{d\varphi}{dn} = \sum_{p=1}^6 n_p v_p.$$

While n_p are geometric entities depending only on P from (Σ) and not on t , v_p are known functions of time, independent of P from (Σ) .

Let us admit that there are the functions $\varphi^{(p)}(x_1, x_2, x_3)$ harmonic outside of (\mathcal{D}) so that $\frac{d\varphi^{(p)}}{dn} = n_p$ on (Σ) and whose *grad* $\varphi^{(p)}$ vanish at far distances. In fact the existence of these functions comes from the solving of a Neumann problem for the exterior of the domain (\mathcal{D}) , with the additional requirements that the first order partial derivative of $\varphi^{(p)}$ tends to zero when the point P tends to infinity.

It is known that such Neumann problems, in quite general conditions, admit one unique solution and only one [52].

Setting then

$$\varphi(x_1, x_2, x_3, t) \equiv \sum_{p=1}^6 v_p(t) \varphi^{(p)}(x_1, x_2, x_3),$$

this function φ satisfies all the conditions of the problem and defines the searched velocity potential for fluid flow outside the obstacle. Once the

function φ is determined, the pressure can be calculated by applying the Bernoulli theorem.

9.4 General Method for Determining of the Fluid Flow Induced by the Displacement of an Arbitrary System of Profiles Embedded in the Fluid in the Presence of an “A Priori” Given Basic Flow

In what follows we intend to give a brief survey on a new method which allows us the solving of any direct problem of plane hydrodynamics, i.e., to determine the fluid flow induced by a general displacement in the inviscid fluid mass, of an arbitrary system of profiles, possibly in the presence of unlimited walls, in the conditions of the pre-existence of an already given “basic” flow which could present even a (finite) number of singularities.

The great advantage of this method consists, not only in its generality but also in the fact that it can be easily adapted to the numerical calculations. A CVBM joined to this general method will be presented later in this book.

From the mathematical point of view, by avoiding the conformal mapping technique, the method solves the proposed problem by using some appropriate singular integral equations which, under our assumptions, lead to a system of regular integral Fredholm equations. By imposing some additional hypotheses on both the profiles and the “a priori” existing basic flow, one establishes also, together with the solving of the involved algebraic system, the existence and uniqueness theorems for the respective integral equations.

9.4.1 The Mathematical Considerations and the Presentation of the Method in the Case of Only One Profile Moving in an Unlimited Fluid

Let us consider¹⁷, as being given, a plane potential inviscid fluid flow called *the basic flow*. Let $w_B(z)$ be the complex velocity of this basic fluid flow.

Let us now imagine the fluid flow induced by a general displacement (roto-translation) of an arbitrary profile in the fluid mass. Of course this flow will superpose on that basic fluid flow. In what follows, we want

¹⁷For more details and even for the consideration of a general case of “ n ” profiles, one could read the paper of T. Petřila [103]. An extension of this method to the case of profiles with sharp trailing edge and of the influence of some unlimited walls on the flow can also be found in the papers of T. Petřila [102], [101].

to present a new method for determining the complex velocity $w_B(z)$ of the fluid flow which results by the just mentioned superposition, a method which could provide simple numerical algorithms for the whole flow pattern.

Concerning the curve C , one admits that its parametrical equation $z = \beta(\psi)$, defined for $\psi \in E_1$ and referred to a fixed system of rectangular Cartesian coordinates Oxy , fulfils the following conditions (I):

- (I)i) it is a 2π periodic bounded function in $[0, 2\pi)$;
- (I)ii) it is a Jordan positively oriented curve for $\psi = [0, 2\pi)$;
- (I)iii) it is a twice continuously differentiable function in $[0, 2\pi)$, with

$\beta(\psi) \neq 0$ and $\beta(\psi) < M$, M being a finite constant.

We remark that the restrictions imposed on the profile (C) will lead to the continuity of its curvature which implies the continuity of the kernels of the involved Fredholm integral equations.

In regard to the given function $w_B(z)$, it belongs to a class (a) of functions with the following properties:

(a) 1) they are holomorphic functions in the domain D_1 (the entire plane), except at a finite number (q) of points z_k placed at a finite distance, and which represent the singular points for these functions; let D_1^* be the domain D_1 from which one has taken out these singular points;

(a) 2) they are continuous and bounded functions in $\overline{D_1^*} \setminus \{z_k\}_{k=\overline{1,q}}$, a domain which contains also the point at infinity; let

$$w_B(\infty) = \lim_{|z| \rightarrow \infty} w_B(z);$$

(a) 3) they are Hölderian functions at the points of the curve $(C)^{18}$.

Let Γ_B be the circulation of the basic fluid flow which equals $\sum_{k=1}^q \Gamma_k$, that is equals the sum of the circulations of all the given singularities of the fluid flow.

Regarding the unknown function $w(z)$, the complex velocity of the resultant flow, it will be looked for in a class of functions (b) which satisfies the requirements:

(b) 1) it is a holomorphic function in the domain $D = D_1 \setminus \{\overline{intC}\}$, except the same points $\{z_k\}_{k=\overline{1,q}}$ which are singular points of the same nature as for $w_B(z)$ (i.e., the corresponding Laurent developments have the same principal parts);

¹⁸ Suppose that, during the displacement of the profile, we have $\{\overline{intC}\} \subset D_1^*$, which means the curves C do not intersect the points $\{z_k\}_{k=\overline{1,q}}$ which stay all the time outside of these curves.

(b) 2) it is a continuous and bounded function in $\overline{D^*} = \overline{D_1^*} \setminus \{z_k\}_{k=1,q} \setminus \{\text{int}C\}$, which also contains the point of infinity where $\lim_{|z| \rightarrow \infty} w(z) = w(\infty) = w_B(\infty)$;

(b) 3) it is a Hölderian function at the points of the contour (C) where it also satisfies the boundary condition:

There is a real function $v(\psi)$ such that for any $\psi \in [0, 2\pi)$, we have

$$\overline{w(\beta(\psi))} = v(\psi) \frac{\dot{\beta}(\psi)}{|\dot{\beta}(\psi)|} + l + im + i\omega [\beta(\psi) - z_A],$$

where l and m are given functions of time corresponding to the components of the transport (translation) velocity at the point $z_A \in \{\text{int}C\}$ while ω is also a function of time defining the instantaneous rotation;

(b) 4) it satisfies the equality $\int_C w(z) dz = \Gamma$, where Γ is an “a priori” given function.

Once all these mathematical assumptions have been introduced, the (unknown) function $w(z)$ is sought among the solutions of the following singular integral equation with a Cauchy kernel, namely

$$w(\xi) = w_B(\xi) - \frac{1}{2\pi i} \int_C \frac{w(z)}{z - \xi} dz, \tag{2.12}$$

where $\xi \in D^*$ ¹⁹.

In order to use the boundary (slip) condition on C , we now let $\xi \rightarrow z_0 = \beta(\psi^*) \in C$ and so we get

$$w(\beta(\psi^*)) = 2w_B(\beta(\psi^*)) - \frac{1}{\pi i} \int_0^{2\pi} \frac{w(\beta(\psi)) \dot{\beta}(\psi)}{\beta(\psi) - \beta(\psi^*)} d\psi$$

¹⁹The above representation for the complex velocity introduces a corrective complex potential (corresponding to the presence of the profile (C)) in the form of a continuous distribution of point vortices along the curve (C). We would get the same representation using Cauchy’s formula for the function $w(z) - w_B(z)$ and for the domain D_R , the cross-section of D with a disk centered at the point $\xi \in D$ and of radius R . Setting then $R \rightarrow \infty$ and taking into account that $\lim_{|z| \rightarrow \infty} (w(z) - w_B(z)) = 0$, we are necessarily led to the following relation for the desired function $w(z)$

$$w(\xi) = w_B(\xi) - \frac{1}{2\pi i} \int_C \frac{w(z)}{z - \xi} dz + \frac{1}{2\pi i} \int \frac{w_B(z)}{z - \xi} dz.$$

As regards the last term, it doesn’t play an essential role because the solution of the Fredholm integral equation (to which we are led), and which satisfies the condition with the “a priori” given circulation, is independent of it.

where \oint is the principal value (in the Cauchy sense) of the involved integral. Denoting then $\nu(\psi) \left| \dot{\beta}(\psi) \right|$ by $\gamma(\psi)$ and

$$\dot{\beta}(\psi) \{l - im - i\omega[\beta(\psi) - z_A]\}$$

by $\nu(\psi)^{20}$, we could write

$$\begin{aligned} & \frac{\gamma(\psi^*)}{\dot{\beta}(\psi^*)} + \frac{1}{\pi i} \oint_0^{2\pi} \frac{\gamma(\psi) d\psi}{\beta(\psi) - \beta(\psi^*)} \\ &= 2w_B(\beta(\psi^*)) - \frac{\nu(\psi^*)}{\beta(\psi^*)} - \frac{1}{\pi i} \oint_0^{2\pi} \frac{\nu(\psi) d\psi}{\beta(\psi) - \beta(\psi^*)}, \end{aligned}$$

that is

$$\begin{aligned} & \gamma(\psi^*) - \frac{1}{2\pi i} \oint_0^{2\pi} \gamma(\psi) \frac{2\dot{\beta}(\psi^*)}{\beta(\psi^*) - \beta(\psi)} d\psi \\ &= 2w_B(\beta(\psi^*)) \dot{\beta}(\psi^*) - \nu(\psi^*) + \frac{1}{2\pi i} \oint_0^{2\pi} \nu(\psi) \frac{2\dot{\beta}(\psi^*)}{\beta(\psi^*) - \beta(\psi)} d\psi. \end{aligned}$$

Separating now the real parts of both sides, we obtain the following integral equation of Fredholm type with continuous kernel, precisely

$$\gamma(\psi^*) - \frac{1}{2\pi} \oint_0^{2\pi} \gamma(\psi) K_{\beta\beta}(\psi^*, \psi) d\psi \tag{2.13}$$

$$= \text{Re} \left\{ 2w_B[\beta(\psi^*)] \dot{\beta}(\psi^*) - \nu(\psi^*) \right\} +$$

$$+ \frac{1}{2\pi} \oint_0^{2\pi} [\nu(\psi) K_{\beta\beta}(\psi^*, \psi) + \text{Im} \nu(\psi) L_{\beta\beta}(\psi^*, \psi)] d\psi = f(\psi^*),$$

where we have denoted

$$L_{\beta\beta}(\psi^*, \psi) + iK_{\beta\beta}(\psi^*, \psi) = \frac{2\dot{\beta}(\psi^*)}{\beta(\psi^*) - \beta(\psi)}.$$

²⁰ With this notation we could also write

$$\Gamma = \int_0^{2\pi} \gamma(\psi) d\psi + \int_0^{2\pi} \nu(\psi) d\psi = \int_0^{2\pi} \gamma(\psi) d\psi + 2\omega S, \text{ } S \text{ being the area bounded by } (C).$$

We remark that according to the above hypotheses, the right side is a Hölderian function, which implies that the solutions of this equation (if they exist), are also Hölderian functions.

To study the existence of the solution of this integral equation we will use the Fredholm alternative which is now applicable. According to this alternative, the existence of the solution is related to the fulfilment of the condition

$$\int_0^{2\pi} f(\psi^*) d\psi^* = 0.$$

Actually, the uniformity of the complex function $w_B(z)$ in the vicinity of C leads to

$$\int_0^{2\pi} \operatorname{Re} \left\{ 2w_B [\beta(\psi^*)] \dot{\beta}(\psi^*) - \nu(\psi^*) \right\} d\psi^* = 0;$$

meanwhile we also have

$$\begin{aligned} \int_0^{2\pi} d\psi^* \left[\operatorname{Re} \frac{1}{2\pi i} \int_0^{2\pi} \nu(\psi) \frac{2\dot{\beta}(\psi^*)}{\beta(\psi^*) - \beta(\psi)} d\psi \right] &= \int_0^{2\pi} d\psi \left[\operatorname{Re} \frac{\nu(\psi)}{2\pi i} \int_0^{2\pi} \frac{2\dot{\beta}(\psi^*)}{\beta(\psi^*) - \beta(\psi)} d\psi^* \right] \\ &= \int_0^{2\pi} \operatorname{Re} \nu(\psi) d\psi \end{aligned}$$

which proves that the condition

$$\int_0^{2\pi} f(\psi^*) d\psi^* = 0$$

is satisfied²¹.

Consequently the equation (2.13) admits a set of solutions of the form $\gamma = k\gamma^0 + \tilde{\gamma}$ where k is a real arbitrary constant, γ^0 is the unique non-zero solution of the homogeneous equation which also satisfies the condition $\int_0^{2\pi} \gamma^0(\psi) d\psi \neq 0$ while $\tilde{\gamma}$ is a particular solution of the non-homogeneous equation. It is easy to see that we can always choose one solution (that is the corresponding k) such that

²¹To interchange (commute) the integrals is possible due to the Bertrand–Poincaré formula.

$$\int_0^{2\pi} \gamma(\psi) d\psi + 2\omega S = \Gamma,$$

Γ being “a priori” given.

The previous results can be concisely formulated in both mathematical and fluid dynamics language, i.e., we have:

THEOREM 2.3. *For any curve C and complex function $w_B(z)$ belonging to the class (I) and (a) respectively, and for any continuous system of four real functions of time (l, m, ω, Γ) , there is only one solution of the above singular integral equation (2.12) which satisfies the conditions (b).*

Or, in hydrodynamical language,

For any profile C and a basic potential incompressible inviscid fluid flow with complex velocity $w_B(z)$, satisfying the conditions (I) and (a) respectively, and for any continuous displacement of this profile in the mass of the fluid, there is only one resultant fluid flow with an “a priori” given circulation which satisfies also the conditions (b).

10. **Notions on the Steady Compressible Barotropic Flows**

Suppose now that the inviscid fluid is compressible but limiting our interest to the case of the steady irrotational flow of a barotropic fluid. Further, for sake of simplicity, we will neglect the external mass forces $f(\mathbf{M})$.

10.1 **Immediate Consequences of the Bernoulli Theorem**

Our working hypotheses allow us to use the second Bernoulli theorem which can be written here in a very simple form, namely $h + \frac{q^2}{2} = h_0$, h_0 being a constant in the whole mass of the fluid and q the velocity modulus (magnitude). In this relation h is a function of p defined up to an additive constant, by the differential equality $\rho dh = dp$. Introducing now the equation of state under the form $p = g(\rho)$ (the fluid being compressible barotropic) we have also

$$dh = \frac{dp}{\rho} = \frac{1}{\rho} \frac{dg}{d\rho} d\rho = \frac{c^2}{\rho} d\rho,$$

c being the speed of sound in the fluid and which is defined as $c^2 = \frac{dg}{d\rho}$. So that it comes out that h will be an increasing function not only of p (see the above definition) but also of ρ .

When h_0 is known the Bernoulli theorem allows, by using also the equation of state, the calculation of h, p, ρ as functions of the velocity modulus q .

Now we shall show that the functions h, p, ρ, c are always decreasing functions with respect to q . For h it comes directly from the above Bernoulli theorem; p and ρ being also increasing functions of h (as inverse functions of increasing ones), they will be decreasing functions of q too. Finally, from $c^2 = \frac{dg}{d\rho}$ and from the hypotheses made on the state equation $p = g(\rho)$ ($\frac{dq}{d\rho} > 0$ and $\frac{d^2g}{d\rho^2} \geq 0$), it can deduce that c^2 is non-decreasing with respect to ρ and hence the above stated property is valid for c too.

The Mach number denoted by M , is the ratio q/c ; so that M is always an increasing function of q .

A last entity which plays an important role in the study of these fluid flows is the mass flux density ρq . For this we have

$$\frac{d(\rho q)}{\rho q} = \frac{dq}{q} + \frac{d\rho}{\rho} \frac{dp}{dp} = \frac{dq}{q} + \frac{dh}{c^2} = \left(1 - \frac{q^2}{c^2}\right) \frac{dq}{q}.$$

We remark that ρq is an increasing function of q (although ρ is decreasing with respect to q) if $M < 1$, that is *the flow is subsonic* while it is a decreasing function of q if $M > 1$, that is *the flow is supersonic*.

In the current applications we will presume that the barotropic fluid is an ideal gas in an adiabatic evolution so that $p = k\rho^\gamma$, k being a positive constant and γ , the adiabatic index, being also a constant greater than unity (for air $\gamma = 1.4$). In this case we have $c^2 = k\gamma\rho^{\gamma-1} \frac{p}{\rho} = \gamma \frac{p}{\rho}$ and, correspondingly, since $\rho dh = dp$, we could take for h the assessment

$$h = \int \frac{k\gamma\rho^{\gamma-1}}{\rho} d\rho = \frac{k\gamma\rho^{\gamma-1}\rho}{(\gamma-1)\rho} = \frac{\gamma}{\gamma-1} \frac{p}{\rho} = \frac{c^2}{\gamma-1}.$$

Denoting by p_0, ρ_0, c_0 the values taken by p, ρ, c at the point of zero velocity ($q = 0$), we could also write

$$\frac{h}{h_0} = \frac{c^2}{c_0^2} = \left(\frac{\rho}{\rho_0}\right)^{\gamma-1} = \left(\frac{p}{p_0}\right)^{\frac{\gamma-1}{\gamma}},$$

relations which, together with the Bernoulli theorem already written at the beginning of the section ($h = h_0 - \frac{q^2}{2}$), lead to

$$\begin{cases} c^2 = c_0^2 - \frac{(\gamma-1)q^2}{2} = c_0^2 \left(1 - \frac{(\gamma-1)q^2}{2c_0^2}\right) \\ \rho = \rho_0 \left(1 - \frac{(\gamma-1)q^2}{2c_0^2}\right)^{\frac{1}{\gamma-1}} \\ p = p_0 \left(1 - \frac{(\gamma-1)q^2}{2c_0^2}\right)^{\frac{\gamma}{\gamma-1}}, \end{cases} \quad (2.14)$$

i.e., to the formulas which give explicitly the dependences $c(q)$, $\rho(q)$ and $p(q)$. The functions (2.14) point out an important property which is specific only to the compressible fluid flows: the constant c_0 being known, it will be impossible for the fluid to overtake during its flow, a certain maximum velocity q_m , given by

$$q_m^2 = \frac{2}{\gamma-1} c_0^2.$$

Such a restriction does not occur in the case of the incompressible flow. When $q \rightarrow q_m$, the quantities p , ρ , c defined by (2.14) tend to zero and so the Mach number increases indefinitely. On the other hand, if at a point of the flow domain the fluid velocity is equal to the sound speed, that is $q = c = c^*$, then, from the same (2.14), we get

$$c^{*2} = \frac{2c_0^2}{\gamma+1} = \frac{\gamma-1}{\gamma+1} q_m^2.$$

The quantity c^* will be a constant called *the critical sound speed in fluid*. In virtue of the already established properties (with regard to the Mach number, for instance), at a certain point the flow is subsonic or supersonic as q is inferior or superior of c^* .

We remark that if our compressible fluid is also perfect, in the sense of the Clapeyron law acceptance together with the constancy of the specific heats C_P and C_V , we will also have $h = C_p T$. But then, in the same conditions of an adiabatic process, we could deduce that

$$\frac{\rho}{\rho_0} = \left(\frac{T}{T_0}\right)^{\frac{1}{\gamma-1}} \text{ and the previous relations should be completed with } T = T_0 \left(1 - \frac{(\gamma-1)q^2}{2c_0^2}\right).$$

It is important to understand in which context the incompressible fluid flow could approximate the compressible fluid flows. If we denote by G the inverse function of $p = g(\rho)$, that is $\rho = G(p)$, the incompressible case corresponds to $G \equiv \text{constant}$. As $\frac{dq}{d\rho} = c^2$ is the inverse of $\frac{dG}{dp}$, we can

see that an incompressible fluid shows up as a limit case of compressible barotropic fluid when *the sound speed is infinity large, i.e., the Mach number is zero everywhere.*

In the adiabatic case, in a domain where q is sufficiently small to support the development

$$p = p_0 \left(1 - \frac{\gamma q^2}{2c_0^2} + \frac{\gamma q^4}{8c_0^4} + \dots \right) = p_0 - \frac{1}{2} \rho_0 q^2 \left(1 - \frac{q^2}{4c_0^2} \right) + \dots,$$

with $c_0^2 = \gamma \frac{p_0}{\rho_0}$, we can see that, in the case when the velocity q is such that the quantity $\frac{q^2}{4c_0^2}$ could be neglected versus the unity, we reobtain the Bernoulli theorem for the incompressible fluid, which means $p = p_0 - \frac{1}{2} \rho_0 q^2$, so that the compressibility effects don't arise.

10.2 The Equation of Velocity Potential (Steichen)

The envisaged flows being irrotational, the velocity vector \mathbf{v} depends on a velocity potential $\Phi(x_1, x_2, x_3)$, i.e., there is the representation $\mathbf{v} = \text{grad} \Phi$ or $v_i = \Phi_{,i}$. This function, as in the incompressible case, will satisfy a partial differential equation which could be determined, for instance, by introducing the above representation into the equation of continuity. More precisely, the equation of continuity could be written (the flow being steady) as

$$\rho v_{i,i} + \rho_{,i} v_i = 0, \quad (\text{div}(\rho \mathbf{v}) = \rho \text{div} \mathbf{v} + \text{grad} \rho \cdot \mathbf{v} = 0).$$

But, using the Bernoulli theorem already written in the previous section ($h + \frac{q^2}{2} = h_0$) and the definitions of c^2 and h as well, we have in the entire fluid mass

$$0 = dh + qdq \implies dp + \rho q dq = dp \frac{dp}{d\rho} + \rho q dq = c^2 d\rho + \rho q dq = 0$$

so that

$$c^2 \rho_{,i} + \rho \left(\frac{q^2}{2} \right)_{,i} = 0$$

or

$$c^2 \text{grad} \rho = -\rho \text{grad} \frac{q^2}{2}.$$

Correspondingly, the equation of continuity, after a division by ρ and a multiplication by c^2 , becomes

$$c^2 \operatorname{div} \mathbf{v} - \mathbf{v} \cdot \operatorname{grad} \left(\frac{q^2}{2} \right) = 0.$$

The flow being irrotational we also have $v_i = \Phi_{,i}$ and $q^2 = (\operatorname{grad} \Phi)^2 = \Phi_{,k} \Phi_{,k}$ so that we can write

$$c^2 \Phi_{,ii} - \frac{1}{2} \Phi_{,i} (\Phi_{,k} \Phi_{,k})_{,i} = (c^2 \delta_{ik} - \Phi_{,i} \Phi_{,k}) \Phi_{,ik} = 0,$$

which represents the looked for equation. This partial differential equation of second order is obviously nonlinear and contains only the derivatives of Φ since we have established that c^2 is a function of q^2 , that is of $\Phi_{,k} \Phi_{,k}$. In the case of the 2-dimensional flows, by setting $x_1 = x$, $x_2 = y$, ($u_1 = u$, $u_2 = v$) we can see that $\Phi(x, y)$ is the solution of the equation

$$\left(1 - \frac{u^2}{c^2} \right) \frac{\partial^2 \Phi}{\partial x^2} - \frac{2uv}{c^2} \frac{\partial^2 \Phi}{\partial x \partial y} + \left(1 - \frac{v^2}{c^2} \right) \frac{\partial^2 \Phi}{\partial y^2} = 0.$$

The type of this equation, called also *the Steichen equation*, depends on the position of the Mach number versus the unity²² which reflects, from the mathematical (analytical) point of view, the profound difference that exists between the subsonic and supersonic flows. So, if $q < c$, the subsonic flows, the equation is of elliptic type while if $q > c$, the supersonic flows, the equation is of hyperbolic type. In the case when for certain regions we have $q < c$ and for others $q > c$, the equation is of mixed type and the associated flow is called *transonic*; in this situation the curves along which the transition from a type to another takes place, that is the curves $q = c \equiv c_m$, are called the *sonic lines*.

As regards the asymptotic behaviour of Φ at far distances, Finn and Gilbarg have proved that, in the subsonic case [46]

$$\Phi = V_\infty (X \cos \alpha + Y \sin \alpha) + \frac{m}{4\pi\beta} \ln(x^2 + \beta^2 y^2) + \frac{\Gamma}{2\pi} \operatorname{arctg} \frac{\beta y}{x}$$

where V_∞ is the constant magnitude of the attack (free-stream) velocity with the incidence α versus OX , $\beta^2 = 1 - M_\infty^2 > 0$, $z = Ze^{-i\alpha}$, $\frac{m}{\beta}$ is the flow-rate and Γ the circulation.

²² The determinant of this equation being $\left(1 - \frac{u^2}{c^2} \right) \left(1 - \frac{v^2}{c^2} \right) - \frac{u^2 v^2}{c^4} = 1 - \frac{u^2 + v^2}{c^2} = 1 - \frac{q^2}{c^2} = 1 - M^2$.

Before focussing on a simple application of this equation we remark that the Steichen equation is equivalent with the system

$$u = \frac{\partial \Phi}{\partial x} = \frac{\rho_0}{\rho} \frac{\partial \psi}{\partial y}, \quad v = \frac{\partial \Phi}{\partial y} = -\frac{\rho_0}{\rho} \frac{\partial \psi}{\partial x}$$

where $\psi(x, y)$ is the stream function which can be directly introduced through the continuity equation, a system which is not a Cauchy–Riemann system any more but a nonlinear one, ρ being a function of $q^2 = (\text{grad } \Phi)^2$. Obviously, in the incompressible case because $\rho = \rho_0$, we reobtain the Cauchy–Riemann system.

Finally, by expressing the Steichen equation through the stream function ψ , we remark the invariance of the form of this equation, which means

$$\left(1 - \frac{u^2}{c^2}\right) \frac{\partial^2 \psi}{\partial x^2} - \frac{2uv}{c^2} \frac{\partial^2 \psi}{\partial x \partial y} + \left(1 - \frac{v^2}{c^2}\right) \frac{\partial^2 \psi}{\partial y^2} = 0,$$

where u and v are now considered as functions of ψ .

Concerning the boundary condition attached to these equations, they come to $\mathbf{v} \cdot \mathbf{n} = \frac{d\Phi}{dn} = 0 = \frac{d\psi}{ds}$ and so $\psi = \text{constant}$ on the fixed obstacle (wall) while, at far field, supposing that the velocity \mathbf{v}_∞ is parallel to the Ox axis, we have

$$\left(\frac{\partial \Phi}{\partial x}\right)_\infty = v_\infty, \quad \left(\frac{\partial \Phi}{\partial y}\right)_\infty = 0,$$

respectively

$$\frac{\rho_0}{\rho_\infty} \left(\frac{\partial \psi}{\partial y}\right)_\infty = v_\infty, \quad \left(\frac{\partial \psi}{\partial x}\right)_\infty = 0, \quad (\rho_\infty = \rho(v_\infty^2)).$$

We remark that if we accept, instead of barotropy, an equation of state under the form $p = p(\rho, s)$ while the fluid flow is now rotationally steady, the equation for the stream function becomes [153]

$$(c^2 - u^2) \psi_{xx} - 2uv\psi_{xy} + (c^2 - v^2) \psi_{yy} = -\rho \left[c^2 \omega + q^2 \left(\frac{\partial p}{\partial s}\right)_p \frac{ds}{d\psi} \right].$$

Obviously, in the irrotational ($\omega \equiv 0$) and homentropic ($s = \text{constant}$) case, we reobtain the above determined equation.

10.3 Prandtl–Meyer (Simple Wave) Flow

Consider now the plane fluid flows whose velocity potential is of the form $\Phi = ru(\theta)$, the variables r and θ being the polar coordinates of

a current point P of the plane. Let \mathbf{i} be the unit vector of OP while \mathbf{j} is the unit vector which is obtained by rotating \mathbf{i} with $+\frac{\pi}{2}$. Since $d\Phi = udr + ru'd\theta$, we can write

$$\mathbf{v} = \text{grad}\Phi = u(\theta)\mathbf{i} + u'(\theta)\mathbf{j}$$

(r being the Lamé coefficient for the variable θ).

Then the velocity vector will remain equipotent with itself along any half-straight line emanating from the origin ($\theta = \text{const}$).

Conversely, it is proved that any irrotational flow with the above property admits a velocity potential of the form $\Phi = ru(\theta)$.²³

Remarking that $q^2 = u^2 + u'^2$ and therefore $\text{grad}\frac{q^2}{2} = \frac{u'}{r}(u + u'')\mathbf{j}$ (because $\text{grad}q^2 = \frac{\partial q^2}{\partial r}\mathbf{i} + \frac{1}{r}\frac{\partial q^2}{\partial\theta}\mathbf{j}$ while $\frac{\partial q^2}{\partial r}$ is obviously zero) together with $\text{div}\mathbf{v} = \text{divgrad}\Phi = \Delta\Phi = \frac{u+u''}{r}$, the equation $c^2\text{div}\mathbf{v} - \mathbf{v}\text{grad}\left(\frac{q^2}{2}\right) = 0$ which is often written as $\Delta\Phi - \frac{1}{2c^2}\text{grad}\Phi \cdot \text{grad}q^2 = 0$, becomes $(c^2 - u'^2)(u + u'') = 0$.

If $u + u'' = 0$, u will be a linear function of $\sin\theta$ and $\cos\theta$ while Φ is a linear function of x and y , the flow being thus uniform. By avoiding this trivial solution, we keep necessarily $u'^2 = c^2$ so that the modulus of the normal component to \mathbf{OP} of the velocity is equal with the local speed of sound²⁴. The flow will be thus supersonic.

Denoting by α the angle made by \mathbf{v} to \mathbf{OP} ($0 < \alpha < \frac{\pi}{2}$), then $\sin\alpha = \frac{1}{M} = \left(\frac{c}{q}\right)$, M being the Mach number at P . The angle α is, by definition, the *Mach angle* at the same point P .

Finally, let us write again the Bernoulli equation $h + \frac{q^2}{2} = h_0$. Admitting that the fluid flow is barotropic in adiabatic evolution, this becomes

$$\frac{q^2}{2} + \frac{c^2}{\gamma - 1} \equiv \frac{u^2 + u'^2}{2} + \frac{u'^2}{\gamma - 1} = \frac{c_0^2}{\gamma - 1},$$

which is a differential equation for determining of $u(\theta)$. To solve this equation we shall introduce the parametric representations

$$u = \sqrt{\frac{2}{\gamma - 1}}c_0 \cos\chi \quad \text{and} \quad u' = \sqrt{\frac{2}{\gamma + 1}}c_0 \sin\chi$$

which finally lead to a representation of the solution in the form

$$u(\theta) = q_m \cos\mu(\theta - \theta_m),$$

²³ The expression for the Laplacian in polar coordinates, being

$$\Delta\Phi = \frac{\partial^2\Phi}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\theta^2} + \frac{1}{r}\frac{\partial\Phi}{\partial r}.$$

²⁴ The curves with this property are also called *Mach lines*.

q_m being the maximum of the fluid velocity while $\mu = \sqrt{\frac{\gamma-1}{\gamma+1}}$.

A flow of this type is called a *simple wave* or *Prandtl–Meyer flow*; it occurs, for example, in the conditions of a supersonic flow past a sharp convex corner (dihedron) made by plane walls (see Figure 2.9). The

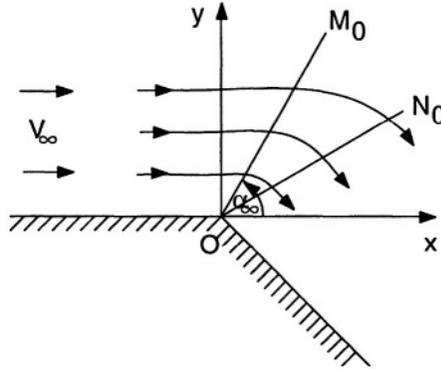


Figure 2.9. The simple wave flow past a convex dihedral

involved flow is uniform in the region delimited by the first horizontal wall and the Mach line OM_0 of equation $\theta = \alpha_\infty^{25}$, where α_∞ is the Mach angle corresponding to v_∞ ; along this Mach line a “matching” with a simple wave flow takes place, this simple wave flow acting in the “fan” (OM_0, ON_0) .²⁶ Once the “expansion” is achieved, the flow becomes again uniform and parallel with the second wall OE .

For details one can consult [69].

10.4 Quasi-Uniform Steady Plane Flows

The examples envisaged in the previous sections have shown that the complete solving of many problems arising from fluid dynamics seems to be extremely difficult even in the case of an inviscid fluid. The main difficulty comes from the nonlinear character of the appropriate mathematical problem, which is obvious in the case of a compressible flow.

²⁵ The existence of such a line is supported by the fact that the perturbation induced by the dihedral vertex could not be transmitted upstream (the sound speed c_∞ being less than the velocity v_∞ which is downstream oriented) and so it will propagate just along OM_0 .

²⁶ Along ON_0 , the radius limiting the fan-expansion, the velocity either takes its maximum value q_m or is parallel with the wall OE , the flow becoming uniform. In the case of a “cuspidal” dihedral (i.e. with an upstream oriented concavity) instead of a fan-expansion we will have a “compression”, i.e. a supersonic flow with a shock wave (a velocity discontinuities line) located in the vicinity of the corresponding half-straight line ON_0 .

If the flow is incompressible and irrotational, the equations are linear while the boundary conditions could become, sometimes, nonlinear such that the “superposition” principle does not apply any more. Finally, even if the problem is entirely linear, it is very often impossible to get an explicit analytical solution.

Due to all these difficulties, sometimes it is advisable to reasonably involve “deep” schemas which allow a better approach to such problems. In this view the linearizing method behaves like a very useful study tool which allows us, by simplifying the problem formulation, to get explicit (approximate) solutions in many and various situations. Naturally, we should always analyze the validity of the obtained results.

10.5 General Formulation of the Linearized Theory

Suppose that as an “unperturbed” flow, a uniform flow of velocity $\mathbf{v}_\infty(U_\infty, 0)$, parallel to the Ox axis is considered. In this flow, the mass density and the pressure are denoted by ρ_∞ and p_∞ respectively and, if the flow is compressible, we denote by c_∞ the sound speed (which is the same at any point of the flow domain). To simplify the writing of the below formulas, one could choose U_∞ as a velocity unit and in this case c_∞ is the inverse of the Mach number which is simply denoted by M . Suppose now that this given uniform flow (stream) is perturbed by introducing of some disturbance factors²⁷, thus having for velocity, pressure and mass density respectively, the representations of the type $U_\infty + \eta u$, ηv , $p_\infty + \eta p$, $\rho_\infty + \eta \rho$, defining entities which characterize the new (perturbed) fluid flow. Here η is a small parameter whose mechanical significance should be made precise in every particular problem.

It easy to see that the determination of this new flow comes to precise these functions u, v, p, ρ . But the equations connecting the unknown functions $u(x, y)$, $v(x, y)$, $p(x, y)$, $\rho(x, y)$ could be obtained by pointing out that the total derivative of a quantity, which is zero in the unperturbed flow, comes now to the operator $U_\infty \frac{\partial}{\partial x}$ ²⁸.

So that the equation of continuity and the Euler equations become, keeping only the main terms (of first order) in η (which agrees with the linearizing principles)

²⁷ Such a perturbation could occur when, for instance, the uniform stream meets a profile, etc.

²⁸ Really, from $\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \mathbf{grad}$, by using both the flow steadiness and the expression linearizing, we get this result.

$$U_\infty \frac{\partial \rho}{\partial x} + \rho_\infty \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = 0,$$

$$U_\infty \frac{\partial u}{\partial x} + \frac{1}{\rho_\infty} \frac{\partial p}{\partial x} = 0, \quad U_\infty \frac{\partial v}{\partial x} + \frac{\partial p}{\partial y} = 0,$$

assuming obviously that the mass (external) forces \mathbf{f} can be neglected²⁹.

If the fluid is incompressible we have $\rho = \rho_\infty$ and the above three equations form a linear system in the three unknown functions u, v, p . If the fluid is barotropic compressible, from the state equation we have

$$p_\infty + \eta p = g (\rho_\infty + \eta \rho),$$

which means, keeping only the principal (main) terms in η ,

$$p = \left(\frac{dg}{d\rho} \right)_\infty \rho = c_\infty^2 \rho,$$

an equation which completes the above system of three equations.

In what follows we will focus on the case when the perturbation of the uniform flow is due to the presence, in this uniform stream, of an obstacle (profile). Before analyzing the boundary conditions on the obstacle we will make precise the conditions joined to the fluid behaviour at infinity.

10.6 Far Field (Infinity) Conditions

Obviously, the entities u, v, p, ρ which characterize the perturbed flow will tend to zero upstream (in an exact formulation, it is possible to find an abscissa x_0 such that for $x < x_0$ these entities are arbitrarily small). This condition allows us to simplify the above written system. Thus, the second equation

$$\rho_\infty U_\infty \frac{\partial u}{\partial x} + \frac{\partial p}{\partial x} = 0,$$

shows that $p + \rho_\infty U_\infty u$ is a function only of y ; but from the imposed condition, this function is necessarily zero because it tends to zero when $x \rightarrow -\infty$ and therefore

$$p = -\rho_\infty U_\infty u.$$

If this value of p is introduced in the third equation of the system, that is in

²⁹ Here, the obvious equalities $\frac{d\rho_\infty}{dt} = 0$; $\frac{\partial \rho}{\partial t} = 0$; $\frac{\partial U_\infty}{\partial x} = 0$ have been used.

$$\rho_{\infty} U_{\infty} \frac{\partial v}{\partial x} + \frac{\partial p}{\partial x} = 0,$$

we reobtain the irrotational feature of the flow $\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = 0\right)$, so that there is a potential $\varphi(x, y)$ of the perturbation velocity, i.e., the components of the (perturbation) velocity admit the representation

$$u = \frac{\partial \varphi}{\partial x}, v = \frac{\partial \varphi}{\partial y}.$$

In principle, φ is precise to within an additive constant; we could fix this constant by defining φ as $\varphi(x, y) = \int_{-\infty}^x u(\xi, y) d\xi$ ³⁰ which implies that $\varphi \rightarrow 0$ when $x \rightarrow -\infty$ (y being fixed).

We shall also admit that $v(x, y)$ could be expressed by the derivative of the above integral, the commutation of the derivative and of the integral being ensured.

At last, the first equation of the system (that of continuity), taking into account all we have already obtained, leads to the following partial differential equation for the function φ ³¹

$$(1 - M^2) \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0,$$

M being the Mach number of the unperturbed flow. Conversely, any solution of this equation defines through the above formulas, a perturbed flow.

10.7 The Slip-Condition on the Obstacle

Let there be, in the fluid mass, an unbounded (of infinite span) cylindrical obstacle whose right section in the plane Oxy is (Σ) . To legitimise the linearization, the tangent drawn at any point of the contour of this section (Σ) must make a very small angle to the Ox axis, the velocity vector being oriented just along this tangent. More precisely, we suppose that the section (Σ) is delimited by a closed contour, infinitely close to the segment $-\frac{l}{2} \leq x \leq \frac{l}{2}$ of the Ox axis and which is defined by the equations

$$y = \eta F^+(x), y = \eta F^-(x), \quad (F^+(x) \geq F^-(x)),$$

³⁰ It is assumed that the written integral exists (it has "a sense"); it is a moment hypothesis which should be checked once the effective solution is obtained.

³¹ Taking into account that $\rho = \frac{p}{c_2^2} = -\frac{\rho_{\infty} U_{\infty}}{c_2^2}$.

where F^+ and F^- are some given functions defined on $-\frac{l}{2} \leq x \leq \frac{l}{2}$, sufficiently smooth on this interval and taking equal values at the ends of it³².

Once these considerations are made, always within the linearized theory, the unknown functions of the problem (of the “perturbed” flow) will be supposed defined in the whole plane (x, y) except the cut $y = 0$, $-\frac{l}{2} \leq x \leq \frac{l}{2}$.

But then, the slip-condition along the profile surface, expressed on the two “sides” of the cut, could be written as

$$v(x, +0) = U_\infty \delta^+(x), v(x, -0) = U_\infty \delta^-(x),$$

where $\delta^+(x)$ and $\delta^-(x)$ are the derivatives of F^+ and F^- with respect to x ³³.

10.8 The Similitude of the Linearized Flows. The Glauert–Prandtl Rule

Suppose, for instance, that we deal with the subsonic flows.

By setting $\beta^2 = 1 - M^2$, $\varphi(x, y)$ will be the solution of the elliptic equation

$$\beta^2 \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0.$$

Let us now consider a change of variables and functions, defined by $\bar{x} = x$, $\bar{y} = \beta y$, $\bar{\varphi}(\bar{x}, \bar{y}) = \beta \varphi(x, y)$. The function $\bar{\varphi}(\bar{x}, \bar{y})$ is a harmonic function in the variable \bar{x} and \bar{y} , which means

$$\frac{\partial^2 \bar{\varphi}}{\partial \bar{x}^2} + \frac{\partial^2 \bar{\varphi}}{\partial \bar{y}^2} = 0$$

Further, we also have

$$\bar{v}(\bar{x}, +0) = \frac{\partial \bar{\varphi}}{\partial \bar{y}}(\bar{x}, +0) = \frac{\partial \varphi}{\partial y}(x, +0) = v(x, +0)$$

and, analogously,

$$\bar{v}(\bar{x}, -0) = v(x, -0).$$

³² It says (in aerodynamics) that $y = \eta F^+$ defines the *upperside* of the profile while $y = \eta F^-$ defines its *lowerside*.

³³ Really the slip-condition $\mathbf{v} \cdot \mathbf{n} = 0$ expressed, for instance, on the upperside will be written as $(U_\infty + \eta u) \eta \delta^+(x) - \eta v = 0$ what leads, by linearizing, to the above result.

Thus, if the potential $\varphi(x, y)$ defines a “neighboring” flow versus another one of Mach number M , i.e., the function φ satisfying the equation of the perturbed flow together with the conditions at far field and the slip condition along the contour of the given profile (Σ), then the function $\bar{\varphi}(\bar{x}, \bar{y})$ will define a perturbed flow governed by the harmonic equation of the incompressible fluid ($M = 0$), with the *same* conditions at far distances and slip-condition along *the same* contour of the profile (Σ).

In this respect the study of a linear subsonic flow could be always reduced to that of an attached incompressible fluid flow. This result is of great practical importance, the study being essentially simplified by reducing the compressible problem to an incompressible one. By collecting all the formulas which allow the complete determination of the compressible case using the data of the attached incompressible problem, we get the so-called *Glauert–Prandtl rule (method)*.

More details on this parallelism of the mentioned flows can be found, for example, in the book of C. Iacob [69].

Obviously, in the conditions of a supersonic flow with $M > 1$, if again $\beta^2 = M^2 - 1$, we will obtain the equations $\beta^2\Phi_{xx} - \Phi_{yy} = 0$ or $\beta^2\bar{\psi}_{xx} - \bar{\psi}_{yy} = 0$, both of them being hyperbolic. A general solution of these equations is

$$\Phi = F_1(x - \beta y) + F_2(x + \beta y),$$

with F_1 and F_2 sufficiently smooth arbitrary functions. The curve $\xi_{\pm} = x \pm \beta y = \text{constant}$, the characteristics of our hyperbolic equations (and which are, generally, weak discontinuities curves) are the Mach lines (or waves).

We can see that the inclination of these curves is given by $tg\theta = \pm \frac{1}{\beta} = \pm \frac{1}{(M^2 - 1)^{\frac{1}{2}}}$, that is $\theta = \pm \arcsin\left(\frac{1}{M}\right)$ and therefore θ is the Mach angle.

Under these circumstances, the propagation velocity \mathbf{v} , joined to the presence of an obstacle in the fluid mass, satisfies the same equation such that we have $\mathbf{v}' = \text{grad}(F_1 + F_2)$ while the total velocity is given by $\mathbf{v} = \mathbf{V}_{\infty} + \mathbf{v}'$ (\mathbf{V}_{∞} being the attack velocity).

A simple calculation points out that the projections on the Mach lines of this total velocity, are constant in the sense that along a Mach line from *a family* (of Mach lines), the projection of the velocity on the Mach lines from *another family* remains constant.

The linearization of the supersonic flow equations is known as the method of J. Ackeret, the equivalent of the Glauert–Prandtl method for the subsonic flows [69].

11. Mach Lines. Weak Discontinuity Surfaces

Let us reconsider the Steichen equation to which we attach a Cauchy condition. In the hydrodynamical language this Cauchy problem applies to the determination of the fluid flow in the proximity of a given analytical arc C , of equation $x = x(\alpha)$, $y = y(\alpha)$, $\alpha \in [\alpha_0, \alpha_1]$, by knowing a distribution of velocity along this arc, given by $u = u(\alpha)$, $v = v(\alpha)$.

Obviously once $u = \frac{\partial \varphi}{\partial x}$ and $v = \frac{\partial \varphi}{\partial y}$ on the arc C have been determined, the velocity potential will be also known on this arc. But for the effective determination of φ (the flow) in a vicinity of the arc C (which is synonymous with the possibility to envisage a Taylor development for φ) it is important that both the arc C and the data on it satisfy some regularity requirements.

It is shown [69] that the Steichen equation being of Monge type, the Cauchy problem is *not* possible for those arcs and data which satisfy the differential relation

$$\left(1 - \frac{u^2}{c^2}\right) dy^2 + \frac{2uv}{c^2} dx dy + \left(1 - \frac{v^2}{c^2}\right) dx^2 = 0.$$

If $\lambda_1(u, v)$ and $\lambda_2(u, v)$ are the solutions of the associated algebraic equation in λ , which means of the equation $(c^2 - u^2)\lambda^2 + 2uv\lambda + c^2 - v^2 = 0$ whose roots are real only if $v^2 \geq c^2$ (supersonic flows), then the characteristic strips are given by [69]

$$\begin{aligned} d\varphi &= u dx + v dy, & d\varphi &= u dx + v dy, \\ du + \lambda_2(u, v) dv &= 0, & du + \lambda_1(u, v) dv &= 0, \\ dy - \lambda_1(u, v) dx &= 0, & dy - \lambda_2(u, v) dx &= 0. \end{aligned}$$

By integrating the equations of the second row we are led to the prime integrals $A(u, v) = C_1$ and $B(u, v) = C_2$ which being basically some partial differential equations of first order, could provide a particular class of solutions (integral surfaces) for the Steichen equation.

If one considers the projection of the characteristic strip (corresponding to a given solution φ) on the flow plane Oxy , the respective curves are (called) the *characteristics*. One of the family of characteristics, corresponding to the above particular solutions, is made by straight lines along which $\mathbf{v}(u, v)$ will be constant. But these are the simple wave flows already envisaged in the case of the expansion around a dihedron (Prandtl–Meyer flows), the flows for which the bijectivity between the

physical plane (x, y) and the hodograph plane (u, v) is absent, which means $\frac{D(u,v)}{D(x,y)} = 0$, and which will not be considered in what follows³⁴.

Generally, if φ is an arbitrary solution of the Steichen equation, the projections of the characteristic strips on the plane Oxy will be defined by the last equation of the two groups or, obviously, by the unique equation $(c^2 - u^2) dy^2 + 2uv dx dy + (c^2 - v^2) dx^2 = 0$ (where $u = \frac{\partial \varphi}{\partial x}$ and $v = \frac{\partial \varphi}{\partial y}$). These projections – the characteristic curves (lines) – are real only if $v \geq c$ (or $v \geq c^*$) and they are called *Mach lines*. From the theory of differential equations it is known that the locus of the cuspidal (“returning”) points of the Mach lines is the sonic line $v = c$.

Therefore through every point of the supersonic flow region $c^* < v \leq q_{\max}$, a Mach line from each family is passing and along it the fluid velocity satisfies the equations from the second row. At any point of a Mach line the projections of the fluid particle velocity on the normal direction are equal to the local speed of sound. Really, from

$$(c^2 - u^2) dy^2 + 2uv dx dy + (c^2 - v^2) dx^2 = 0,$$

if $ds = \sqrt{dx^2 + dy^2}$ is the elemental arc along a Mach line we also have that $c^2 = (v \frac{dx}{ds} - u \frac{dy}{ds})^2 = (\mathbf{v} \cdot \mathbf{n})^2$. This result being valid for both characteristics at a certain point, leads to the fact that the direction of the velocity vector (that is the tangent drawn to the streamline at a point) is the bisecting line of the angle made by the Mach lines at that point, an angle which is the double of the Mach angle $\alpha = \text{Arc sin } \frac{c}{v}$.

Any surface (curve) of *weak discontinuity* (that is across it there are no discontinuities for the velocity field but there are discontinuities for the first order derivatives of the velocity components) is compulsory among the characteristic surfaces (curves), an expected result according to the unsolvability of the Cauchy problem in this case.

Consider now a linear or quasilinear system of first order partial differential equations, written under the form $A^0 \mathbf{U}_{,0} + (\mathbf{A} \cdot \nabla) \mathbf{U} + \mathbf{B} = 0$, where $\mathbf{A} = (A^1, A^2, \dots, A^n)$, $\nabla = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n} \right)$, the matrix of unknowns being \mathbf{U} , the matrix (column) of the “free” terms is \mathbf{B} while the

³⁴ Considering a hyperbolic system of the type $U_{,t} + A(U)U_{,x} = 0$ and defining a solution of the simple wave type as a solution of the form $U(\mathbf{r}, t) = U[h(\mathbf{r}, t)]$ – which means the dependence on the Euclidian variables is made by the same function h – Friedrichs has shown, in a famous theorem, that within the class of continuous solutions only a solution of the simple wave type could be joined (it is adjacent) to a constant state (corresponding to the rest or to a uniform flow).

matrices of the system are A^k ,

$$\mathbf{U} = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{pmatrix}, \quad A^k = \begin{pmatrix} a_{11}^k & \cdot & \cdot & \cdot & a_{1n}^k \\ a_{21}^k & \cdot & \cdot & \cdot & a_{2n}^k \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ a_{n1}^k & \cdot & \cdot & \cdot & a_{nn}^k \end{pmatrix}.$$

Obviously, either the coefficients a_{ij}^k or the terms b_i could depend on the independent variables $x_0 = t, x_1, \dots, x_n$ (a linear system) plus, possibly, on the unknowns u_1, u_2, \dots, u_n (a quasilinear system). We will see immediately that the compressible inviscid fluid (Euler) system is of the above form.

Let us now consider a Cauchy condition associated to the above written system, a condition which implies the specification of the solution \mathbf{U} on a hypersurface Σ of equation

$$G_0(x_0, x_1, \dots, x_n) = 0,$$

that is $U|_{\Sigma} = F$, F being a given column vector. Similarly, as in the case of the Steichen equation, the solvability of this problem is connected with the possibility of the evaluation of the higher order derivatives of \mathbf{U} on the surface Σ (that is the possibility of a Taylorian expansion) what is not possible if $\det \left(a_{ij}^k \frac{\partial G_0}{\partial x_k} \right) = 0$ [91], a relation which defines

the characteristic hypersurfaces. In other terms, if $P_{ij} = \sum_{k=0}^m a_{ij}^k \alpha_k$ and $P = \det(P_{ij})$, then P being also a homogeneous polynomial function of n degrees in $\alpha_0, \alpha_1, \dots, \alpha_m$, if this is zero only when at $\alpha_0 = \alpha_1 = \dots = \alpha_m = 0$, the system will be elliptic (it does not have real characteristic hypersurfaces) or if the equation $P = 0$ (in α_0) has n real roots (for any given values for $\alpha_1, \dots, \alpha_m$) the system will be completely hyperbolic.

Finally, a hypersurface Σ is a weak discontinuity surface (when passing across it u_i are continuous while at least one of its derivatives $u_{i,k}$ has a discontinuity of first kind), if and only if $P = \det(P_{ij}) = 0$ [33]. As this represents also the equation of characteristic hypersurfaces we get the above mentioned result.

The theory of weak discontinuity surfaces is very important in fluid mechanics since the perturbations propagate along the discontinuity surfaces. If we accept, for instance, that a uniform stream of velocity \mathbf{v} is perturbed at a fixed point O , then this perturbation will be transported by the fluid and then it propagates with the sound speed c following a direction \mathbf{n} . In the subsonic case, $v < c$, this perturbation may reach any

point from upstream or downstream, there not being real characteristics. In the case $v > c$, the perturbation propagates in a region which is strictly delimited by the real characteristics (Mach lines) which pass through O , thus delimitating a cone with the vertex at O and whose span is the double of the Mach angle. Outside this cone there is no perturbation interference linked to the fixed point O .

If we recall the Euler equations, in an adiabatic regime and in the absence of the mass forces, then considering as independent thermodynamical variables ρ and s , from $p = p(\rho, s)$ we have

$$\frac{\partial p}{\partial x_j} = \frac{\partial p}{\partial \rho} \frac{\partial \rho}{\partial x_j} + \frac{\partial p}{\partial s} \frac{\partial s}{\partial x_j} = c^2 \frac{\partial \rho}{\partial x_j} + \frac{\partial p}{\partial s} \frac{\partial s}{\partial x_j},$$

such that the Euler equations become

$$\begin{aligned} \frac{\partial \rho}{\partial t} + v_i \frac{\partial \rho}{\partial x_i} + \rho \frac{\partial v_i}{\partial x_i} &= 0, \\ \frac{\partial v_j}{\partial t} + v_i \frac{\partial v_j}{\partial x_i} &= -\frac{1}{\rho} \left(c^2 \frac{\partial \rho}{\partial x_j} + \frac{\partial p}{\partial s} \frac{\partial s}{\partial x_j} \right), \\ \frac{\partial s}{\partial t} + v_i \frac{\partial s}{\partial x_i} &= 0. \end{aligned}$$

Considering again the matrices A_1, A_2, A_3 and the vector of the unknown functions \mathbf{U} by

$$\begin{aligned} A^1 &= \begin{bmatrix} v_1 & \rho & 0 & 0 & 0 \\ \frac{c^2}{\rho} & v_1 & 0 & 0 & \frac{1}{\rho} \frac{\partial p}{\partial s} \\ 0 & 0 & v_1 & 0 & 0 \\ 0 & 0 & 0 & v_1 & 0 \\ 0 & 0 & 0 & 0 & v_1 \end{bmatrix}, & A^2 &= \begin{bmatrix} v_2 & 0 & \rho & 0 & 0 \\ 0 & v_2 & 0 & 0 & 0 \\ \frac{c^2}{\rho} & 0 & v_2 & 0 & \frac{1}{\rho} \frac{\partial p}{\partial s} \\ 0 & 0 & 0 & v_2 & 0 \\ 0 & 0 & 0 & 0 & v_2 \end{bmatrix}, \\ A^3 &= \begin{bmatrix} v_3 & 0 & 0 & \rho & 0 \\ 0 & v_3 & 0 & 0 & 0 \\ 0 & 0 & v_3 & 0 & 0 \\ \frac{c^2}{\rho} & 0 & 0 & v_3 & \frac{1}{\rho} \frac{\partial p}{\partial s} \\ 0 & 0 & 0 & 0 & v_3 \end{bmatrix}, & U &= \begin{bmatrix} \rho \\ v_1 \\ v_2 \\ v_3 \\ s \end{bmatrix}, \end{aligned}$$

the above system can be rewritten as $\mathbf{U}_{,t} + (\mathbf{A} \cdot \nabla) \mathbf{U} = 0$, where $\mathbf{A} = (A^1, A^2, A^3)$. Following the result from the above general frame, the characteristic equation $G(x_1, x_2, x_3, t) = 0$ will be given by $P = 0$, where $w = G_{,t}$, $\alpha_i = G_{,i}$, ($\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$) while P is

$$P = \begin{bmatrix} d & \alpha_1 \rho & \alpha_2 \rho & \alpha_3 \rho & 0 \\ \alpha_1 \frac{c^2}{\rho} & d & 0 & 0 & \frac{\alpha_1}{\rho} \frac{\partial p}{\partial s} \\ \alpha_2 \frac{c^2}{\rho} & 0 & d & 0 & \frac{\alpha_2}{\rho} \frac{\partial p}{\partial s} \\ \alpha_3 \frac{c^2}{\rho} & 0 & 0 & d & \frac{\alpha_3}{\rho} \frac{\partial p}{\partial s} \\ 0 & 0 & 0 & 0 & d \end{bmatrix}$$

and where $d = w + \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3$. Developing the determinant using the last row, we have that

$$P = (w + \alpha \cdot \mathbf{v})^3 \left[(w + \alpha \cdot \mathbf{v})^2 - c^2 \alpha^2 \right] = 0$$

which, for any $\alpha_1, \alpha_2, \alpha_3$, has all the roots (in w) real and so the system of the compressible inviscid fluid equations, in adiabatic evolution, is of hyperbolic type.

As regards the possibly discontinuity surfaces, which are among the characteristic surfaces, by denoting the propagation velocity of such a surface with p ($p = -\frac{G_{,t} + \mathbf{v} \text{grad} G}{|\text{grad} G|}$) from the above equation we get

$$-p^3 (p^2 - c^2) = 0,$$

that is $p = 0$ or $p = \pm c$. Meanwhile the surface of velocity $p = 0$ (the entropy wave) is a material surface (which is moving together with the fluid) and along which an entropy discontinuity could occur while the pressure is constant, the surfaces which propagate with the sound speed ($\pm c$) (called the sound waves) will be the loci for pressure discontinuities, the entropy remaining there constant.

12. Direct and Hodograph Methods for the Study of the Compressible Inviscid Fluid Equations

In what follows we will give a brief overview of some of the methods for approaching the “generalized” Cauchy–Riemann system for the steady irrotational plane flows, i.e., the system

$$\frac{\partial \varphi}{\partial x} = \frac{\rho_0}{\rho} \frac{\partial \psi}{\partial y}; \quad \frac{\partial \varphi}{\partial y} = -\frac{\rho_0}{\rho} \frac{\partial \psi}{\partial x},$$

with the classical slip-condition on the surface of the embedded bodies together with the condition at infinity (in the case of the unbounded domains).

The above system is obviously nonlinear since

$$\rho = \rho \left(\left(\frac{\partial \varphi}{\partial x} \right)^2 + \left(\frac{\partial \varphi}{\partial y} \right)^2 \right).$$

Concerning the existence of the solution of this system, a system equivalent with the Steichen equation, C. Morawetz and A. Busemann have proved that, at least in the transonic case, it is ensured provided that one gives up the usual continuity requirements.

Now we will briefly present either a direct method or some hodograph methods to approach the above system. For sake of simplicity we will deal with the subsonic (elliptic) case when any discontinuity surface is avoided.

12.1 **A Direct Method [115]**

The direct method we intend to present briefly in the sequel is important by its possibilities to be used for approaching *other* nonlinear systems too.

Suppose, from the beginning, that the functions φ and ψ are in the form $\varphi = f(u)$ and $\psi = g(v)$, where u and v represent, respectively, the velocity potential and the stream function of the *same* flow but considered incompressible. Using the Cauchy–Riemann system for u and v , we will get

$$\varphi_x = \frac{f_u}{g_v} \psi_y, \psi_y = -\frac{f_u}{g_v} \psi_x$$

where $\frac{f_u}{g_v} = \Phi \left[f_u^2 (\text{grad } u)^2 \right]^{35}$.

Now we will get, using this direct method, the classical solutions of the source and of the (point) vortex in the compressible case.

By imposing that $\psi = g(\theta) = \frac{c}{2\pi}\theta + c'$ (c and c' are constant), we would try to determine a $\varphi = f(\ln r)$ such that the above system is fulfilled. Simple calculations show that this φ should be of the form

$$\varphi = \frac{c}{2\pi} \int v \left(\frac{\rho}{v}\right)'_v dv + c'' \quad (c'' \text{ constant}), \text{ that is } v^2 = \varphi_x^2 + \varphi_y^2 = \left[\frac{f'(\ln r)}{r} \right]^2$$

while $\rho = \frac{f'(\ln r)}{\frac{c}{2\pi}}$ depends on this v^2 . But this solution is just the compressible source. Analogously, if $\varphi = f(\theta) = \frac{c}{2\pi}\theta + c''$ is given, then the corresponding solution ψ of the obvious structure $\psi = g(\ln \frac{1}{r})$, will be necessarily $\psi = \frac{c}{2\pi} \int \frac{dv}{\rho v} + c'$ and $\rho = \frac{\frac{c}{2\pi}}{g'(\ln \frac{1}{r})} = \rho(v^2)$.

The last triplet (φ, ψ, ρ) corresponds to the compressible point vortex.

³⁵ With respect to the explicit form for Φ it is, for instance,

$\Phi[\] = \left\{ \frac{K_1}{K_1 - K_2[\]} \right\}$, $K_1, K_2, \alpha > 0$ (for adiabatic flows) or $\Phi[\] = \sqrt{1 + A[\]}$, $A > 0$ (Chaplygin fluid), etc. Obviously this functional dependence should fulfil the restrictions implied by its significance, namely $\Phi > 0$, $\Phi(v=0) = 1$, $\frac{d\Phi}{dv^2} > 0$, $\Phi < \Phi(c^{*2})$ (c^* being the critical velocity).

Let us now extend the above procedure by considering either the pair of functions $A(x, y), B(x, y) : D_1 \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $A, B \in C^2(D_1)$, $\frac{D(A,B)}{D(x,y)} \neq 0$ or the nonholomorphic function $g(z) = A(x, y) + iB(x, y)$, $g : D_1 \rightarrow D_2$. If we introduce also the function $F : D_2 \rightarrow \mathbb{C}$, $F = U(A, B) + iV(A, B)$, a holomorphic function in D_2 , it is obvious that the composed function $F(g(z)) \equiv f(z) = \varphi(x, y) + i\psi(x, y)$ will be nonholomorphic in D_1 . So we have in our hands a pair of complex functions g and F with the above mentioned properties, which should be formulated such that their composition satisfies the focussed system and, more, the functional dependence $\rho = \rho(v^2)$ is ensured. Basically all these lead, through the Cauchy–Riemann system which is satisfied by U and V , to the fulfilment of the condition

$$\frac{1}{\Delta} \left(A_x^2 + B_x^2 - \delta \frac{\psi_x}{\psi_y} \right) = \frac{1}{\Delta} \left(A_y^2 + B_y^2 - \delta \frac{\psi_y}{\psi_x} \right) \equiv \rho,$$

where $\Delta = \frac{D(A,B)}{D(x,y)}$ and $\delta = A_x A_y + B_x B_y$, each side of this equality depending on $\varphi_x^2 + \varphi_y^2$.

In the particular case of a subsonic stream past a circular obstacle with a velocity at far field $(v_\infty, 0)$, by accepting the adiabatic law $\rho = (1 - kv^2)^{-\alpha}$ and choosing $F(A, B) = -\arctg \frac{A}{B} + i \ln \sqrt{A^2 + B^2}$, the above system leads, by an approximate solving, to a solution which has been already established through the Imai–Lamla method but which now satisfies exactly the boundary conditions [115].

12.2 Chaplygin Hodograph Method. Molenbroek–Chaplygin equation

The hodograph (plane) method, as in the incompressible case, leads us to a study of the flow in the “hodograph” plane (u, v) and, consequently, the independent variables x and y are replaced by u and v or V and θ (the velocity polar coordinates) while φ and ψ should be expressed with these new coordinates³⁶. It is also possible to try, conversely, to express x, y, V and θ as functions of φ and ψ , considered now independent variables, which has the advantage of knowing, in general, the variation domain for the point (φ, ψ) of the plane $O\varphi\psi$ while the corresponding domain from the hodograph plane is not known yet.

We remark that if we make the change of variable defined by $u = \frac{\partial \varphi}{\partial x}$ and $v = \frac{\partial \varphi}{\partial y}$ together with the change of function $\Phi = ux + vy - \varphi$,

³⁶Details on the “hodograph” plane techniques can be found, for instance, in [69].

the Steichen equation will transform into the following linear partial differential equation (Prandtl equation)

$$\left(1 - \frac{v^2}{c^2}\right) \frac{\partial^2 \Phi}{\partial u^2} + 2 \frac{uv}{c^2} \frac{\partial^2 \Phi}{\partial u \partial v} + \left(1 - \frac{u^2}{c^2}\right) \frac{\partial^2 \Phi}{\partial v^2} = 0,$$

to which one could apply the classical methods of integration (Riemann). The inconvenience of such change of variable and function consists in the lack of a simple mechanical interpretation for Φ , while φ and ψ have several interpretations.

If we keep only the passage to the hodograph plane, by setting $z = x + iy$ we have $dz = dx + idy$ and $d\varphi = udx + vdy$, $d\psi = -\frac{\rho}{\rho_0}vdx + \frac{\rho}{\rho_0}udy$ and from here, by eliminating dx and dy and replacing $u = V \cos \theta$ and $v = V \sin \theta$, we get

$$dz = \frac{e^{i\theta}}{V} \left(d\varphi + i \frac{\rho_0}{\rho} d\psi \right).$$

Imposing that the right side should be a total (exact) differential and separating then the real and imaginary parts, we obtain the system

$$\frac{\partial \theta}{\partial \varphi} = \frac{\rho}{\rho_0 V} \frac{\partial V}{\partial \psi}, \quad \frac{\partial \theta}{\partial \psi} = V \left(\frac{\rho_0}{\rho V} \right)'_V \frac{\partial V}{\partial \varphi}$$

which, by “inversion”, could be written (Chaplygin)

$$\frac{\partial \varphi}{\partial \theta} = \frac{\rho_0 V}{\rho} \frac{\partial \psi}{\partial V}, \quad \frac{\partial \varphi}{\partial V} = V \left(\frac{\rho_0}{\rho V} \right)'_V \frac{\partial \psi}{\partial \theta} \equiv -\frac{\rho_0}{\rho V} \left(1 - \frac{V^2}{c^2} \right) \frac{\partial \psi}{\partial \theta}.$$

If we manage to solve this system, we will have $\varphi(v, \theta)$ and $\psi(v, \theta)$, defined in a domain of the hodograph plane contained in the disk $u^2 + v^2 = V_{\max}^2$ ³⁷. From the “connection” formulas $dz = \frac{e^{i\theta}}{V} \left(d\varphi + i \frac{\rho_0}{\rho} d\psi \right)$, by integrating, we can obtain $x(V, \theta)$ and $y(V, \theta)$, that is $x(u, v)$ and $y(u, v)$, and therefore, by inversion (the condition $\frac{D(x,y)}{D(u,v)} \neq 0$ making this possible), one finally gets $u(x, y)$ and $v(x, y)$.

Suppose now that, from the last two equations of the system, we have eliminated φ , thus obtaining the so-called Molenbroek–Chaplygin equation

³⁷ We denote the magnitude of the maximum velocity $q_m \equiv V_{\max}$ while the critical velocity $c^* = V^*$.

$$\frac{\partial}{\partial V} \left(\frac{\rho_0 v}{\rho} \frac{\partial \psi}{\partial V} \right) + \frac{\rho_0}{\rho V} \left(1 - \frac{V^2}{c^2} \right) \frac{\partial^2 \psi}{\partial \theta^2} = 0,$$

an equation which could be rewritten, in an equivalent form

$$V^2 \frac{\partial^2 \psi}{\partial V^2} + \left(1 - \frac{V^2}{c^2} \right) \frac{\partial^2 \psi}{\partial \theta^2} + V \left(1 + \frac{V^2}{c^2} \right) \frac{\partial \psi}{\partial V} = 0.$$

The last form is a linear elliptic or hyperbolic equation, according to $V < c$ or $V > c$, and whose characteristics in the hodograph plane will not depend on $\varphi(V, \theta)$ or $\psi(V, \theta)$. These characteristics called also *hodograph characteristics*, will be defined by the equation $d\theta^2 + \frac{1}{V^2} \left(1 - \frac{v^2}{c^2} \right) dV^2 = 0$.

It is shown that these hodograph characteristics, in fact the characteristics of the Prandtl equation in the coordinates V and θ , have perpendicular directions (tangents) vis-a-vis the Mach lines of the other family from the physical plane.

Before ending this last section of Chapter 2, we intend to present, briefly, other useful forms of the Chaplygin system or of the Molenbroek–Chaplygin equation.

If, for instance, in the plane of the variable V , we introduce $r = \int \frac{\sqrt{1-M^2}}{V} dV$, we obtain $\frac{\partial \varphi}{\partial \theta} = k(r) \frac{\partial \psi}{\partial r}$, $\frac{\partial \varphi}{\partial r} = -k(r) \frac{\partial \psi}{\partial \theta}$ with $k(r) = \frac{\rho_0}{\rho} \sqrt{1-M^2}$ and the Molenbroek–Chaplygin equation becomes $\frac{\partial^2 \psi}{\partial r^2} + \frac{\partial^2 \psi}{\partial \theta^2} + \left(\frac{d}{dr} \ln r \right) \frac{\partial \psi}{\partial r} = 0$.

If in the place of V , we consider now the variable $\sigma = - \int_{V^*}^V \frac{\rho}{\rho_0} \frac{dV}{V}$, ρ being a function of V , then the Molenbroek–Chaplygin equation gives us $\frac{\partial^2 \psi}{\partial \sigma^2} + k(\sigma) \frac{\partial^2 \psi}{\partial \theta^2} = 0$ with $k(\sigma) \equiv (1 - M^2) \frac{\rho_0^2}{\rho^2}$, which is used specially in the transonic flows. The case $k(\sigma) = \sigma$ corresponds to the Tricomi equation.

Finally, in the adiabatic case, by introducing the nondimensional variable τ and the constant β so that $\tau = \frac{\gamma-1}{2} \frac{v^2}{c_0^2} = \frac{V^2}{V_{\max}^2}$ and $\beta = \frac{1}{\gamma-1}$, to the interval of variation $0 \leq V \leq V_{\max}$ corresponds the interval $0 \leq \tau \leq 1$ while to the critical value V^* of the velocity corresponds $\tau^* = \frac{1}{2\beta+1}$. We also have $\rho = \rho_0 (1 - \tau)^\beta$, $p = p_0 (1 - \tau)^{\beta+1}$, $c^2 = c_0^2 (1 - \tau)$ and the Chaplygin system and the Molenbroek–Chaplygin equation become respectively,

$$\frac{\partial \varphi}{\partial \theta} = \frac{2\tau}{(1 - \tau)^\beta} \frac{\partial \psi}{\partial \tau}; \quad \frac{\partial \varphi}{\partial \tau} = - \frac{1 - (2\beta + 1)\tau}{2\tau(1 - \tau)^{\beta+1}} \frac{\partial \psi}{\partial \theta},$$

and

$$\frac{\partial}{\partial \tau} \left(\frac{2\tau}{(1-\tau)^\beta} \frac{\partial \psi}{\partial \tau} \right) + \frac{1 - (2\beta + 1)\tau}{2\tau(1-\tau)^{\beta+1}} \frac{\partial^2 \varphi}{\partial \theta^2} = 0.$$

Using the method of separating variables, Chaplygin has succeeded in obtaining the exact general solution for the above equations by means of the hypergeometric series [69].

Chapter 3

VISCOUS INCOMPRESSIBLE FLUID DYNAMICS

In what follows we will give a short survey on some features related to the viscous incompressible fluid flows and their equations (Navier–Stokes), all considered within the context of building of some numerical algorithms to approach these flows. Thus, after a brief overview of some uniqueness and existence results, we will focus on different formulations used for Navier–Stokes equations. A special role will be played by the so-called integral conditions for the rotation which replaces the non-existence of a “classical” boundary condition.

Aspects connected with the nondimensionalization of the involved equations, followed by some approximate models in the case of small, respectively great, Reynolds number, are then envisaged. From the large variety of approaches to the important concept of boundary layer, we will chose the probabilistic way which, apart from a higher rigor, is a source of efficient numerical algorithms.

Everywhere in this chapter the laminar character of the flow is accepted.

1. The Equation of Vorticity (Rotation) and the Circulation Variation

We have seen that for a viscous incompressible fluid, the stress tensor is given by the constitutive law $[\mathbf{T}] = -p[\mathbf{I}] + 2\mu[\mathbf{D}]$, that is $[\boldsymbol{\sigma}] = 2\mu[\mathbf{D}]$.

We suppose, in the sequel, that the viscosity coefficient μ is constant (by accepting the Stokes hypothesis $3\lambda + 2\mu = 0$, λ should be constant as well). Since

$$2\bar{\mu} \operatorname{div}[\mathbf{D}] = \mu \left[\operatorname{div}(\operatorname{grad} \mathbf{v}) + \operatorname{div}(\operatorname{grad} \mathbf{v})^T \right]$$

$$= \mu [\operatorname{div}(\operatorname{grad} \mathbf{v}) + \operatorname{grad}(\operatorname{div} \mathbf{v})] = \mu \nabla^2 \mathbf{v} = -\mu \operatorname{rot} \boldsymbol{\omega},$$

by introducing also the kinematic viscosity coefficient $\nu = \frac{\mu}{\rho}$, the equations which govern the fluid flow (more precisely, the equations of linear momentum) could be rewritten, as we have previously seen, in one of the following equivalent forms:

$$\frac{\partial \mathbf{v}}{\partial t} + \operatorname{div}(\mathbf{v} \otimes \mathbf{v}) = -\frac{1}{\rho} \operatorname{grad} p + \nu \nabla^2 \mathbf{v} + \mathbf{f},$$

or

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \operatorname{grad}) \mathbf{v} = -\frac{1}{\rho} \operatorname{grad} p + \nu \nabla^2 \mathbf{v} + \mathbf{f},$$

or

$$\frac{\partial \mathbf{v}}{\partial t} + \operatorname{grad} \left(\frac{1}{2} v^2 \right) + \boldsymbol{\omega} \times \mathbf{v} = -\frac{1}{\rho} \operatorname{grad} p + \nu \nabla^2 \mathbf{v} + \mathbf{f}.$$

These equations of mixed type, are known also as the *Navier–Stokes equations*. Obviously, in order to define precisely the whole pattern of the flow, they should be completed by the equation of continuity and the equation of energy together with some initial and boundary (adherence or no-slip conditions) plus, eventually (in the case of unbounded domains), the behaviour conditions at far field (infinity).

In what follows we will search new formulations for the Navier–Stokes system or even different “approximations” for it in order to solve some practical problems.

Before doing that we need some results about the vorticity (rotation) and circulation.

For a viscous compressible fluid, by applying the operator *rot* to both sides of the flow equation under the Helmholtz form, which means to the equation

$$\rho \left[\frac{\partial \mathbf{v}}{\partial t} + \operatorname{grad} \left(\frac{1}{2} v^2 \right) + \boldsymbol{\omega} \times \mathbf{v} \right] = \rho \mathbf{f} - \operatorname{grad} p + \operatorname{div}[\boldsymbol{\sigma}],$$

in the hypothesis that the external forces come from a potential U , that is $\mathbf{f} = -\operatorname{grad} U$, we get

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \operatorname{rot}(\boldsymbol{\omega} \times \mathbf{v}) = \operatorname{rot} \left(-\frac{\operatorname{grad} p}{\rho} + \frac{1}{\rho} \operatorname{div}[\boldsymbol{\sigma}] \right).$$

However, according to Appendix A,

$$\text{rot}(\boldsymbol{\omega} \times \mathbf{v}) = (\mathbf{v} \cdot \text{grad}) \boldsymbol{\omega} - (\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v} + \boldsymbol{\omega} (\text{div} \mathbf{v}) - \mathbf{v} (\text{div} \boldsymbol{\omega})$$

and $\text{div} \boldsymbol{\omega} = 0$, so we will also have

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + (\mathbf{v} \cdot \text{grad}) \boldsymbol{\omega} = (\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v} - \boldsymbol{\omega} (\text{div} \mathbf{v}) + \text{rot} \left(-\frac{\text{grad} p}{\rho} + \frac{1}{\rho} \text{div}[\boldsymbol{\sigma}] \right),$$

i.e., the rate of change of vorticity for an observer who is moving with the fluid is

$$\frac{D\boldsymbol{\omega}}{Dt} = (\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v} - \boldsymbol{\omega} (\text{div} \mathbf{v}) + \text{rot} \left(-\frac{\text{grad} p}{\rho} + \frac{1}{\rho} \text{div}[\boldsymbol{\sigma}] \right),$$

$\text{div}[\boldsymbol{\sigma}]$ having the expression already formulated within the study of viscous compressible fluid flows.

On the other hand, we know that the circulation along a closed fluid contour C , is defined by $\Gamma = \int_C \mathbf{v} \cdot d\mathbf{r}$ and $\frac{D\Gamma}{Dt} = \int_C \mathbf{a} \cdot d\mathbf{r}$. But since

$\mathbf{a} = \mathbf{f} - \frac{\text{grad} p}{\rho} + \frac{\text{div}[\boldsymbol{\sigma}]}{\rho}$ and \mathbf{f} comes from the potential U we obtain

$$\frac{D\Gamma}{Dt} = \int_C \left(-\frac{\text{grad} p}{\rho} + \frac{1}{\rho} \text{div}[\boldsymbol{\sigma}] \right) \cdot d\mathbf{r},$$

which provides the rate of change of circulation for the considered fluid flow.

Obviously, in the conditions of a viscous incompressible flow ($\text{div} \mathbf{v} = 0$) and under the same hypothesis on the conservative character of the external forces ($\mathbf{f} = -\text{grad}U$), by applying again the operator rot to both sides of the Navier–Stokes system, that is to

$$\frac{\partial \mathbf{v}}{\partial t} + \text{grad} \left(\frac{1}{2} v^2 \right) + \boldsymbol{\omega} \times \mathbf{v} = -\frac{1}{\rho} \text{grad} p + \nu \nabla^2 \mathbf{v} + \mathbf{f},$$

we get

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \text{rot}(\boldsymbol{\omega} \times \mathbf{v}) = \nu \nabla^2 \boldsymbol{\omega}.$$

As $\text{rot}(\boldsymbol{\omega} \times \mathbf{v})$ is given this time ($\text{div} \mathbf{v} = 0$) by

$$(\mathbf{v} \cdot \text{grad}) \boldsymbol{\omega} - (\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v}$$

and

$$(\boldsymbol{\omega} \cdot \text{grad}) \mathbf{v} = (\text{grad } \mathbf{v}) \boldsymbol{\omega} = ([\mathbf{D}] + [\boldsymbol{\Omega}]) \boldsymbol{\omega} = [\mathbf{D}] \boldsymbol{\omega},$$

we finally have

$$\frac{D\boldsymbol{\omega}}{Dt} = [\mathbf{D}] \boldsymbol{\omega} + \nu \nabla^2 \boldsymbol{\omega}.$$

We remark that the vorticity changes due to the term $[\mathbf{D}] \boldsymbol{\omega}$ are related to either the “stretching” of the vortex line or to the “angular turning” of the vortex line. In the plane case these aspects of stretching or turning are completely absent ($[\mathbf{D}] \boldsymbol{\omega} = 0$) and the vorticity equation is simply

$$\frac{D\boldsymbol{\omega}}{Dt} = \nu \nabla^2 \boldsymbol{\omega}.$$

The above equations, which have been established assuming incompressibility, will have the same structure even in the case of barotropic compressibility when there is a function $h(\rho)$ such that

$$\text{grad } h = \frac{\text{grad } p}{\rho}.$$

The same equations anticipate vorticity conservation in the plane case, which will not be true in the three-dimensional case. This remark would back support the non-existence of some general uniqueness and existence results (with the continuous dependence on data) for the three-dimensional Navier–Stokes equations when only some local results, that is for small intervals of time, exist.

2. Some Existence and Uniqueness Results

The Navier–Stokes equations (the equations of viscous incompressible fluid flows) have had the attention of many mathematicians who have approached them in their study of the mathematical coherence of the corresponding model, i.e., the search for the existence and uniqueness of the solution which depends continuously on data.

In a famous paper published in 1933 [82], J. Leray established the existence of the steady state solution (but not its uniqueness) in a bounded domain Ω , for the Navier–Stokes system by using an “a priori” assessment of the Dirichlet integral in the form $\int_{\Omega} (\text{grad } \mathbf{v})^2 \leq M$, where M depends on Ω , the Reynolds number and the data of the problem (the external mass forces and the transport velocity of the domain boundary). In the same paper Leray investigated also the case when Ω is an external unbounded domain (the complement of a compact set) by completing the Navier–Stokes equations with a condition of the type $\lim_{\mathbf{r} \rightarrow \infty} \mathbf{v}(\mathbf{r}) = \mathbf{v}_{\infty}$ (i.e., a far field condition).

Although in the three-dimensional case the respective behaviour condition is satisfied (Finn [45]), this will be not always fulfilled in the plane case so that the problem of the mentioned Leray solutions is still open.

In the particular case when $\mathbf{v}_\infty \neq 0$ and the Reynolds number is sufficiently small, G.P.Galdi has given an existence and uniqueness result within some suitable function spaces. The same author has established an existence and uniqueness result for the Oseen problem [48]. Concerning the Stokes problem for an exterior domain whose boundary is Lipschitzian, Galdi and Simander have proved, in $L^2(\Omega)$, the existence and the uniqueness of a solution which depends continuously on data [50].

As regards the Cauchy problem for the unsteady Navier–Stokes system, the existence and the uniqueness of the classical solution has been established in both plane and axially symmetric cases while in space the existence has been proved only locally, i.e., for limited time intervals and for sufficiently small Cauchy data (in a suitable topology) [77], E. Hopf pointing out that this problem is not “well-posed” [66]. The same E. Hopf has also proved the existence of weak solutions for the Navier–Stokes equations [67].

An overview of the existence and uniqueness results has been made by R.K.Zeytonian [159] and more recently by P.L. Lions [85].

In the sequel we will touch upon the some uniqueness results of the classical solution which, as we have pointed out in the case of the inviscid fluids, are of the greatest practical interest.

Thus, in the conditions of the domains which are bounded by surfaces made by a finite number of closed boundaries of rigid bodies (possibly in motion), a Dirichlet–Cauchy condition for the Navier–Stokes equations (i.e., the adherence condition together with an initial condition for velocity) has a unique (classical) solution in quite non-restrictive hypotheses (Foa, [47]). D. Graffi and J.Serrin have extended this result to the case of the compressible fluids too [57], [135]. At the same time, following a procedure given by Rionero and Maiellaro for the inviscid fluids, the uniqueness of the classical solution is also established under the assumptions of the boundedness at infinity of the velocity gradients [130].

Concerning the unbounded domains (the exterior of a closed and bounded surface), a situation which often occurs within practical problems, Dario Graffi has shown the uniqueness of the solution for a Dirichlet–Cauchy problem provided that the velocity and pressure fields are continuous and bounded with respect to the spatial variables and the time, while the velocity second order derivatives are continuous a.e. with re-

spect to the same variables and, at far distances, the pressure p behaves as $p - p_0 = o\left(\frac{1}{r}\right)$.

Some extensions of this result, for the case of compressible fluids, may be found in D. Graffi [59], S. Rionero and P. Galdi [129].

Obviously if a classical solution initially exists (that is on a small interval of time, starting from t_0) and it is steady, then, if this solution will not be sufficiently smooth at an ulterior moment t (which means, basically, it will not exist) the uniqueness will collapse.

3. **The Stokes System**

Let D be a plane or spatial region with a fixed smooth boundary ∂D and \mathbf{w} a vectorial field defined on D . It is known that such a vectorial field \mathbf{w} could be uniquely decomposed into the sum $\mathbf{u} + \text{grad } p = \mathbf{w}$, where \mathbf{u} is a vector satisfying $\text{div } \mathbf{u} = 0$ (solenoidal) being also “parallel” to the boundary ∂D , that is $\mathbf{u} \cdot \mathbf{n}|_{\partial D} = 0$, while p is a scalar (defined up to an additive constant) [19].

Due to this result we may define the operator P , called the *orthogonal projection operator*, which maps every vector \mathbf{w} into the vector \mathbf{u} , i.e., into its part of zero divergence which is also “parallel” to the boundary. According to the above result this operator P is well-defined.

We notice that P is, by construction, a linear operator satisfying the equality $\mathbf{w} = P\mathbf{w} + \text{grad } p$, whose *fixed points* are the vectors \mathbf{u} fulfilling $\text{div } \mathbf{u} = 0$, $\mathbf{u} \cdot \mathbf{n}|_{\partial D} = 0$ and, of course, $P\mathbf{u} = \mathbf{u}$ while its *zeros* are the vectors $\text{grad } p$ because, obviously, $P(\text{grad } p) = 0$.

Let us consider the Navier–Stokes system, under the assumptions of the external (mass) forces absence or of their derivation from a potential U , and let us apply to this system the operator P . As

$$P\left(\frac{1}{\rho}\text{grad } p\right) = P(\text{grad } U) = 0$$

we have

$$P\left(\frac{\partial \mathbf{v}}{\partial t}\right) = P\left(-(\mathbf{v} \cdot \nabla)\mathbf{v} + \nu \nabla^2 \mathbf{v}\right).$$

But if \mathbf{v} satisfies the ircompressibility condition ($\text{div } \mathbf{v} = 0$) and the necessary condition on the fixed boundary ∂D ($\mathbf{v} \cdot \mathbf{n}|_{\partial D} = 0$) as well, the same result does hold for $\frac{\partial \mathbf{v}}{\partial t}$ and it does not for $\nabla^2 \mathbf{v}$ (this fulfils $\text{div}(\nabla^2 \mathbf{v}) = 0$ but, in general, $\nabla^2 \mathbf{v} \cdot \mathbf{n}|_{\partial D} \neq 0$). With this remark we are led to the following equation of evolution type (an important feature which allows the construction of numerical temporal algorithms)

$$\frac{\partial \mathbf{v}}{\partial t} = P \left(-(\mathbf{v} \cdot \nabla) \mathbf{v} + \frac{1}{R} \nabla^2 \mathbf{v} \right),$$

where $\frac{1}{R} = \nu$, R being the so-called Reynolds number (generally $R = \frac{V_c L_c}{\nu}$, V_c and L_c being the characteristic (reference) velocity and length respectively or, in other terms, it is the ratio between the weight of the inertial forces and that of the viscosity forces).

The importance of this equation consists first in the pressure elimination, the pressure being then constructed “a posteriori” as the “gradient” part of

$$-(\mathbf{v} \cdot \nabla) \mathbf{v} + \frac{1}{R} \nabla^2 \mathbf{v}.$$

Further, this consequence of the Navier–Stokes equation is of a great importance in elaborating on a class of numerical algorithms¹.

If R is small (the case of the slow flows or the very viscous fluids, etc.) the right side of the above equation could be approximated by

$$\frac{\partial \mathbf{v}}{\partial t} = P \left(\frac{1}{R} \nabla^2 \mathbf{v} \right)$$

and hence we have the approximate system

$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} + \text{grad } p = \frac{1}{R} \nabla^2 \mathbf{v} \\ \text{div } \mathbf{v} = 0, \mathbf{v} \cdot \mathbf{n}|_{\partial D} = 0. \end{cases}$$

This system which represents a good approximation of the Navier–Stokes equations (in the above mentioned hypotheses) is of parabolic type and it is called *the Stokes system*.

The Stokes system is a first (classical) linearized form of the viscous fluid equations. In fact, to the equations of this system one associates corresponding adherence (no-slip) conditions $\mathbf{v}|_{\partial D} = 0$ and initial conditions under the form $\mathbf{v}|_{t=t_0} = \mathbf{h}$ and $\frac{\partial \mathbf{v}}{\partial t}|_{t=t_0} = \mathbf{g}$ as well.

Applying the divergence (*div*) operator to both sides of the previous system we get $\Delta p = 0$ in D , that is, within the Stokes model, the pressure is a harmonic function. If the flow is steady we will have that

¹In fact, except the incompressible case, all the unsteady flow equations for both viscous and inviscid fluid are of evolution type. Even in the incompressible case, one could restore this evolution character by introducing an “artificial compressibility” which later tends to zero. For instance, the equation of continuity becomes $\varepsilon \frac{\partial p}{\partial t} + \text{div } \mathbf{v} = 0$, with ε a small parameter which ultimately is obliged to tend to zero.

the fluid velocity is a biharmonic function ($\nabla^4 \mathbf{v} = 0$) while the vorticity $\boldsymbol{\omega} = \text{rot } \mathbf{v}$ is also a harmonic function ($\nabla^2 \boldsymbol{\omega} = 0$).

In this book we will come back to the Stokes system within the context of certain applications to practical problems.

We cannot finalize this section without pointing out what is known as *Stokes paradox*. Basically this paradox shows up that, in the conditions of a plane steady uniform (at far field) flow around a circular cylinder, the Stokes model fails².

The failure of the approximation at far distances through the Stokes model (in fact there is not a valid uniform approximation of the exact equations), leads to the consideration of some nonlinear effects within the Stokes equations. Some details on this new approach which leads to the so-called *Oseen model*, can be found in the sequel and, for instance, in [98].

4. **Equivalent Formulations for the Navier–Stokes Equations in Primitive Variables**

There are two main distinct ways to proceed in the construction of some equivalent formulations for the Navier-Stokes equations, both being of great use in the numerical approach to these equations.

The first is the pressure-velocity or (only) pressure formulation, known also as the formulation in “primitive” (“genuine”) variables. The second is the vorticity-potential or stream function formulation (with its variants) known as the formulation in “non-primitive” variables. In the sequel we will give a brief survey on the most important features of both formulations, focussing on some recent results about the integral conditions for vorticity which interfere within the formulation in “non-primitive” variables.

4.1 **Pressure Formulation**

In what follows we will envisage an equivalent formulation of the Navier–Stokes system which allows evaluation of the pressure as a function of velocity field. For this we first consider the Navier–Stokes equations under the form

$$\frac{\partial \mathbf{v}}{\partial t} + \text{div} (\mathbf{v} \otimes \mathbf{v}) = -\frac{1}{\rho} \text{grad } p + \nu \nabla^2 \mathbf{v} + \mathbf{f},$$

to which one applies the divergence operator. Using then the formulae (see Appendix A)

²The first rigorous proof of the Stokes paradox can be found in the first edition of the Kocin, Kibel, Rose book [74].

$$\operatorname{div} (\mathbf{v} \otimes \mathbf{v}) = (\operatorname{grad} \mathbf{v}) \mathbf{v} + \mathbf{v} (\operatorname{div} \mathbf{v})$$

and

$$\operatorname{div} ([\mathbf{A}]\mathbf{v}) = (\operatorname{div}[\mathbf{A}]^T) \cdot \mathbf{v} + [\mathbf{A}]^T \cdot (\operatorname{grad} \mathbf{v}),$$

we also have

$$\begin{aligned} \operatorname{div} [\operatorname{div} (\mathbf{v} \otimes \mathbf{v})] &= \operatorname{div} (\operatorname{grad} \mathbf{v}^T) \cdot \mathbf{v} + (\operatorname{grad} \mathbf{v})^T \cdot \operatorname{grad} \mathbf{v} \\ &+ (\operatorname{div} \mathbf{v})^2 + \mathbf{v} \cdot \operatorname{grad} (\operatorname{div} \mathbf{v}) = \operatorname{grad} (\operatorname{div} \mathbf{v}) \cdot \mathbf{v} + (\operatorname{grad} \mathbf{v})^T \cdot (\operatorname{grad} \mathbf{v}) \\ &+ (\operatorname{div} \mathbf{v})^2 + \operatorname{grad} (\operatorname{div} \mathbf{v}) \cdot \mathbf{v} \end{aligned}$$

and consequently

$$-\frac{1}{\rho} \nabla^2 p = \frac{\partial \alpha}{\partial t} + \alpha^2 + 2 (\operatorname{grad} \alpha) \cdot \mathbf{v} + (\operatorname{grad} \mathbf{v})^T \cdot \operatorname{grad} \mathbf{v} - \nu \nabla^2 \alpha - \operatorname{div} \mathbf{f},$$

where $\alpha = \operatorname{div} \mathbf{v}$.

Using now the decomposition of the gradient tensor, that is $\operatorname{grad} \mathbf{v} = [\mathbf{D}] + [\mathbf{\Omega}]$, where $[\mathbf{D}]$ is the symmetric rate-of-strain tensor ($D_{ij} = \frac{1}{2}(v_{i,j} + v_{j,i})$) and $[\mathbf{\Omega}]$ is the skew-symmetric rotation tensor ($\Omega_{ij} = \frac{1}{2}(v_{i,j} - v_{j,i})$), we may check by direct calculations, that

$$[\mathbf{D}] \cdot [\mathbf{D}] = [\mathbf{D}] \cdot \operatorname{grad} \mathbf{v}$$

and $\frac{1}{2}\omega^2 = [\mathbf{\Omega}] \cdot \operatorname{grad} \mathbf{v}$ so that $(\operatorname{grad} \mathbf{v})^T \cdot (\operatorname{grad} \mathbf{v}) = [\mathbf{D}] \cdot [\mathbf{D}] - \frac{1}{2}\omega^2$.

By introducing now Truesdell's number for vorticity M_T , defined through $M_T = |\boldsymbol{\omega}| / (2[\mathbf{D}] \cdot [\mathbf{D}])^{\frac{1}{2}}$ (and which is seen as a measure for the fluid vorticity), the above equation could be also rewritten

$$-\frac{1}{\rho} \nabla^2 p = \frac{\partial \alpha}{\partial t} + 2 (\operatorname{grad} \alpha) \cdot \mathbf{v} + (1 - M_T^2) [\mathbf{D}] \cdot [\mathbf{D}] - \nu \nabla^2 \alpha + \alpha^2 - \operatorname{div} \mathbf{f}.$$

As $\alpha = 0$ together with the incompressibility assumption, we get the following equation for the pressure determining

$$-\frac{1}{\rho} \nabla^2 p = (1 - M_T^2) [\mathbf{D}] \cdot [\mathbf{D}] - \operatorname{div} \mathbf{f},$$

an equation to which one should join the appropriate boundary conditions. We remark that $M_T = 0$ for the irrotational flow while $M_T = \infty$ for the rigid bodies ($[\mathbf{D}] = 0$), so that $0 \leq M_T < \infty$.

4.2 Pressure-Velocity Formulation

The Navier–Stokes equations, in the absence of external (body) forces, written in the form

$$\frac{\partial \mathbf{v}}{\partial t} + \operatorname{div}(\mathbf{v} \otimes \mathbf{v}) = -\frac{1}{\rho} \operatorname{grad} p + \nu \nabla^2 \mathbf{v},$$

$$\operatorname{div} \mathbf{v} = 0,$$

will be completed in what follows by the equation of internal energy

$$\rho \frac{De}{Dt} = \Phi + \chi \nabla^2 T$$

where $\Phi = [\mathbf{T}] \cdot [\mathbf{D}] = 2\mu[\mathbf{D}] \cdot [\mathbf{D}]$ is the so-called *dissipation* (function) which measures the rate of work done by the “viscous part” of the stresses during the deformation process of a unit volume of fluid in order to increase the internal energy and hence the temperature of the fluid. Since Φ should be negative, from its explicit structure $\Phi = \mu \frac{\partial v_i}{\partial x_j} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right)$, it turns out that $3\lambda + 2\mu \geq 0$ and $\mu \geq 0$ which is obviously satisfied.

On the other side, ρ being constant for the incompressible fluid, the thermodynamics equations lead to $\frac{De}{Dt} = C$ with C the specific heat. Hence the internal energy equation becomes

$$\frac{DT}{Dt} = \frac{\Phi}{\rho C} + k \nabla^2 T,$$

where $k = \frac{\chi}{\rho C}$ is the thermal diffusion.

From the pressure equation (see *pressure formulation*) where $\alpha = \operatorname{div} \mathbf{v} = 0$, we now have

$$\nabla^2 p = -\rho (\operatorname{grad} \mathbf{v})^T \cdot (\operatorname{grad} \mathbf{v}),$$

an equation which should be (numerically) solved simultaneously with the flow and continuity equations of the Navier–Stokes system. The use of the no-slip condition on a solid fixed surface $\mathbf{v}|_{\partial D} = 0$ in the flow equation, yields³

$$\operatorname{grad} p|_{\partial D} = \mu \Delta \mathbf{v}|_{\partial D} = -\mu \operatorname{rot} \boldsymbol{\omega}|_{\partial D}$$

or, by taking the dot product with \mathbf{n} , the unit outward normal drawn to ∂D we get

³We have used the vector identity $\nabla^2 \mathbf{v} = \operatorname{grad}(\operatorname{div} \mathbf{v}) - \operatorname{rot} \boldsymbol{\omega}$ with $\boldsymbol{\omega} = \operatorname{rot} \mathbf{v}$.

$$\left. \frac{dp}{dn} \right|_{\partial D} = -\mu \mathbf{n} \cdot (\text{rot } \boldsymbol{\omega})|_{\partial D} .$$

Thus we have to solve a Neumann problem for the Poisson equation of the pressure, a problem which creates much inconvenience due to the nonlinear character (in velocity derivatives) of the boundary condition. To overcome most of these shortcomings it is recommended, for instance, the use of $\alpha = 0$ everywhere except for the evolution term $\rho \frac{\partial \alpha}{\partial t}$, that is to replace the above pressure equation by

$$\nabla^2 p = -\rho (\text{grad } \mathbf{v})^T \cdot (\text{grad } \mathbf{v}) - \frac{\partial \alpha}{\partial t} \rho$$

or, equivalently, by

$$\nabla^2 p = -\rho (1 - M_T^2) [\mathbf{D}] \cdot [\mathbf{D}] - \frac{\partial \alpha}{\partial t} \rho,$$

where M_T is Truesdell's number.

These equations should be solved (at time steps) simultaneously with the flow equation, the pressure for $\alpha = 0$ being taken as the "right" pressure.

Chorin has suggested another method which avoids completely the pressure equation. Replacing the equation of continuity $\alpha = \text{div } \mathbf{v} = 0$ by the equation

$$\beta \frac{\partial p}{\partial t} + \text{div } \mathbf{v} = 0,$$

where β is an artificial compressibility and $p = \frac{1}{\beta}$ is the corresponding artificial equation of state, Chorin solves only this equation together with the flow equation, the incompressibility being achieved by a dynamic relaxation in time so that $\frac{\partial p}{\partial t} \rightarrow 0$ and the steady state is attained.

5. Equivalent Formulations for the Navier–Stokes Equations in “Non-Primitive” Variables

In what follows we intend to present some alternative formulations for the Navier–Stokes equations which, besides a certain theoretical interest, will lead to remarkable advantages in the numerical and computational approach. We will focus on the unsteady cases when we try to “split” the equations vis-a-vis the involved unknowns while the incompressibility condition implies the Laplace operator. This approach allows us to avoid the compatibility condition between the boundary and initial data (a condition which does not occur in the steady state case) but it requires the formulation of some integral type conditions for vorticity which will replace certain adherence conditions on the boundary.

Let us recall the Navier–Stokes equations in the domain D , of solid boundary ∂D (with the unit outward normal \mathbf{n}), that is

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\nabla p + \nu \nabla^2 \mathbf{v},$$

$$\nabla \cdot \mathbf{v} = 0$$

with the initial conditions $\mathbf{v}|_{t=0} = \mathbf{v}_0$ and the boundary conditions $\mathbf{v}|_{\partial D} = \mathbf{b}$, \mathbf{b} being the displacement velocity of the wall (boundary) which satisfies also (for every $t \geq 0$) the global condition $\int_{\partial D} \mathbf{b} \cdot \mathbf{n} ds = 0$.

Obviously the initial velocity \mathbf{v}_0 should fulfil the condition $\nabla \cdot \mathbf{v}_0 = 0$ (solenoidal vector) while \mathbf{b} and \mathbf{v}_0 should also satisfy the *compatibility condition* $\mathbf{n} \cdot \mathbf{b}|_{t=0} = \mathbf{n} \cdot \mathbf{v}_0|_{\partial D}$.

This compatibility condition, due only to the incompressibility, was used by Kato in 1967 [73] to establish the existence and uniqueness of the classical solution for the inviscid fluids (Euler equation) in the bidimensional case. At the same time, this compatibility condition together with the solenoidal character of the initial velocity, allows us to identify the appropriate linear space of the initial velocities which is finally H^1 [140].

In the following, by limiting ourselves to the plane case, we will try to give a new formulation for the Navier–Stokes equations using other variables than the “genuine” (“primitive”) ones. At the beginning we will write the Navier–Stokes system in orthogonal generalized (curvilinear) coordinates, followed by the stream function formulation. Then we will establish the equivalent equations in vorticity and stream function (the “ $\zeta - \Psi$ ” formulation) which reduces obviously the number of unknowns and eliminates the incompressibility condition whose numerical fulfilment could be extremely difficult. This formulation, the most used to approach the viscous incompressible fluids, has a weak point by the lack of the boundary condition for vorticity. We will show how it is possible to bypass this inconvenience by introducing a so-called *integral type condition for vorticity*.

5.1 Navier–Stokes Equations in Orthogonal Generalized Coordinates. Stream Function Formulation

The complexity of different practical problems, the diminution of the computational effort as well, lead to the choice of appropriate systems of reference (coordinates) which would simplify both the formulation and the solving of the problems.

In what follows we will write the Navier–Stokes equations in orthogonal curvilinear (generalized) coordinates. (For supplementary details, the consideration of non-orthogonal coordinates included, see, for instance, [153]). As a direct application, in the same orthogonal coordinates, we will give the transcription of the envisaged equations under the stream function form (formulation).

Let us now consider the generalized coordinates ξ_1, ξ_2, ξ_3 and, at a given point $\mathbf{r} = \mathbf{r}(\xi_1, \xi_2, \xi_3)$, let there be a triplet of unit vectors $\mathbf{e}_1 = \frac{1}{H_1} \frac{\partial \mathbf{r}}{\partial \xi_1}$, $\mathbf{e}_2 = \frac{1}{H_2} \frac{\partial \mathbf{r}}{\partial \xi_2}$, $\mathbf{e}_3 = \frac{1}{H_3} \frac{\partial \mathbf{r}}{\partial \xi_3}$, which are respectively tangent to the coordinate curves ξ_1, ξ_2, ξ_3 , and where $H_i = \left| \frac{\partial \mathbf{r}}{\partial \xi_i} \right|$ are the so-called *Lamé coefficients*. The fact that ξ_1, ξ_2, ξ_3 are generalized orthogonal coordinates implies automatically that $\mathbf{e}_i \cdot \mathbf{e}_k = \delta_{ik}$.

We know that the gradient, divergence, rotor (curl) and Laplacian operators have respectively (in these coordinates) the expressions [153]

$$\operatorname{div} \mathbf{A} = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial \xi_1} (h_2 h_3 A_1) + \frac{\partial}{\partial \xi_2} (h_1 h_3 A_2) + \frac{\partial}{\partial \xi_3} (h_1 h_2 A_3) \right]$$

where $\mathbf{A}(A_1, A_2, A_3)$,

$$\operatorname{grad} \Phi = \frac{\partial \Phi}{\partial \xi_1} \frac{\mathbf{e}_1}{h_1} + \frac{\partial \Phi}{\partial \xi_2} \frac{\mathbf{e}_2}{h_2} + \frac{\partial \Phi}{\partial \xi_3} \frac{\mathbf{e}_3}{h_3},$$

$$\operatorname{rot} \mathbf{A} = \frac{\mathbf{e}_1}{h_2 h_3} \left[\frac{\partial}{\partial \xi_2} (h_3 A_3) - \frac{\partial}{\partial \xi_3} (h_2 A_2) \right]$$

$$+ \frac{\mathbf{e}_2}{h_1 h_3} \left[\frac{\partial}{\partial \xi_3} (h_1 A_1) - \frac{\partial}{\partial \xi_1} (h_3 A_3) \right]$$

$$+ \frac{\mathbf{e}_3}{h_1 h_2} \left[\frac{\partial}{\partial \xi_1} (h_2 A_2) - \frac{\partial}{\partial \xi_2} (h_1 A_1) \right],$$

$$\Delta \Phi = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial \xi_1} \left(\frac{h_2 h_3}{h_1} \frac{\partial \Phi}{\partial \xi_1} \right) + \frac{\partial}{\partial \xi_2} \left(\frac{h_3 h_1}{h_2} \frac{\partial \Phi}{\partial \xi_2} \right) + \frac{\partial}{\partial \xi_3} \left(\frac{h_1 h_2}{h_3} \frac{\partial \Phi}{\partial \xi_3} \right) \right]$$

where Φ is any scalar function while \mathbf{A} is an arbitrary vector $\mathbf{A} = A_1 \mathbf{e}_1 + A_2 \mathbf{e}_2 + A_3 \mathbf{e}_3$.

But then we can rewrite the Navier–Stokes equations as

$$\frac{\partial \rho}{\partial t} + \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial \xi_1} (\rho h_2 h_3 u_1) + \frac{\partial}{\partial \xi_2} (\rho h_1 h_3 u_2) + \frac{\partial}{\partial \xi_3} (\rho h_1 h_2 u_3) \right] = 0,$$

$$\begin{aligned}
 & \rho \left(\frac{\partial u_1}{\partial t} + \frac{u_1}{h_1} \frac{\partial u_1}{\partial \xi_1} + \frac{u_2}{h_2} \frac{\partial u_1}{\partial \xi_2} + \frac{u_3}{h_3} \frac{\partial u_1}{\partial \xi_3} + \frac{u_1 u_2}{h_1 h_2} \frac{\partial h_1}{\partial \xi_2} \right. \\
 & \quad \left. - \frac{u_2^2}{h_1 h_2} \frac{\partial h_2}{\partial \xi_1} + \frac{u_1 u_3}{h_1 h_3} \frac{\partial h_1}{\partial \xi_3} - \frac{u_3^2}{h_1 h_3} \frac{\partial h_3}{\partial \xi_1} \right) \\
 &= \frac{1}{h_1^2 h_2 h_3} \left[\frac{\partial}{\partial \xi_1} (h_1 h_2 h_3 \tau_{\xi_1 \xi_1}) + \frac{\partial}{\partial \xi_2} (h_1^2 h_3 \tau_{\xi_1 \xi_2}) + \frac{\partial}{\partial \xi_3} (h_1^2 h_2 \tau_{\xi_1 \xi_3}) \right] \\
 & \quad - \frac{1}{h_1^2} \frac{\partial h_1}{\partial \xi_1} \tau_{\xi_1 \xi_1} - \frac{1}{h_1 h_2} \frac{\partial h_2}{\partial \xi_1} \tau_{\xi_2 \xi_2} - \frac{1}{h_1 h_3} \frac{\partial h_3}{\partial \xi_1} \tau_{\xi_3 \xi_3}, \\
 & \rho \left(\frac{\partial u_2}{\partial t} + \frac{u_1}{h_1} \frac{\partial u_2}{\partial \xi_1} + \frac{u_2}{h_2} \frac{\partial u_2}{\partial \xi_2} + \frac{u_3}{h_3} \frac{\partial u_2}{\partial \xi_3} + \frac{u_1 u_2}{h_1 h_2} \frac{\partial h_2}{\partial \xi_1} \right. \\
 & \quad \left. - \frac{u_1^2}{h_1 h_2} \frac{\partial h_1}{\partial \xi_2} - \frac{u_3^2}{h_2 h_3} \frac{\partial h_3}{\partial \xi_2} + \frac{u_2 u_3}{h_2 h_3} \frac{\partial h_2}{\partial \xi_3} \right) \\
 &= \frac{1}{h_1 h_2^2 h_3} \left[\frac{\partial}{\partial \xi_1} (h_2^2 h_3 \tau_{\xi_1 \xi_2}) + \frac{\partial}{\partial \xi_2} (h_1 h_2 h_3 \tau_{\xi_2 \xi_2}) + \frac{\partial}{\partial \xi_3} (h_1 h_2^2 \tau_{\xi_2 \xi_3}) \right] \\
 & \quad - \frac{1}{h_1 h_2} \frac{\partial h_1}{\partial \xi_2} \tau_{\xi_1 \xi_1} - \frac{1}{h_2^2} \frac{\partial h_2}{\partial \xi_2} \tau_{\xi_2 \xi_2} - \frac{1}{h_2 h_3} \frac{\partial h_3}{\partial \xi_2} \tau_{\xi_3 \xi_3}, \\
 & \rho \left(\frac{\partial u_3}{\partial t} + \frac{u_1}{h_1} \frac{\partial u_3}{\partial \xi_1} + \frac{u_2}{h_2} \frac{\partial u_3}{\partial \xi_2} + \frac{u_3}{h_3} \frac{\partial u_3}{\partial \xi_3} + \frac{u_1 u_3}{h_1 h_3} \frac{\partial h_3}{\partial \xi_1} \right. \\
 & \quad \left. - \frac{u_1^2}{h_1 h_3} \frac{\partial h_1}{\partial \xi_3} + \frac{u_2 u_3}{h_2 h_3} \frac{\partial h_3}{\partial \xi_2} - \frac{u_3^2}{h_2 h_3} \frac{\partial h_2}{\partial \xi_3} \right) \\
 &= \frac{1}{h_1 h_2 h_3^2} \left[\frac{\partial}{\partial \xi_1} (h_2 h_3^2 \tau_{\xi_1 \xi_3}) + \frac{\partial}{\partial \xi_2} (h_1 h_3^2 \tau_{\xi_2 \xi_3}) + \frac{\partial}{\partial \xi_3} (h_1 h_2 h_3 \tau_{\xi_3 \xi_3}) \right] \\
 & \quad - \frac{1}{h_1 h_3} \frac{\partial h_1}{\partial \xi_3} \tau_{\xi_1 \xi_1} - \frac{1}{h_2 h_3} \frac{\partial h_2}{\partial \xi_3} \tau_{\xi_2 \xi_2} - \frac{1}{h_3^2} \frac{\partial h_3}{\partial \xi_3} \tau_{\xi_3 \xi_3}.
 \end{aligned}$$

As regards the entities $\tau_{\xi_i \xi_j}$ they become

$$\tau_{\xi_1 \xi_1} = -p + 2\mu \left(\frac{1}{h_1} \frac{\partial u_1}{\partial \xi_1} + \frac{u_2}{h_1 h_2} \frac{\partial h_1}{\partial \xi_2} + \frac{u_1}{h_1 h_3} \frac{\partial h_1}{\partial \xi_3} \right) + \lambda (\operatorname{div} \mathbf{v}),$$

$$\tau_{\xi_2\xi_2} = -p + 2\mu \left(\frac{1}{h_2} \frac{\partial u_2}{\partial \xi_2} + \frac{u_1}{h_1 h_2} \frac{\partial h_2}{\partial \xi_1} + \frac{u_3}{h_2 h_3} \frac{\partial h_2}{\partial \xi_3} \right) + \lambda (\operatorname{div} \mathbf{v}),$$

$$\tau_{\xi_3\xi_3} = -p + 2\mu \left(\frac{1}{h_1} \frac{\partial u_1}{\partial \xi_3} + \frac{u_1}{h_1 h_3} \frac{\partial h_3}{\partial \xi_1} + \frac{u_2}{h_2 h_3} \frac{\partial h_3}{\partial \xi_2} \right) + \lambda (\operatorname{div} \mathbf{v}),$$

$$\tau_{\xi_1\xi_2} = \mu \left(\frac{1}{h_1} \frac{\partial u_2}{\partial \xi_1} + \frac{1}{h_2} \frac{\partial u_1}{\partial \xi_2} - \frac{u_2}{h_1 h_2} \frac{\partial h_2}{\partial \xi_1} - \frac{u_1}{h_1 h_2} \frac{\partial h_1}{\partial \xi_2} \right),$$

$$\tau_{\xi_1\xi_3} = \mu \left(\frac{1}{h_1} \frac{\partial u_3}{\partial \xi_1} + \frac{1}{h_3} \frac{\partial u_1}{\partial \xi_3} - \frac{u_3}{h_1 h_3} \frac{\partial h_3}{\partial \xi_1} - \frac{u_1}{h_1 h_3} \frac{\partial h_1}{\partial \xi_3} \right),$$

$$\tau_{\xi_2\xi_3} = \mu \left(\frac{1}{h_2} \frac{\partial u_3}{\partial \xi_2} + \frac{1}{h_3} \frac{\partial u_2}{\partial \xi_3} - \frac{u_3}{h_2 h_3} \frac{\partial h_3}{\partial \xi_2} - \frac{u_2}{h_2 h_3} \frac{\partial h_2}{\partial \xi_3} \right),$$

and where $\operatorname{div} \mathbf{v}$ is expressed as above (the writing of the divergence operator in generalized coordinates). These equations are used when their conservative form is not wanted.

In the following we will focus on the stream function formulation for the Navier–Stokes equations, a form used by certain numerical methods due to the advantage of the automatic fulfilment of the equation of continuity. At the beginning we deal with the plane and axially symmetric flows and then, by using the scalar and vectorial potentials, we will extend our search to the three-dimensional case.

Let us consider again the fluid velocity $\mathbf{v} = u_1 \mathbf{e}_1 + u_2 \mathbf{e}_2 + u_3 \mathbf{e}_3$. If these velocity components are independent on a certain coordinate (as the other flow parameters), the fluid flow is either plane (bidimensional) or axially-symmetric (revolution).

For sake of simplicity, suppose that all the parameters associated to the flow are, for example, independent of ξ_3 . In the plane case the flow will be the same as on the surface $\xi_3 = \text{constant}$, the component $u_3 = 0$ and $h_3 = 1$. On the other hand, in the axially-symmetric case, ξ_3 is the azimuthal angle and the ξ_3 derivatives are zero although the component u_3 is or is not zero while $h_3 \neq 1$.

In the axially-symmetric case with the azimuthal angle ξ_3 constant, the above written (in generalized coordinates) continuity equation will be identically satisfied by

$$h_3 u_1 = \frac{1}{h_2} \frac{\partial \psi}{\partial \xi_2}, \quad h_3 u_2 = -\frac{1}{h_1} \frac{\partial \psi}{\partial \xi_1}.$$

Denoting $h_3 u_3 = w$, we find (from the above expression for $\operatorname{rot} \mathbf{v}$) that the vorticity components are

$$\omega_1 = \frac{1}{h_2 h_3} \frac{\partial w}{\partial \xi_2}, \omega_2 = -\frac{1}{h_1 h_2} \frac{\partial w}{\partial \xi_1}, \omega_3 = -\frac{1}{h_3} D^2 \psi$$

where the differential operator D^2 is

$$D^2 = \frac{h_3}{h_1 h_2} \left[\frac{\partial}{\partial \xi_1} \left(\frac{h_2}{h_1 h_3} \frac{\partial}{\partial \xi_1} \right) + \frac{\partial}{\partial \xi_2} \left(\frac{h_1}{h_2 h_3} \frac{\partial}{\partial \xi_2} \right) \right].$$

If now we consider the Navier–Stokes equation which corresponds to ξ_3 together with the equation of rotation, we have

$$\frac{\partial w}{\partial t} - \frac{1}{h_1 h_2 h_3} \frac{\partial (\psi, w)}{\partial (\xi_1, \xi_2)} = \nu D^2 w,$$

$$\begin{aligned} \frac{\partial}{\partial t} (D^2 \psi) + \frac{2w}{h_1 h_2 h_3} \frac{\partial (w, h_3)}{\partial (\xi_1, \xi_2)} - \frac{1}{h_1 h_2 h_3} \frac{\partial (\psi, D^2 \psi)}{\partial (\xi_1, \xi_2)} \\ + \frac{2D^2 \psi}{h_1 h_2 h_3^2} \frac{\partial (\psi, h_3)}{\partial (\xi_1, \xi_2)} = \nu D^4 \psi \end{aligned}$$

where

$$\frac{\partial (\alpha, \beta)}{\partial (\xi, \eta)} = \frac{\partial \alpha}{\partial \xi} \frac{\partial \beta}{\partial \eta} - \frac{\partial \alpha}{\partial \eta} \frac{\partial \beta}{\partial \xi}.$$

In two dimensions, $w = 0$, $h_3 = 1$, $D^2 = \nabla^2$ and the previous equations become

$$\frac{\partial}{\partial t} (\nabla^2 \psi) - \frac{1}{h_1 h_2} \frac{\partial (\psi, \nabla^2 \psi)}{\partial (\xi_1, \xi_2)} = \nu \nabla^4 \psi$$

and

$$\nabla^2 \psi = -\omega.$$

In the tridimensional case, we start with the following representation for the velocity \mathbf{v} , namely $\mathbf{v} = \text{grad } \Phi + \text{rot } \mathbf{A}$, where $\nabla^2 \Phi = 0$ and the vector \mathbf{A} is solenoidal, that is $\text{div } \mathbf{A} = 0$. The last requirement could be satisfied by looking for \mathbf{A} under the form $S \text{grad } N$, which means to fulfil

$$\text{div } \mathbf{A} = S \Delta N + (\text{grad } S) \cdot (\text{grad } N) = 0,$$

or, in other terms, N should be a harmonic function while the surfaces $S = \text{constant}$ and $N = \text{constant}$ have to be orthogonal.

Obviously, the above representation satisfies implicitly the continuity equation. Applying the rotor (curl) operator to this representation we also get

$$\operatorname{div}(\operatorname{grad} \mathbf{A}) = -\boldsymbol{\omega}.$$

In other words, in the three-dimensional case, the writing of the involved equations (using the scalar and vectorial potentials) comes to the consideration of the last equation together with $\nabla^2 \Phi = 0$ and the equation of vorticity $\boldsymbol{\omega}$.

At the same time, by substituting the expression $\mathbf{v} = \nabla\psi \times \mathbf{k}$ into the definition of vorticity, we obtain

$$-\nabla^2 \psi = \zeta.$$

Concerning the boundary conditions for these two scalar equations, they could be deduced from those already known by a separate consideration of the normal and tangential velocity components at the boundary points. If \mathbf{n} is the unit outward normal vector drawn to the boundary ∂D , $\boldsymbol{\tau}$ is the unit tangent vector counterclockwise oriented, s is the natural parameter (the arc length) on the boundary, then the condition $\mathbf{v}|_{\partial D} = \mathbf{b}$ implies

$$\mathbf{n} \cdot (\nabla\psi \times \mathbf{k})|_{\partial D} = (\mathbf{k} \times \mathbf{n}) \cdot \nabla\psi|_{\partial D} = \boldsymbol{\tau} \cdot \nabla\psi|_{\partial D} = \left. \frac{d\psi}{ds} \right|_{\partial D} = \mathbf{n} \cdot \mathbf{b},$$

respectively

$$\boldsymbol{\tau} \cdot (\nabla\psi \times \mathbf{k})|_{\partial D} = (\mathbf{k} \times \boldsymbol{\tau}) \cdot \nabla\psi|_{\partial D} = -\mathbf{n} \cdot \nabla\psi|_{\partial D} = - \left. \frac{d\psi}{dn} \right|_{\partial D} = \boldsymbol{\tau} \cdot \mathbf{b}.$$

The first of these conditions, after integrating along the boundary, leads to a Dirichlet condition for ψ . By accepting that D is a simply connected domain, from the global condition $\int_{\partial D} \mathbf{b} \cdot \mathbf{n} ds = 0$ we get the warranty that the respective integral along the boundary defines a uniform function $a(s, t)$, to within an additive function of time $A(t)$, such that

$$a(s, t) = \int_{s_1}^s \mathbf{n}(s') \cdot \mathbf{b}(s', t) ds' + A(t),$$

where s_1 is the natural coordinate of a fixed point of ∂D .

To simplify the form of the boundary condition for ψ we will not take into consideration the term $A(t)$ and, denoting by $b(s, t) = -\boldsymbol{\tau}(s) \cdot \mathbf{b}(s, t)$, the two conditions could be written

$$\Psi|_{\partial D} = a, \quad \left. \frac{d\psi}{dn} \right|_{\partial D} = b.$$

Regarding the initial condition, this implies a vorticity condition at the instant $t = 0$, precisely

$$\zeta|_{t=0} = (\nabla \times \mathbf{v}_0) \cdot \mathbf{k}.$$

With respect to the compatibility condition attached to the Navier–Stokes system, that is $\mathbf{n} \cdot \mathbf{b}|_{t=0} = \mathbf{n} \cdot \mathbf{v}_0|_{\partial D}$, it could be rewritten in the form

$$\frac{\partial a(s, 0)}{\partial s} = \mathbf{n} \cdot \mathbf{v}_0|_{\partial D},$$

where also $\nabla \cdot \mathbf{v}_0 = 0$.

If these last two conditions on the data are satisfied, Guermond J. L. and Quartapelle L. have rigorously established in 1993 [126], the equivalence between the genuine formulation of the Navier–Stokes equations and the “ $\zeta - \psi$ formulation”, which means with the system

$$\frac{\partial \zeta}{\partial t} - \nu \nabla^2 \zeta + J(\zeta, \psi) = 0,$$

$$-\nabla^2 \psi = \zeta,$$

$$\psi|_{\partial D} = a, \quad \left. \frac{d\psi}{dn} \right|_{\partial D} = b,$$

$$\zeta|_{t=0} = (\nabla \times \mathbf{v}_0) \cdot \mathbf{k}.$$

Obviously this formulation is nonlinear due to the presence of the Jacobian which is “coupling” the equations in ζ and ψ ; further, there are two boundary conditions for ψ and none for ζ . If the difficulties caused by this nonlinearity can be overtaken by combining some explicit or implicit step-time algorithms within suitable iterative procedures, those connected with the boundary conditions will be avoided by one of the following methods (formulations) which are presented in the sequel.

5.1.1 The Biharmonic Formulation

The simplest way to avoid the lack of a boundary condition for vorticity is to eliminate, from the previous system, the vorticity itself. By substituting the expression for vorticity $\zeta = -\nabla^2\psi$ into the transport equation for it (the vorticity equation) we reach the problem

$$\frac{\partial^2 \nabla^2 \psi}{\partial t} - \nu \nabla^4 \psi + J(\nabla^2 \psi, \psi) = 0,$$

$$\psi|_{\partial D} = a, \quad \left. \frac{d\psi}{dn} \right|_{\partial D} = b, \quad \psi|_{t=0} = \psi_0,$$

where ψ_0 is the solution of the Dirichlet problem

$$-\nabla^2 \psi_0 = (\nabla \times \mathbf{v}_0) \cdot \mathbf{k}, \quad \psi_0|_{\partial D} = a(s, t),$$

the data \mathbf{v}_0 and a satisfying both the compatibility and solenoidal condition. In the above formulation the boundary conditions don't lead to an overdetermined problem (as they seemed to in the " $\zeta - \psi$ " formulation) because the equation in ψ is of fourth order. There are many numerical procedures either in finite differences or in boundary elements (for the linearized variants). This problem could also be written in the following variational form (which is essential for a finite element type method):

“To find a function $\psi \in H^2(D)$ such that $\psi|_{\partial D} = a$ and $\left. \frac{d\psi}{dn} \right|_{\partial D} = b$ and

$$\left(\nabla \varphi, \frac{\partial}{\partial t} \nabla \psi \right) + \nu (\nabla^2 \varphi, \nabla^2 \psi) + (J(\varphi, \psi), \nabla^2 \psi) = 0, \quad \forall \varphi \in H_0^2(D),$$

where (\cdot, \cdot) denotes the inner (dot) product in L^2 while $H^2(D)$ and $H_0^2(D)$ are the standard notations for the corresponding Sobolev spaces”.

5.2 A “Coupled” Formulation in Vorticity and Stream Function

This new formulation envisages a new way to avoid the difficulties joined to a double condition for ψ on ∂D and to a total absence of conditions for ζ .

We remark that, even in the absence of the non-linear term from the vorticity equation, the involved equations should be considered as being coupled through the boundary conditions. In other terms, one of the conditions for ψ must be “associated” with the vorticity equation but this equation is not sufficient to determine alone the unique ζ . Therefore,

in this approach, a boundary condition for ζ is not needed but the two equations should be solved necessarily as coupled.

More precisely, the Dirichlet condition $\psi|_{\partial D} = a$ will be attached to the rotation (vorticity) equation

$$-\nu \nabla^2 \zeta + \frac{\partial \zeta}{\partial t} + J(\zeta, \psi) = 0$$

while the Neumann condition $\frac{d\psi}{dn}\Big|_{\partial D} = b$ is associated with the equation $-\nabla^2 \psi = \zeta$. But this last equation will not be a real Poisson equation since ζ is an unknown and so the compatibility condition for such a Neumann problem

$$\int_D \zeta dv = \int_{\partial D} b ds \tag{3.1}$$

is not required anymore. Obviously we also have the initial condition for vorticity, i.e., $\zeta|_{t=0} = (\nabla \times \mathbf{v}_0) \cdot \mathbf{k}$.

To such a formulation one could join either ADI techniques with finite differences (Napolitano) [94] or Chebyshev spectral approximations (Heinrichs)[63].

At the same time, in view of the construction of some finite element type methods, one could state the following variational (mixed) formulation for the above equations:

“To determine $\zeta \in H^1(D)$ and $\psi \in H^1(D)$ such that $\psi|_{\partial D} = a$ and

$$\nu (\nabla \varphi, \nabla \zeta) + \left(\varphi, \frac{\partial \zeta}{\partial t} \right) + (\varphi, J(\zeta, \psi)) = 0, \forall \varphi \in H_0^1(D),$$

$$(\nabla \xi, \nabla \psi) - (\xi, \zeta) = \int_{\partial D} \xi b ds, \forall \xi \in H^1(D)$$

where again (\cdot, \cdot) denotes the inner (dot) product in L^2 while $H^1(D)$ and $H_0^1(D)$ are the standard notations for the Sobolev spaces”.

We finally remark that in this formulation one of the two conditions on ψ is imposed implicitly as a natural condition.

5.3 The Separated (Uncoupled) Formulation in Vorticity and Stream Function

In what follows we will try to separate the equations from the “ $\zeta - \psi$ formulation”. To do that we need some supplementary conditions for vorticity which should replace the boundary conditions for it. These supplementary conditions will be stated in a different form versus the classical boundary conditions, since they have an integral character.

Due to L. Quartapelle and Valz-Cris we have the following result [127]:

THEOREM A function ζ defined on D , is such that $\zeta = -\nabla^2\psi$ with $\psi|_{\partial D} = a$ and $\left.\frac{d\psi}{dn}\right|_{\partial D} = b$ if and only if

$$\int_D \zeta \eta dv = \int_{\partial D} a \left(\frac{d\eta}{dn} - b\eta \right) ds,$$

for any harmonic function η on D , that is $\nabla^2\eta = 0$ in D .

This integral condition, whose existence has been anticipated by other scientists, has to be considered as a condition of a unique type vis-a-vis the usual classical boundary conditions. This is not a boundary integral formulation due to the presence of the volume integral.

If we introduce the fundamental solution $G(\mathbf{r}, \mathbf{r}')$ for the Laplace operator (the Green function) through the equation $-\nabla'^2 G(\mathbf{r}, \mathbf{r}') = 4\pi\delta(\mathbf{r} - \mathbf{r}')$ where $\delta(\mathbf{r} - \mathbf{r}')$ is the Dirac distribution in two dimensions, by using the Poisson (Green) formula for a pair of regular functions α and β (on D), that is

$$\int_D (\alpha \nabla^2 \beta - \beta \nabla^2 \alpha) dv = \int_{\partial D} \left(\alpha \frac{d\beta}{dn} - \beta \frac{d\alpha}{dn} \right) ds$$

where now $\alpha = G(\mathbf{r}, \mathbf{r}')$ while ψ satisfies $-\nabla^2\psi = \zeta$, $\psi|_{\partial D} = a$ and $\left.\frac{d\psi}{dn}\right|_{\partial D} = b$, we obtain the following new form of the integral condition

$$\begin{aligned} & -4\pi\psi(\mathbf{r})\gamma(\mathbf{r}) + \int_D G(\mathbf{r}, \mathbf{r}') \zeta(\mathbf{r}) dv' \\ & = \int_{\partial D} \left[a(s') \frac{dG(\mathbf{r}, \mathbf{r}')}{dn'} - b(s') G(\mathbf{r}, \mathbf{r}') \right] ds', \end{aligned}$$

with $\gamma(\mathbf{r}) = 1, 0, \frac{1}{2}$ as \mathbf{r} is inside, outside or on the boundary point of D .

The introduction of the above integral condition allows us to break the “ $\zeta - \psi$ formulation” into the two problems

$$\left(-\nu \nabla^2 + \frac{\partial}{\partial t} \right) \zeta = -J(\zeta, \psi), \quad \int_D \zeta \eta dv = \int_{\partial D} \left(a \frac{\partial \eta}{\partial n} - b\eta \right) ds$$

and

$$-\nabla^2\psi = \zeta, \quad \psi|_{\partial D} = a$$

where η is an arbitrary harmonic function.

Obviously, in the absence of the nonlinear term, a complete separation of the two equations may be achieved so that they could be solved successively (one by one) in the indicated order.

At the same time, if the second equation is accompanied by the Neumann condition

$$\left. \frac{d\psi}{dn} \right|_{\partial D} = b$$

the result is completely equivalent. The same thing happens if we consider also the arbitrary function of time $A(t)$, the integral condition being invariant with respect to this choice.

Among the applications of the vorticity integral condition we should mention the works of Dennis and his collaborators where one has studied the fluid flows past flat plates of finite size and which are “aligned” with the stream, the fluid flows around circular cylinders or spheres and even the Oseen model [25], [26], [27], [28]. More precisely, in all these researches, one deals with series expansions for ζ and ψ , with respect to different suitable orthogonal function systems and then one keeps only a finite number of series terms. The final results agree well with the classical ones [42].

Now we will make some considerations on the equivalent formulations in the three-dimensional case. For these flows some additional difficulties occur due to the fact that the components of the velocity vector (which is solenoidal) are, in general, different from zero and two of them (the tangential components) should be determined on the solid boundaries.

We would limit ourselves to the “ $\varphi - \zeta - \mathbf{A}$ ” formulation, backed by the (always possible) vector decomposition

$$\mathbf{v} = \nabla\varphi + \nabla \times \mathbf{A} \quad \text{where} \quad (\nabla \times \mathbf{A}) \cdot \mathbf{n}|_{\partial D} = 0.$$

Concerning the transport equation for vorticity (the rotation equation), it is known that now it has the form (we denoted $\zeta \equiv \boldsymbol{\zeta}$)

$$\frac{\partial \zeta}{\partial t} + \nabla \times (\zeta \times \mathbf{v}) = \nu \nabla^2 \zeta$$

with an initial condition (corresponding to the initial condition for \mathbf{v}) of the type $\zeta|_{t=0} = \nabla \times \mathbf{v}_0$.

By applying then the divergence operator to the vorticity equation we get

$$\frac{\partial (\nabla \cdot \zeta)}{\partial t} = \nu \nabla^2 (\nabla \cdot \zeta)$$

with the supplementary initial condition $\nabla \cdot \zeta|_{t=0} = \nabla \cdot (\nabla \times \mathbf{v}_0) = 0$.

If this equation is also completed by the supplementary homogeneous condition $\nabla \cdot \zeta|_{\partial D} = 0$, for $t > 0$, the unique solution of the above equation will be identical to zero, which means ζ should be a solenoidal vector for $t > 0$ ($\nabla \cdot \zeta = 0$). This last condition, introduced by Lighthill, together with the initial condition for ζ , are the necessary and sufficient requirements for ζ to be solenoidal, a condition demanded by the definition itself of ζ ($\zeta = \nabla \times \mathbf{v}$).

In the sequel we will limit ourselves to considering the “ $\varphi - \zeta - \mathbf{A}$ ” formulation based on the unique (always possible) “splitting” of the velocity vector by $\mathbf{v} = \nabla\varphi + \nabla \times \mathbf{A}$ where the vector \mathbf{A} is determined up to the gradient of a scalar function ψ and it fulfils the condition $(\nabla \times \mathbf{A}) \cdot \mathbf{n}|_{\partial D} = 0$.

Obviously the above representation and the incompressibility condition lead also to

$$-\nabla^2\varphi = 0,$$

which means φ will be harmonic in D .

The boundary conditions which are imposed on φ and \mathbf{A} will be derived from those imposed on \mathbf{v} by separation of the normal and tangential components from $\mathbf{v}|_{\partial D} = \mathbf{b}$.

We accept, together with Hirasaki and Hellums, that regarding the boundary condition on the normal direction, it will be satisfied by

$$\mathbf{n} \cdot \nabla\varphi|_{\partial D} = \mathbf{n} \cdot \mathbf{b}$$

and

$$\mathbf{n} \cdot (\nabla \times \mathbf{A})|_{\partial D} = 0,$$

the last condition being, in fact, synonymous with the orthogonality condition

$$\int_D \nabla\varphi \cdot (\nabla \times \mathbf{A}) \, dv = 0.$$

The determination of φ leads to solving a Neumann problem which, taking into consideration the global condition

$$\int_{\partial D} \mathbf{b} \cdot \mathbf{n} \, ds = 0,$$

can be uniquely solved to within an arbitrary function of time $\Phi(t)$. Once φ is determined, the tangential part of the boundary condition for \mathbf{v} , that is $\mathbf{n} \times \mathbf{v}|_{\partial D} = \mathbf{n} \times \mathbf{b}$, becomes also

$$\mathbf{n} \times \nabla \times \mathbf{A}|_{\partial D} = \mathbf{n} \times (-\nabla\varphi|_{\partial D} + \mathbf{b})$$

By applying now the rotor (curl) operator to both sides of the decomposition $\mathbf{v} = \nabla\varphi + \nabla \times \mathbf{A}$, we get for \mathbf{A} the equation $\nabla \times \nabla \times \mathbf{A} = \boldsymbol{\zeta}$ and the attached boundary conditions

$$\mathbf{n} \cdot (\nabla \times \mathbf{A})|_{\partial D} = 0, \quad \mathbf{n} \times \nabla \times \mathbf{A}|_{\partial D} = \mathbf{n} \times (-\nabla\varphi|_{\partial D} + \mathbf{b}).$$

But the above system is equivalent with

$$-\nabla^2 \mathbf{A} = \boldsymbol{\zeta}, \quad \mathbf{n} \times \mathbf{A}|_{\partial D} = 0,$$

$$\mathbf{n} \times (\nabla \times \mathbf{A})|_{\partial D} = \mathbf{n} \times (-\nabla\varphi|_{\partial D} + \mathbf{b}), \quad \nabla \cdot \mathbf{A}|_{\partial D} = 0.$$

Finally the following results hold:

THEOREM 3.1. *The Navier–Stokes system written in the genuine variables \mathbf{v} and p together with a Cauchy–Dirichlet (initial-boundary) condition is equivalent with the following system in variables $\varphi, \boldsymbol{\zeta}$ and \mathbf{A} ,*

$$-\nabla^2 \varphi = 0, \quad \mathbf{n} \cdot \nabla\varphi|_{\partial D} = \mathbf{n} \cdot \mathbf{b},$$

$$\frac{\partial \boldsymbol{\zeta}}{\partial t} - \nu \nabla^2 \boldsymbol{\zeta} + \nabla \times [(\boldsymbol{\zeta} + \nabla \times \mathbf{A})] = 0, \quad \boldsymbol{\zeta}|_{t=0} = \nabla \times \mathbf{v}_0, \quad \nabla \cdot \boldsymbol{\zeta}|_{\partial D} = 0,$$

$$-\nabla^2 \mathbf{A} = \boldsymbol{\zeta}, \quad \mathbf{n} \times \mathbf{A}|_{\partial D} = 0,$$

$$\mathbf{n} \times \nabla \times \mathbf{A}|_{\partial D} = \mathbf{n} \times (-\nabla\varphi|_{\partial D} + \mathbf{b}), \quad \nabla \cdot \mathbf{A}|_{\partial D} = 0,$$

provided that the data $\mathbf{n} \cdot \mathbf{b}$ and \mathbf{v}_0 satisfy the restrictions

$$\int_{\partial D} \mathbf{b} \cdot \mathbf{n} ds = 0, \quad \nabla \cdot \mathbf{v}_0 = 0, \quad \mathbf{n} \cdot \mathbf{b}|_{t=0} = \mathbf{n} \cdot \mathbf{v}_0|_{\partial D}.$$

As regards the vorticity integral condition this could be written now in the form [126]

$$\int_D \boldsymbol{\zeta} \cdot \boldsymbol{\eta} dv = \int_{\partial D} [\mathbf{n} \times (-\nabla\varphi|_{\partial D} + \mathbf{b})] \cdot \boldsymbol{\eta} ds$$

where $\boldsymbol{\eta}$ is an arbitrary solenoidal vector.

Correspondingly, the “uncoupled” (separated) form in the “ $\varphi - \boldsymbol{\zeta} - \mathbf{A}$ ” formulation would be [126]

$$\begin{aligned} -\nabla^2\varphi &= 0, \quad \mathbf{n} \cdot \nabla\varphi|_{\partial D} = \mathbf{n} \cdot \mathbf{b}, \\ -\nu\nabla^2\boldsymbol{\zeta} + \frac{\partial\boldsymbol{\zeta}}{\partial t} + \nabla \times [\boldsymbol{\zeta} \times (\nabla\varphi \times \nabla \times \mathbf{A})] &= 0, \quad \boldsymbol{\zeta}|_{t=0} = \nabla \times \mathbf{v}_0, \\ \int_D \boldsymbol{\zeta} \cdot \boldsymbol{\eta} dv &= \int_{\partial D} [\mathbf{n} \times (-\nabla\varphi|_{\partial D} + \mathbf{b})] \cdot \boldsymbol{\eta} ds, \quad \nabla \cdot \boldsymbol{\zeta}|_{\partial D} = 0, \\ &[-\nabla^2\boldsymbol{\eta} = 0, \quad \nabla \cdot \boldsymbol{\eta}|_{\partial D} = 0], \\ -\nabla^2\mathbf{A} &= \boldsymbol{\zeta}, \quad \mathbf{n} \times \mathbf{A}|_{\partial D} = 0, \quad \nabla \cdot \mathbf{A}|_{\partial D} = 0. \end{aligned}$$

It is important to remark that, in three dimensions, the equation of rotation (vorticity) has been completed by both boundary and integral conditions, the last of them implying all the three components of vorticity.

5.4 An Integro-Differential Formulation

The establishing of a unique integro-differential equation which is equivalent with the Navier–Stokes system is due to Wu [157], [158].

Basically, the procedure uses both the rotation equation

$$\frac{\partial\boldsymbol{\omega}}{\partial t} + \text{rot}(\boldsymbol{\omega} \times \mathbf{v}) = \nu\Delta\boldsymbol{\omega},$$

and the Poisson equation

$$\Delta\mathbf{v} = -\text{rot}\boldsymbol{\omega},$$

the last one being the consequence of the consideration of the condition $\text{div } \mathbf{v} = 0$ into the identity

$$\Delta\mathbf{v} = -\text{grad}(\text{div } \mathbf{v}) - \text{rot}\boldsymbol{\omega}.$$

Let now $\boldsymbol{\xi}$ and \mathbf{x} be a variable and a fixed point respectively, both belonging to the flow domain, while $\mathbf{r} = |\boldsymbol{\xi} - \mathbf{x}|$. It is known that the solution of the above Poisson equation is given (in three and then in two dimensions) by

$$\mathbf{v}(\mathbf{x}) = \frac{1}{4\pi} \int_D \frac{\text{rot } \boldsymbol{\omega}(\boldsymbol{\xi})}{r} dv(\boldsymbol{\xi}) + \frac{1}{4\pi} \int_{\partial D} \left[\frac{1}{r} \frac{d\mathbf{v}(\boldsymbol{\xi})}{dn} - \mathbf{v}(\boldsymbol{\xi}) \frac{d}{dn} \left(\frac{1}{r} \right) \right] ds(\boldsymbol{\xi}),$$

respectively

$$\begin{aligned} \mathbf{v}(\mathbf{x}) &= \frac{1}{2\pi} \int_D \ln \left(\frac{1}{r} \right) \text{rot } \boldsymbol{\omega}(\boldsymbol{\xi}) dv(\boldsymbol{\xi}) \\ &+ \frac{1}{2\pi} \int_{\partial D} \left[\ln \left(\frac{1}{r} \right) \frac{d\mathbf{v}(\boldsymbol{\xi})}{dn} - \mathbf{v}(\boldsymbol{\xi}) \frac{d}{dn} \ln \left(\frac{1}{r} \right) \right] ds(\boldsymbol{\xi}). \end{aligned}$$

But we also know that

$$\frac{d\mathbf{v}(\boldsymbol{\xi})}{dn} = (\mathbf{n} \cdot \text{grad}) \mathbf{v}(\boldsymbol{\xi}) = (\text{grad}_{\boldsymbol{\xi}} \mathbf{v}) \cdot \mathbf{n}$$

and both the identity

$$(\text{grad } \mathbf{v}) \cdot \mathbf{n} = \boldsymbol{\omega} \times \mathbf{n} + (\mathbf{n} \times \text{grad}) \times \mathbf{v} + \mathbf{n} (\text{div } \mathbf{v})$$

and the (Gauss) theorem

$$\int_D \text{rot } \mathbf{P} dv = - \iint_{\partial D} (\mathbf{P} \times \mathbf{n}) ds$$

are valid.

Making then successively $\mathbf{P} = \frac{\boldsymbol{\omega}}{r}$ and $\mathbf{P} = \boldsymbol{\omega} \ln \left(\frac{1}{r} \right)$, we get

$$\int_D \frac{\text{rot}_{\boldsymbol{\xi}} \boldsymbol{\omega}}{r} dv = - \int_{\partial D} \frac{\boldsymbol{\omega} \times \mathbf{n}}{r} ds + \int_D \frac{(\boldsymbol{\xi} - \mathbf{x}) \times \boldsymbol{\omega}}{r^3} dv,$$

respectively

$$\int_D \ln \left(\frac{1}{r} \right) \text{rot}_{\boldsymbol{\xi}} \boldsymbol{\omega} dv = - \int_{\partial D} \boldsymbol{\omega} \times \mathbf{n} \ln \left(\frac{1}{r} \right) ds + \int_D \frac{(\boldsymbol{\xi} - \mathbf{x}) \times \boldsymbol{\omega}}{r^2} dv$$

which leads to

$$\begin{aligned} \mathbf{v}(\mathbf{x}) &= \frac{1}{4\pi} \int_D \frac{(\boldsymbol{\xi} - \mathbf{x}) \times \boldsymbol{\omega}}{r^3} dv(\boldsymbol{\xi}) \\ &+ \frac{1}{4\pi} \int_{\partial D} \left[\frac{(\mathbf{n} \times \text{grad}_{\boldsymbol{\xi}}) \times \mathbf{v}}{r} + \frac{\mathbf{n} \cdot (\boldsymbol{\xi} - \mathbf{x})}{r^3} \mathbf{v}(\boldsymbol{\xi}) \right] ds(\boldsymbol{\xi}), \end{aligned}$$

respectively

$$\begin{aligned} \mathbf{v}(\mathbf{x}) &= \frac{1}{2\pi} \int_D \frac{(\boldsymbol{\xi} - \mathbf{x}) \times \boldsymbol{\omega}}{r^2} dv(\boldsymbol{\xi}) \\ &+ \frac{1}{2\pi} \int_{\partial D} \left[(\mathbf{n} \times \text{grad}_{\boldsymbol{\xi}}) \times \mathbf{v} \ln\left(\frac{1}{r}\right) + \frac{\mathbf{n} \cdot (\boldsymbol{\xi} - \mathbf{x})}{r^2} \mathbf{v}(\boldsymbol{\xi}) \right] ds(\boldsymbol{\xi}). \end{aligned}$$

Using the adherence (no-slip) condition $\mathbf{v}|_{\partial D} = 0$ and the consequence $(\mathbf{n} \times \text{grad}) \times \mathbf{v}|_{\partial D} = 0$, for interior flow we would have

$$\mathbf{v}(\boldsymbol{\xi}) = \frac{1}{A} \int_D \frac{(\boldsymbol{\xi} - \mathbf{x}) \times \boldsymbol{\omega}}{r^d} dv$$

while for the exterior flow with the free-stream velocity \mathbf{v}_{∞} ,

$$\mathbf{v}(\boldsymbol{\xi}) = \mathbf{v}_{\infty} + \frac{1}{A} \int_D \frac{(\boldsymbol{\xi} - \mathbf{x}) \times \boldsymbol{\omega}}{r^d} dv,$$

with $A = 4\pi$ or 2π and $d = 3$ or 2 (according to the tri or bidimensional case).

The substitution of these representations in the vorticity equation gives rise to the integro-differential equation.

6. Similarity of the Viscous Incompressible Fluid Flows

The *(dynamic) similarity method* is a very useful tool not only in aerohydrodynamics but even in the approach to many other physical or technical problems. This method allows us to specify all the conditions which should be imposed on some laboratory models such that the information obtained from laboratory experiments could be extended to the real situations. At the same time this method provides a special technique for getting a whole class of solutions (depending on certain parameters), starting with a solution of the system of equations which governs the respective problem (process).

This method will also support the possibility of the construction of some nondimensional solutions, a fundamental feature in the numerical approach to the equations associated to the process (problem).

Generally speaking, two physical phenomena are said to be *(dynamically) similar* if the parameters characterizing one of these phenomena could be directly obtained from the same parameters for the second phenomenon (and which are, obviously, evaluated at the “similar” spatial

points and at the same moments) by a simple multiplication with some unchanged factors called the *similarity coefficients*.

Let us now establish the similarity conditions for two viscous incompressible fluid flow without any heat interchange with the surroundings (isothermal). Considering then a system of characteristic (reference) values for time, length (coordinates), velocity, pressure and mass (body) forces, denoted respectively by t_c , L_c , v_c , p_c , f_c and operating the variable and function change

$$t = t_c \bar{t}, x_i = L_c \bar{x}_i, u_i = v_c \bar{u}_i, p = p_c \bar{p}, f = f_c \bar{f},$$

where the quantities with “bar” are obviously nondimensionalized, the Navier–Stokes system becomes⁴

$$\frac{v_c}{t_c} \bar{u}_{i,\bar{i}} + \frac{v_c^2}{t_c} (\bar{v} \cdot \bar{\nabla}) \bar{u}_i = f_c \bar{f}_i - \frac{p_c}{\rho L_c} \bar{\nabla} \bar{p} + \frac{\nu v_c}{L_c^2} \bar{\Delta} \bar{u}_i$$

and

$$\frac{v_c}{L_c} \bar{u}_{i,\bar{i}} = 0.$$

Dividing by $\frac{v_c^2}{L_c}$ and supposing that the conservative terms are not neglected, we get

$$Sh \bar{u}_{i,\bar{i}} + (\bar{v} \cdot \bar{\nabla}) \bar{u}_i = \frac{1}{Fr} \bar{f}_i - Eu \bar{\nabla} \bar{p} + \frac{1}{R} \bar{\Delta} \bar{u}_i$$

and

$$\bar{u}_{i,\bar{i}} = 0,$$

where the following nondimensionalized entities (also called the *similarity numbers*) interfere:

$$Sh = \frac{L_c}{v_c t_c} \text{ (Strouhal number); } \quad Eu = \frac{p_c}{\rho v_c^2} \text{ (Euler number);}$$

$$R = \frac{v_c L_c}{\nu} \text{ (Reynolds number); } \quad Fr = \frac{v_c}{f_c L_c} \text{ (Froude number).}$$

The above equations are the nondimensionalized Navier–Stokes equations. To them we should add the nondimensionalized initial and boundary conditions, according to the given problem.

If two viscous incompressible isothermal fluid flows are similar, the parameters (field values) of one of them could be obtained from the same

⁴The components of the velocity \mathbf{v} are now denoted by u_i while those of the nondimensional velocity $\bar{\mathbf{v}}$ are denoted by \bar{u}_i .

parameters of the second flow, by multiplying with the same factor, i.e., both the equations and the initial and boundary conditions (which ensure at least the solution uniqueness) should be identical and, consequently, the similarity numbers Sh , Eu , Fr and R are also the same. Obviously the respective solutions will depend on the parameters t_c , L_c , v_c , p_c , f_c and even on ν and ρ (supposed constant but, generally, different in the two flows), all the parameters being linked by the condition that Sh , Eu , Fr and R take the same values in the two similar flows. Therefore we are led to a class of solutions depending on a reduced (with four) number of free parameters, an important theoretical result.

In the case when we put away both the isothermal and homogeneous character of the flow, but supposing that the variation of the temperature and of the concentration do not influence the viscosity, the thermal conductivity as well as other thermodynamical properties of the fluid then, if the radiation heat is ignored, the equations of the viscous incompressible fluid could be written as

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{1}{\rho} \text{grad} p + \beta \mathbf{g} \Delta T + \nu \nabla^2 \mathbf{v}$$

and

$$\text{div} \mathbf{v} = 0$$

where T is the temperature whose variation is ΔT , β is the thermal coefficient of the fluid expansion connected with the Archimedean force due to the density difference, that is $\mathbf{F} = \beta \mathbf{g} \Delta T$, and \mathbf{g} is the gravity acceleration.

Concerning the equation of heat conduction, it takes the form

$$\frac{\partial T}{\partial t} + \mathbf{v} \cdot \text{grad} T = a \nabla^2 T,$$

where a is a constant which is called the *thermal diffusion coefficient*.

By using again the above equations, the technique of the similarity method, we obtain the nondimensionalized system

$$Sh \frac{\partial \bar{v}}{\partial \bar{t}} + (\bar{v} \cdot \bar{\nabla}) \bar{v} = -Eu \bar{\nabla} \bar{p} + \frac{1}{Fr} \bar{g} + \frac{1}{R} \bar{\nabla}^2 \bar{v},$$

$$Sh \frac{\partial \bar{T}}{\partial \bar{t}} + \bar{v} \cdot \bar{\nabla} \bar{T} = \frac{1}{Pe} \bar{\nabla}^2 \bar{T}$$

where, besides the Strouhal, Euler, Reynolds and Froude numbers (the last being now defined by $Fr = \frac{v_c^2}{g L_c \beta \Delta T}$), there arises also the Peclet number Pe which is defined by $Pe = \frac{v_c L_c}{a}$. Sometimes the Froude and

Peclet numbers are replaced by the Prandtl number Pr and the Grashof number Gr , defined as $Pr = \frac{Pe}{R}$ and $Gr = \frac{R^2}{Fr}$.

Considering now the adjacent phenomenon of the propagation (diffusion) of the involved substance if C is the concentration of the “mixture” and D the diffusion coefficient of it, it is shown [87] that the differential equation of the mixture diffusion has exactly the same form as the equation of heat conduction, namely

$$\frac{\partial C}{\partial t} + \mathbf{v} \cdot \text{grad } C = D \nabla^2 C,$$

which, by nondimensionalizing, becomes

$$Sh \frac{\partial \bar{C}}{\partial t} + \bar{\mathbf{v}} \cdot \bar{\nabla} \bar{C} = \frac{1}{Pe_d} \bar{\nabla}^2 \bar{C}$$

where Pe_d is the diffusion Peclet number which is different from the ordinary Peclet number, defined above (and in which a is replaced by D), namely

$$Pe_d = \frac{v_c L_c}{D}$$

and to which there corresponds a diffusion Prandtl number Pr_d (also-called the Schmidt Sc number) by the relation

$$Sc \equiv Pr_d = \frac{Pe_d}{R} = \frac{\nu}{D}.$$

In the sequel we will consider only the steady flows of the viscous incompressible (homogeneous and isothermal) fluids, in the absence of the external (mass, body) forces. These flows which basically depend only on a unique similarity number (the Reynolds number), are of great practical interest within the context of dividing these fluid flows in two great categories: the fluid flows with small (low) Reynolds number and the fluid flows with high (large) Reynolds number.

6.1 The Steady Flows Case

Let us consider again the Navier–Stokes system in the particular conditions of steadiness and of the absence of external mass (body) forces. Let L , v_c , p_c be respectively, a reference length, velocity and pressure which are characteristic for the envisaged problem.

Let us now make a variable and function change

$$x_i = L \bar{x}_i, u_i = v_c \bar{u}_i, p = p_c \bar{p},$$

which transforms the equation of continuity into $\frac{v_c}{L} \bar{u}_{i,\bar{i}} = 0$, that is $\bar{u}_{i,\bar{i}} = 0$, while the flow equations become

$$\bar{u}_j \bar{u}_{i,\bar{j}} + \frac{p_c}{\rho v_c^2} \bar{p}_{,\bar{i}} = \frac{\nu}{v_c L} \bar{u}_{i,\bar{j}\bar{j}}.$$

The variables and the functions with “bar” will be called *reduced* and the corresponding resultant system of equations is called *the reduced system*. Within this system two nondimensional coefficients arise:

– the first, $\frac{p_c}{\rho v_c^2}$, is not connected to any interesting feature of the (solution) system and that is why we avoid it by choosing $p_c = \rho v_c^2$ (it is possible to make such a choice because the pressure interferes only by its derivatives, which does not happen in the compressible case);

– the second will be the inverse of the Reynolds number $R = \frac{v_c L}{\nu} = \frac{\rho v_c L}{\mu}$ and it characterizes the weight of the viscosity effects (ν) versus those caused by the inertia ($v_c L$).

In this way the reduced system can be written

$$\bar{u}_{i,\bar{i}} = 0,$$

$$\bar{u}_j \bar{u}_{i,\bar{j}} + \bar{p}_{,\bar{i}} = \frac{1}{R} \bar{u}_{i,\bar{j}\bar{j}}.$$

Let there now be a solution of this system (considered for R fixed), namely

$$\bar{u}_i = \bar{u}_i(\bar{x}_i, R), \quad \bar{p} = \bar{p}(\bar{x}_i, R).$$

To this solution there corresponds, by the formulas of variable and function change, a family of solutions for the Navier–Stokes equations and the equation of continuity, a family which depends on four parameters L, v_c, ρ, μ , linked by the condition that R should be fixed (therefore only three parameters are independent).

Hence there is the following family of solutions (associated to a solution of the reduced system)

$$u_i(x_1, x_2, x_3; L, v_c, \rho, \mu) = v_c \bar{u}_i\left(\frac{x_1}{L}, \frac{x_2}{L}, \frac{x_3}{L}; R\right)$$

and

$$p(x_1, x_2, x_3; L, v_c, \rho, \mu) = \rho v_c^2 \bar{p}\left(\frac{x_1}{L}, \frac{x_2}{L}, \frac{x_3}{L}; R\right).$$

The fluid flows which correspond to such a family of solutions, for the same fixed R , are called *similar flows*.

Obviously, if \bar{u}_i are zero on a surface of equation $F(\bar{x}_1, \bar{x}_2, \bar{x}_3) = 0$, any similar flow satisfies also the adherence condition along the surface $F\left(\frac{x_1}{L}, \frac{x_2}{L}, \frac{x_3}{L}\right) = 0$ which represents the equation of the boundary of an obstacle immersed in the fluid.

Besides its exceptional theoretical importance (connected with the construction of a class of solutions of the Navier–Stokes system which depend on three free parameters), the nondimensionalized reduced system is the system we deal with in view of the use of the numerical algorithms and implicitly to simulate the fluid flows on the computer.

We cannot also forget that the similarity principle for the fluid flows backs the laboratory experience on prototypes (as those made in an aerodynamical tunnel) and when, by starting with the measurements performed in some particular conditions, it is possible to anticipate the results in much more general conditions provided that the Reynolds number is constant.

7. Flows With Low Reynolds Number. Stokes Theory

Let $\bar{u}_i = \bar{u}_i(\bar{x}_i, R)$ and $\bar{p} = \bar{p}(\bar{x}_i, R)$ be a solution of the reduced system for a certain fixed R .

Suppose now that we make $R \rightarrow 0$ while \bar{x}_i are fixed. Denoting by $\bar{u}_i^{(1)}$ and $\bar{p}^{(1)}$ the main parts of \bar{u}_i and \bar{p} respectively, it is shown that the following asymptotic behaviours hold, namely

$$\bar{u}_i = \bar{u}_i^{(1)}(\bar{x}_j) + o(f),$$

$$\bar{p} = R^{-\alpha} \bar{p}^{(1)}(\bar{x}_j) + o(R^{-\alpha}),$$

α being a real number (not determined yet) while the notation $o(f)$ designates infinitely small quantities with respect to f .

Using these developments in the reduced system and neglecting those terms which are of higher order (in the small parameter R) than the kept terms, we get

$$\bar{u}_{i,\bar{i}}^{(1)} = 0, \quad R^{1-\alpha} \bar{p}_{,\bar{i}}^{(1)} = \bar{u}_{i,\bar{j}\bar{j}}^{(1)}.$$

It is obvious that only the choice $\alpha = 1$ allows us to watch the problem in what follows (i.e., to keep the maximum number of the unknown functions), such that we are led to the system (we will now omit the writing of superscripts)

$$\bar{u}_{i,\bar{i}} = 0, \quad \bar{p}_{,\bar{i}} = \bar{u}_{i,\bar{j}\bar{j}}.$$

This linear system is the Stokes system for steady flows. By applying the divergence operator to the second equation, we also have $\Delta \bar{p} = 0$, which means the pressure is a harmonic function within this model.

We now remark that if the flow is plane or axially symmetric, there will be a stream function ψ which allows us to express both components of velocity (that is u and v) with the help of this unique function so we have (see also Chapter 1)

$$u = \frac{1}{y^m} \frac{\partial \psi}{\partial y}; \quad v = -\frac{1}{y^m} \frac{\partial \psi}{\partial x}$$

($m = 0$ in the plane case and $m = 1$ in the revolution case).

Nondimensionalizing the steady Navier–Stokes system, starting now from the rotation (vorticity) equation

$$\text{rot}(\boldsymbol{\omega} \times \mathbf{v}) = \nu \Delta \boldsymbol{\omega} \quad (\boldsymbol{\omega} = \omega \mathbf{k} \quad \text{and} \quad \omega = \frac{1}{2} \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right)),$$

by

$$x = L\bar{x}, y = L\bar{y}, u = U_c \bar{u}, v = V_c \bar{v}$$

and corresponding

$$\psi = U_c L^{1+m} \bar{\psi}, \omega = U_c L^{-1} \bar{\omega},$$

we are led (keeping only the main parts in $\bar{\psi}$) to the system

$$\frac{\partial^2 \bar{\psi}}{\partial \bar{x}^2} + \frac{\partial^2 \bar{\psi}}{\partial \bar{y}^2} - \frac{m}{\bar{y}} \frac{\partial \bar{\psi}}{\partial \bar{y}} = -2\bar{\omega} \bar{y}^m,$$

$$\frac{\partial^2 (\bar{\omega} \bar{y}^m)}{\partial \bar{x}^2} + \frac{\partial^2 (\bar{\omega} \bar{y}^m)}{\partial \bar{y}^2} - \frac{m}{\bar{y}} \frac{\partial (\bar{\omega} \bar{y}^m)}{\partial \bar{y}} = 0.$$

The last equation could be also found in the study of the stream function of an inviscid incompressible irrotational fluid flow (Chapter 1). If $\bar{\omega} \bar{y}^m$ is determined, from the second equation, then the first equation allows us to define the function $\bar{\psi}$.

Obviously, in the plane case ($m = 0$), the stream function will be a biharmonic function, that is $\Delta(\Delta \bar{\psi}) = 0$.

Unfortunately the Stokes model which is elliptic in the steady case while it is parabolic in the unsteady one, fails at large distances from the immersed obstacle [33]. This result, known also as *Stokes paradox*, could be proved, in an elegant manner, in the case of the flow past a circular cylinder by pointing out the impossibility of such a steady flow with a nonzero constant velocity at far field [153]. Basically this paradox means

that, irrespective how small is the flow velocity at infinity (at far field), the nonlinear term of the Navier–Stokes system (which is neglected in the Stokes model) cannot be considered small enough vis-a-vis the other terms (uniformly, in the whole cylinder outside). Or, in other words, the Navier–Stokes equations should be considered, basically, nonlinear.

In fact, even if we study the three-dimensional flow past a sphere using the Stokes model, a serious deviation versus the experiment arises at a sufficiently large distance from the sphere. An explanation of this weakness consists in the fact that the simplification considered within the Stokes model is rigorous only if terms $\bar{u}_{i,\bar{j}\bar{j}}$ and $\bar{u}_j\bar{u}_{i,\bar{j}}$ are of the same magnitude order. But at far field such a situation does not always occur (for instance in the case of the sphere, the terms $\bar{u}_{i,\bar{j}\bar{j}}$ are always of the order $o\left(\frac{1}{r^3}\right)$ while the terms $\bar{u}_j\bar{u}_{i,\bar{j}}$ are of order $o\left(\frac{1}{r^2}\right)$). To overtake this inconvenience when we study the fluid flow at large distances, a good suggestion is to choose a reference length L sufficiently great (of the order of the distance between the obstacle and the far points) such that, even in the case of slow flows with high viscosity, the Reynolds number does not become small. Considering then a new variable and function change defined by $\tilde{x}_i = R\bar{x}_i$, $\tilde{u}_i = \bar{u}_i$, $\tilde{p} = \bar{p}$ where \tilde{x}_i are kept constant while $\bar{x}_i \rightarrow \infty$, the initial system

$$\bar{u}_{i,i} = 0, \quad \bar{u}_j\bar{u}_{i,\bar{j}} + \bar{p}_{,\bar{i}} = \frac{1}{R}\bar{u}_{i,\bar{j}\bar{j}}$$

will be rewritten in the form

$$\tilde{u}_{i,\bar{i}} = 0, \quad \tilde{u}_j\tilde{u}_{i,\bar{j}} + \tilde{p}_{,\bar{i}} = \tilde{u}_{i,\bar{j}\bar{j}},$$

a system which, by keeping its non-linearity, does not differ essentially from the Navier–Stokes system.

If we accept that the far field (stream) velocity is parallel with the Ox_1 axis (this means its nondimensionalized components are given by δ_{i1}), the solution of the above system will be sought under the form

$$\tilde{u}_i = \delta_{i1} + R\tilde{u}'_i + \dots \quad \text{and} \quad \tilde{p} = R\tilde{p}' + \dots$$

where \tilde{u}'_i and \tilde{p}' are the main parts of the perturbation terms associated to the presence of the obstacle. Finally, by using these expansions in the above equations and eliminating the terms of higher order in the small parameter R , we arrive at the linear system

$$\tilde{u}'_{i,\bar{i}} = 0, \quad \tilde{u}'_{j,\bar{1}} + \tilde{p}'_{,\bar{i}} = \tilde{u}'_{i,\bar{j}\bar{j}},$$

known as the *Oseen system*.

This system is different from the previous Stokes system only by the presence of the term $u_{i,\bar{i}}$. But just the presence of this term allows us to avoid the Stokes paradox, i.e., it becomes possible to study flow at large distances.

Before stating some appreciation about the Oseen model (system) within a known problem, we remark once again that, the pressure \tilde{p}' is a harmonic function (which we get directly by applying the divergence operator to the second equation).

Let then Φ be a harmonic function in \tilde{x}_i such that $\tilde{p}' = \frac{\partial \Phi}{\partial \tilde{x}_i}$. If we introduce, instead of the function \tilde{u}'_i , the function $v_i = \tilde{u}'_i + \Phi_{,\bar{i}}$, then this new function v_i will satisfy

$$v_{i,\bar{i}} = 0, \quad v_{i,\bar{j}\bar{j}} - v_{i,\bar{i}} = 0,$$

while

$$\tilde{p}_{,\bar{i}} = \Phi_{,\bar{i}\bar{i}}, \quad \Phi_{,\bar{j}\bar{j}} = 0,$$

a system whose unknowns are “separated”.

7.1 The Oseen Model in the Case of the Flows Past a Thin Profile

Let us consider the plane flow of a viscous incompressible fluid with a uniform (parallel to Ox) velocity \mathbf{v}_∞ at far field in the presence of a thin obstacle (profile) whose sketch in the flow plane Oxy is the smooth arc C of continuously differentiable equations $x = x(s)$, $y = y(s)$, $s \in [0, l]$.

Following [33] we accept, if the perturbations induced by the presence of the profile are respectively \mathbf{v}' and p' , that the looked for velocity and pressure fields have the representations

$$\mathbf{v} = v_\infty (\mathbf{i} + \mathbf{v}'), \quad p = p_0 + \rho v_\infty^2 p'$$

where v_∞ is the velocity magnitude at far field, ρ is the constant density and p_0 is the pressure in the unperturbed flow.

Nondimensionalizing, by the introduction of the new variables and functions as follows,

$$x' = L^{-1}x, \quad y' = L^{-1}y \quad \text{and} \quad \rho' = L^{-1}\rho,$$

in the hypothesis of the steadiness and by neglecting the perturbations of higher order, we get a new system for perturbed velocity and pressure, namely

$$\frac{\partial \mathbf{v}}{\partial x} = -\text{grad} p + \frac{1}{R} \Delta \mathbf{v}, \quad \text{with} \quad R = \frac{Lv_\infty}{\nu}$$

(where we have omitted using the “prime” superscript symbol for the perturbed entities, a convention which is kept in the sequel).

But this equation corresponds to the Oseen approximation and it will be completed with the equation of continuity

$$\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} = 0$$

together with the (nondimensionalized) no-slip condition on C ,

$$v_1|_C = -1, \quad v_2|_C = 0$$

and the behaviour condition at far field (infinity)

$$\lim_{r \rightarrow \infty} (\mathbf{v}, p) = 0.$$

As p is a harmonic function, if $q(x, y)$ designates a harmonic conjugate function of it, then $p(x, y) + iq(x, y)$ will be an analytic function whose development in the neighborhood of infinity is of the form

$$p(x, y) + iq(x, y) = \frac{\alpha_1}{z} + \frac{\alpha_2}{z^2} + \dots$$

Let there now be a holomorphic function $P(x, y) + iQ(x, y)$ whose derivative is equal to $R(p + iq)$. According to the derivative definition for such a function we have that P and Q should satisfy the system

$$p = \frac{1}{R} \frac{\partial P}{\partial x} = \frac{1}{R} \frac{\partial Q}{\partial y}, \quad q = \frac{1}{R} \frac{\partial Q}{\partial x} = -\frac{1}{R} \frac{\partial P}{\partial y}.$$

On the other side the stream function $\psi(x, y)$ whose existence is assured by the continuity equation ($v_1 = \frac{\partial \psi}{\partial y}$ and $v_2 = -\frac{\partial \psi}{\partial x}$), is a constant which, vanishing at infinity, is necessarily zero everywhere.

So we are led to the equation

$$\Delta \psi = R \frac{\partial \psi}{\partial x} + \frac{\partial Q}{\partial x}$$

which, completed with $\Delta Q = 0$, would provide a system we deal with in the flow domain.

Considering now the auxiliary functions $\tilde{\psi}$ and \tilde{Q} , defined by

$$Q = 2\tilde{Q}, \quad \frac{1}{2R}\psi = -\tilde{Q} + \tilde{\psi},$$

the above system can be decomposed into the independent equations

$$\Delta \tilde{Q} = 0, \quad \Delta \tilde{\psi} - R \frac{\partial \tilde{\psi}}{\partial x} = 0$$

to which one adds the vanishing conditions at far field of the type

$$\lim_{r \rightarrow \infty} \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right) (\tilde{Q}, \tilde{\psi}) = 0.$$

Let $P_t(x(t), y(t))$ and $P_s(x(s), y(s))$ be two points of the contour C . Since the arc C , swept as s increases from an origin O , is smooth either the distances ρ between these points or the argument α of the joining chord (which means the angle made with the positive sense of Ox) are continuous functions.

Let

$$\xi + i\eta = x(s) - x(t) + i[y(s) - y(t)] = \rho^{i\alpha sgn(s-t)}$$

and analogously

$$\tilde{z} \equiv \tilde{x} + i\tilde{y} = x - x(t) + i[y - y(t)] = \tilde{r}e^{i\tilde{\theta}}$$

where $z = x + iy$ is the affix of a point $M(x, y)$ from the flow plane.

Obviously, the continuity of $\tilde{\theta}(x, y, t)$ requires us to avoid its “growth” which could arise by a complete rotation around P_t , which means we should consider a suitable cut in the flow plane, as for instance the half-straight line $M_\infty P_s P_t$ (M_∞ being the infinity point of the flow plane).

Further, if $M \rightarrow P_s \in C$, and we also have that

$$\tilde{\theta} = \begin{cases} \alpha & , \text{ if } s > t \\ \alpha \pm \pi & , \text{ if } s < t \end{cases} ,$$

the sign \pm corresponding to the “right”, respectively “left”, boundary value.

Let there now be a holomorphic function $Log \tilde{z} = \ln \tilde{r} + i\tilde{\theta}$ where \tilde{r} and $\tilde{\theta}$ satisfy the Cauchy - Riemann system

$$\frac{\partial \tilde{\theta}}{\partial x} = \frac{\partial}{\partial y} \ln \frac{1}{\tilde{r}}, \quad \frac{\partial \tilde{\theta}}{\partial y} = -\frac{\partial}{\partial x} \ln \frac{1}{\tilde{r}}.$$

Our intention is to search the solution $\tilde{Q}(x, y)$ of the modified Oseen system in the form

$$\tilde{Q}(x, y) = \int_0^1 \left[A_1(t)\tilde{\theta} + B_1(t) \ln \frac{1}{\tilde{r}} \right] dt$$

which is obviously a harmonic function while the arbitrary functions A_1 and B_1 are to be determined through the fulfilment of the boundary conditions.

Concerning the second equation of the Oseen system, by the change of function $\tilde{\psi} = Fe^{\frac{R}{2}x}$ it becomes, in the new unknown function F , a Helmholtz equation $\Delta F - \frac{R^2}{4}F = 0$. But it is well known that this Helmholtz equation admits the solution $K_0\left(\frac{Rr}{2}\right)$ where K_0 is the Bessel function of imaginary argument, of the second kind and zero order. If we now introduce the function $\Theta\left(\frac{Rr}{2}, \theta\right)$, connected with K_0 by the system

$$\begin{aligned} \frac{\partial \Theta}{\partial x} &= \frac{\partial}{\partial y} K_0\left(\frac{Rr}{2}\right) - \frac{R}{2} K_0\left(\frac{Rr}{2}\right), \\ \frac{\partial \Theta}{\partial y} &= -\frac{\partial}{\partial x} K_0\left(\frac{Rr}{2}\right) + \frac{R}{2} K_0\left(\frac{Rr}{2}\right), \end{aligned}$$

this function will also verify the above Helmholtz equation, that is⁵

$$\Delta \Theta - \frac{R^2}{4} \Theta = 0.$$

Consequently, the solution of the Oseen equation $\Delta \tilde{\psi} - R \frac{\partial \tilde{\psi}}{\partial x} = 0$ could be represented in the form

$$\tilde{\psi}(x, y) = \int_0^1 \left[A_2(t) \Theta\left(\frac{R}{2} \tilde{r}, \theta\right) + B_2(t) K_0\left(\frac{R}{2} \tilde{r}\right) \right] e^{\frac{R}{2} \tilde{x}} dt,$$

the functions $A_2(t)$ and $B_2(t)$ being distinguished via the boundary conditions.

By introducing also the Bessel function of imaginary argument, of the second kind and the first order, that is $K_1(z)$, which is linked to the previous function $K_0(z)$ by the relation $K'_0(z) = -K_1(z)$ and denoting by $C_1 = A_1 + iB_1$ and $C_2 = A_2 + iB_2$, it is shown that [33]

$$\begin{aligned} v_1 - iv_2 &= -\frac{2}{R} \left(\frac{\partial}{\partial y} + i \frac{\partial}{\partial x} \right) \tilde{Q} + \frac{2}{R} \left(\frac{\partial}{\partial y} + i \frac{\partial}{\partial x} \right) \tilde{\psi} \\ &= -\frac{2}{R} \int_0^1 \bar{C}_1 \frac{e^{-i\tilde{\theta}}}{\tilde{r}} dt + \int_0^1 e^{\frac{R}{2} \tilde{x}} \left[C_2 K_0\left(\frac{R}{2} \tilde{r}\right) + \bar{C}_2 K_1\left(\frac{R}{2} \tilde{r}\right) e^{-i\tilde{\theta}} \right] dt. \end{aligned}$$

⁵The explicit expression of this function Θ , the K'_0 's conjugate, can be found in the literature [154].

By imposing now the adherence condition, either we approach the arc (C) from its left or from its right, that is $v_1 - iv_2 = -1$; using the Plemelj “jump” theorems for the potential of double layer, we obtain the singular integral equation (but which can be reduced to a Fredholm equation with a continuous kernel) of the problem

$$\oint_0^1 \bar{C}(t) \frac{dt}{\xi + i\eta}$$

$$-\frac{R}{2} \oint_0^1 \left[C(t)K_0 \left(\frac{R}{2}\rho \right) + \bar{C}(t) \frac{\rho}{\xi + i\eta} K_1 \left(\frac{R}{2}\rho \right) \right] e^{\frac{R}{2}\xi} dt = \frac{R}{2},$$

where $C = A + iB$ with $A = A_1 = A_2$ and $B = -B_1 = B_2$ (equalities previously proved) while \oint denotes the integral considered in the Cauchy (principal) sense.

When (C) is a flat plate without any incidence (that is, placed on the Ox axis), the equations of this profile are $x = s$, $y = 0$ ($s \in [0, 1]$), $\xi = s - t$, $\eta = 0$, $\rho = |s - t|$ and $\tilde{z} = x - t + iy = \tilde{r}e^{i\theta}$.

Remarking that the factors which multiply the unknown C , in the integral equation of the problem, are zero, by separating the real and imaginary part of this equation we obtain either an equation which has only the trivial solution $B(t) \equiv 0$ or the integral equation of first kind

$$\oint_0^1 A(t)H(s - t)dt = k$$

where

$$H = \frac{1}{s - t} - \frac{R}{2} e^{\frac{R}{2}(s-t)} \left[K_0 \left(\frac{R}{2}|s - t| \right) + \operatorname{sgn}(s - t) K_1 \left(\frac{R}{2}|s - t| \right) \right].$$

Supposing that the Reynolds numbers are low, the singular kernel $H(s - t)$ could be approximated by $\frac{R}{2} \ln \frac{\frac{R}{2}|s-t|}{2} + \frac{R}{2}(\gamma - 1)$, γ being the Euler constant. Consequently the above integral equation becomes

$$\int_0^1 A(t) [\ln 4|s - t| + a - 1] dt = 1$$

where $a = \gamma + \ln \frac{R}{16}$ and whose solution, given by T. Carleman, is

$$A(t) = \frac{1}{\pi (a - 1) \sqrt{t(1 - t)}}.$$

More details on this Oseen system approach can be found in [33].

8. Flows With High (Large) Reynolds Number

If we look again at the reduced system

$$\bar{u}_{i,\bar{i}} = 0, \bar{u}_j \bar{u}_{i,\bar{j}} + \bar{p}_{,\bar{i}} = \frac{1}{R} \bar{u}_{i,\bar{j}\bar{j}} \equiv \frac{1}{R} \bar{\Delta} \bar{\mathbf{u}}, \bar{\mathbf{u}}|_{\partial D} = 0$$

where now the Reynolds number $R = \frac{U_c L}{\nu}$ is supposed large (which could be done also for ν small) a legitimate question will be whether or not the solution of this system is “close” to that of the corresponding Euler system, for the same flow domain, that is to

$$\bar{u}_{i,\bar{i}} = 0, \bar{u}_j \bar{u}_{i,\bar{j}} + \bar{p}_{,\bar{i}} = 0, \bar{\mathbf{u}} \cdot \mathbf{n}|_{\partial D} = 0.$$

In other words, the fact that $\frac{1}{R} \bar{\Delta} \bar{\mathbf{u}} \rightarrow 0$ would imply the convergence of the Navier–Stokes system solution to the corresponding solution of the ideal incompressible fluid (Euler) equations?

We will see that the presence of this viscosity term $\frac{1}{R} \bar{\Delta} \bar{\mathbf{u}}$, irrespective of how small it is, besides retaining of the second order character for the Navier–Stokes system, together with the adherence condition (obviously more complete than the slip condition for inviscid fluids) will determine:

1. The “Procrustian” differentiation of the fluid flow governed by the Navier–Stokes equations (vis-a-vis the flow associated to the Euler equations) in the proximity (vicinity) of the boundary ∂D in a region whose “thickness” is in inverse variation with R .

2. The mentioned region where this differentiation occurs and which persists irrespective of how small R is, could be even separated from the boundary, this separation acting as a source of vorticity.

So that completely new circumstances will arise and they will be fundamental to understanding the limits of the inviscid fluid model, which means the extent to which one could use with good results the hypotheses (schemes) already introduced for this inviscid fluid.

For a better understanding of these ideas we start our study with a simple mathematical model where one analyses the relationship between the solution of the second order differential equation with Dirichlet (biloc) condition and the solution of a Cauchy problem for that first order differential equation which is the “limit” of the first equation when the small parameter $\varepsilon \rightarrow 0$. The conclusions obtained from this abstract mathematical model will be extended to the parallelism between the Navier–Stokes and Euler equations.

8.1 Mathematical Model

Let us consider the second order differential equation

$$\varepsilon f''(x) + f'(x) = a, \quad x \in [0, 1]$$

where $a \in (0, 1)$ and ε is a small positive parameter, to which we join the boundary conditions

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

It is known that the unique solution of this bilocal problem is

$$f(x, \varepsilon) = (1 - a) \frac{1 - e^{-\frac{x}{\varepsilon}}}{1 - e^{-\frac{1}{\varepsilon}}} + ax.$$

Suppose now that, in this solution, we make $\varepsilon \rightarrow 0$ and so we have

$$\lim_{\varepsilon \rightarrow 0} f(x, \varepsilon) \equiv f_0(x) = 1 + a(x - 1)$$

for $x \in (0, 1]$.

A questionable aspect would be the rapport between this limit and the “limit” of the differential equation resulting from the given equation when $\varepsilon \rightarrow 0$, which means the differential equation $f' = a$. In fact $f_0(x)$ will be a solution of the differential equation $f' = a$, more precisely that solution which satisfies the prescribed condition $f_0(1) = 1$ but it does not at the point 0 where $f_0(0) = 1 - a \neq 0$.

In other terms the convergence of $f(x, \varepsilon)$ to $f_0(x)$ when $\varepsilon \rightarrow 0$, is *nonuniform* in the interval $[0, 1]$ and in the neighborhood of zero $f_0(x)$ cannot be considered a correct approximation for the exact solution $f(x, \varepsilon)$ of the initially given bilocal problem for the second order differential equation.

To get a correct approximation of $f(x, \varepsilon)$ in the neighborhood of $x = 0$ we will use a special technique (the “ordinates dilatation”). More precisely, we perform a change of variable $x = \varepsilon \xi$ and then we make $\varepsilon \rightarrow 0$ but keeping $\xi = \frac{x}{\varepsilon}$ to be a constant. So that we obtain

$$\lim_{\substack{\varepsilon \rightarrow 0 \\ (\xi \text{ fixed})}} f(x, \varepsilon) \equiv \tilde{f}_0(\xi) = (1 - a) \left(1 - e^{-\xi}\right).$$

This new limit function $\tilde{f}_0(\xi)$ will be the solution of that differential equation got from the initial one by the change of variable and function $x = \varepsilon \xi$, $f(x, \varepsilon) = \tilde{f}(\xi, \varepsilon)$ and then keeping only the main (of the highest order in ε) terms of it, i.e., of the differential equation $\tilde{f}'' + \tilde{f}' = 0$. We can also see that $\tilde{f}(0) = 0$ and $\lim_{\xi \rightarrow \infty} \tilde{f}_0(\xi) = 1 - a = f_0(0)$, that is a

“matching condition” of the two approximations holds. Obviously the just found function $\tilde{f}_0\left(\frac{x}{\varepsilon}\right)$ represents a good approximation of $f(x, \varepsilon)$ (for ε enough small) in the neighborhood $V(0)$ of the origin (where a boundary condition is lost) while $f_0(x)$ will be a correct approximation for the same function in the complement of the previous region, that is for $x \notin V(0)$.

This simple model could be a guide in introducing the so-called “boundary layer” which corresponds to the region where the approximation of the solution of the Navier–Stokes system through the corresponding Euler solution is not possible. In fact the Navier–Stokes system, with a high Reynolds number, plays the role of the above second order differential equation with $\varepsilon = \frac{1}{R}$, the immediate proximity of the wall (obstacle) corresponds to $V(0)$ and the Euler equation takes the place of the “limit” equation $f' = a$ (when $\varepsilon = \frac{1}{R} \rightarrow 0$).

To get a correct approximation of the Navier–Stokes equations in the vicinity of the obstacle (wall), where the solution of the Euler equation fails (replacing also the adherence condition by the much less rigorous slip condition), one performs again a change of variables and functions (the “ordinate dilatation”) making then $\varepsilon \rightarrow 0$ such that the new just introduced variables keep their constancy. Finally, considering only the main terms in $\varepsilon = \frac{1}{R}$ (and neglecting the rest) we reach the so-called *boundary layer equations*.

As regards the solutions of the Euler system, they match with those of the boundary layer equations at a sufficiently large distance from the obstacle, i.e., on the “border” of this boundary layer whose thickness varies directly with $\varepsilon = \frac{1}{R}$, as we will see later.

The parallelism between the envisaged mathematical model and the approximation of the Navier–Stokes system by the Euler and boundary layer equations is illustrated also in Figure 3.1.

8.2 **The Boundary Layer Equations**

Our purpose is now to determine explicitly the boundary layer equations in the conditions of the existence of an obstacle which could be identified with the positive real semiaxis (the half-infinite flat plate) and which is placed in a viscous incompressible fluid stream with a velocity $\mathbf{v}_\infty(U_\infty, 0)$ at far field. Obviously the same problem for an ideal fluid (a uniform flow) leads, in nondimensional variables, to the solution $\bar{u} = 1, \bar{v} = 0, \bar{p} = 1$, but this solution does not approximate the viscous fluid flow in the boundary layer formed in the proximity of the wall.

To determine the boundary layer equations, we should set up a change of variable and function that implies the “coordinates dilatation” (in this case an “ordinates dilatation”) and then we keep only the main terms in

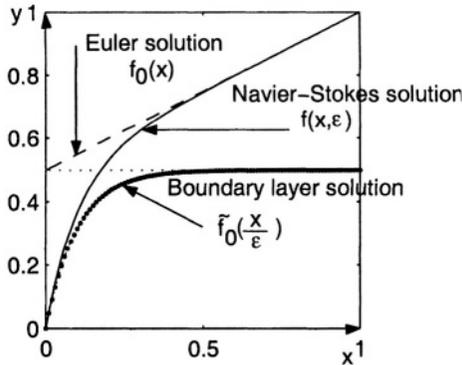


Figure 3.1. The approximation of the Navier–Stokes solutions by the Euler and boundary layer solution

$\frac{1}{R}$. More precisely, in the nondimensionalized equations of the viscous incompressible fluid, that is in

$$\bar{u}_{i,\bar{i}} = 0, \bar{u}_j u_{i,\bar{j}} + \bar{p}_{,\bar{i}} = \frac{1}{R} \bar{u}_{i,\bar{j}\bar{j}}, \quad i, j = 1, 2; \bar{i}, \bar{j} = 1, 2 \text{ and } \bar{u}_1 = \bar{u}; \bar{u}_2 = \bar{v},$$

performing a change of variable and function which allows a clearer appearance of the velocity component normal to the plate (“the ordinates dilatation”), that is

$$\bar{x} = \xi, \bar{y} = R^{-\alpha} \eta, \bar{u} = \tilde{u}, \bar{v} = R^{-\alpha} \tilde{v}, \bar{p} = \tilde{p} \quad \text{with } \alpha > 0,$$

we obtain that (necessarily) $\alpha = \frac{1}{2}$, $\delta = R^{-\frac{1}{2}}$ being just the boundary layer “thickness” (for any other value of α , either the continuity equation would lose a term, becoming trivial, or the terms due to viscosity or those due to the acceleration quantity — from the other two equations — would disappear, in both situations the whole system becoming more “poor”). If now we make $R \rightarrow \infty$, imposing ξ and η to be constant and then keeping only the main terms in $\frac{1}{R}$, one obtains the following system of equations of boundary layer (Prandtl)⁶

$$\frac{\partial \tilde{u}}{\partial \xi} + \frac{\partial \tilde{v}}{\partial \eta} = 0,$$

⁶The boundary layer equations in the case of curved surfaces are much more complicated (see, for instance, S.L. Goldstein [56]).

$$\begin{aligned}\tilde{u} \frac{\partial \tilde{u}}{\partial \xi} + \tilde{v} \frac{\partial \tilde{u}}{\partial \eta} + \frac{\partial \tilde{p}}{\partial \xi} &= \frac{\partial^2 \tilde{u}}{\partial \xi^2}, \\ \frac{\partial \tilde{p}}{\partial \eta} &= 0,\end{aligned}$$

with the boundary conditions which express the adherence to the plate

$$\tilde{u}(\xi, 0) = \tilde{v}(\xi, 0) = 0, \xi > 0$$

together with the matching conditions with the inviscid fluid flow

$$\begin{aligned}\lim_{\eta \rightarrow \infty} \tilde{u}(\xi, \eta) &= \lim_{y \rightarrow 0} \bar{u}(x, y) = 1, \\ \lim_{\eta \rightarrow \infty} \tilde{p}(\xi, \eta) &= \lim_{y \rightarrow 0} \bar{p}(x, y) = 1.\end{aligned}$$

Obviously, the approximation through the boundary layer solution is backed by the existence of some positive constants C and α such that, in a certain norm within the velocity space, the solution of the Navier–Stokes system and the corresponding solution of the boundary layer equations satisfy an estimation of the form⁷

$$\|\bar{\mathbf{u}} - \tilde{\mathbf{u}}\| \leq \frac{C}{R^\alpha}, \text{ for } 0 \leq \bar{y} \leq \delta \text{ and } R \rightarrow \infty.$$

Before giving a brief mathematical study of the Prandtl equations we should make some remarks. First, if we evaluate the circulation along a simple contour (for instance, a rectangular one) which is tangent to the obstacle, being all the time inside the boundary layer, this circulation will vanish. Really, if our rectangular contour $ABCD$ has the side DC tangent to the obstacle at D and the other side AB is obviously parallel with it, from $\tilde{u} = 0$ on the boundary, we have also there $\frac{\partial \tilde{u}}{\partial \xi} = 0$ while the continuity equation leads to $\frac{\partial \tilde{v}}{\partial \eta} = 0$. Thus, since $\tilde{v} = 0$ on the boundary we could suppose that \tilde{v} is small in the proximity of the boundary or more specifically, \tilde{v} is small compared with the value of \tilde{u} along AB while \tilde{u} is near zero along DC . So we have

$$\int_{ABCD} \tilde{\mathbf{v}} \cdot d\mathbf{r} = \int_{DA} \tilde{v} d\eta + \int_{AB} \tilde{u} d\xi - \int_{CB} \tilde{v} d\eta - \int_{DC} \tilde{u} d\xi \approx \int_{AB} \tilde{u} d\xi > 0.$$

⁷There are very few, and only in particular cases, mathematical results on such estimations. Concerning the existence and uniqueness theorems we should mention O.A. Olejnik [96] and P. C. Fife [44] who have shown, under some assumptions, the existence of such an estimation for $\alpha = \frac{1}{2}$.

Implicitly, there will be a source of vorticity, the existence of the boundary layer being associated with a mechanism for producing vorticity in the boundary vicinity.

Experimentally, we can see that, when a boundary layer arises in the neighborhood of an obstacle and an “unfavourable” pressure gradient⁸ occurs, there is a point C where this boundary layer is separating from the obstacle, between the upper delimitation border of the boundary layer and the obstacle surface some *inverse flows* being possibly formed. This separation will be a vorticity source which propagates in the boundary layer which could support the *almost potential* fluid flows model (see the previous chapter), the separating vorticity lines being considered as emanating from the separation points of the boundary layer. It would be plausible to identify the separation points with those points where the vorticity vanishes although there are no mathematical results to support this assertion.

The second matching condition, together with the last equation, shows that $\tilde{p}(\xi, \eta) = 1$, which means the pressure is constant inside the boundary layer and its value equals that of the pressure of the ideal fluid in the adjacent flow.

As a consequence of this remark, $\frac{\partial \tilde{p}}{\partial \xi} = 0$ and the Prandtl system will contain only the velocity components \tilde{u} and \tilde{v} . But the continuity equation (the compressibility condition) allows then the construction of a stream function $\tilde{\psi}$ such that $d\tilde{\psi} = \tilde{u}d\eta - \tilde{v}d\xi$ and so the boundary layer system could be rewritten, in the unique unknown $\tilde{\psi}$, as

$$\frac{\partial^3 \tilde{\psi}}{\partial \eta^3} - \frac{\partial \tilde{\psi}}{\partial \eta} \frac{\partial^2 \tilde{\psi}}{\partial \xi \partial \eta} + \frac{\partial \tilde{\psi}}{\partial \xi} \frac{\partial^2 \tilde{\psi}}{\partial \eta^2} = 0,$$

an equation to which one should attach the conditions

$$\tilde{\psi}(\xi, 0) = \frac{\partial \tilde{\psi}(\xi, 0)}{\partial \eta} = 0, \quad \lim_{\eta \rightarrow \infty} \frac{\partial \tilde{\psi}}{\partial \eta} = 1.$$

To construct the solution of this third order nonlinear partial differential equation, we remark that if $\tilde{\psi}(\xi, \eta)$ is a solution of this equation,

⁸The “unfavourable” pressure gradients are correlated with a pressure increasing in the flow direction which leads to a slower fluid flow in the boundary layer together with an accentuated slenderness of this one, all of them determining the formation of a rest region where a slow inverse flow could arise. As the main fluid stream should avoid this quite significant zone and thus determine the boundary layer separation, in this case we can’t make an exact assesment of the adjacent inviscid flow. In the conditions of the “favourable” pressure gradients, the decrease of the pressure in the sense of the flow together with the continuous slenderness of the boundary layer, make that the outer inviscid fluid model will be not affected anymore and this inviscid model could be “added” without any difficulties.

the same thing happens with the functions $\frac{a}{b}\tilde{\psi}\left(\frac{\xi}{a}, \frac{\eta}{b}\right)$ for any constants a and b . In the particular case when these constants are linked through a relation of the form $b = a^n$ (n rational), together with the solution $\tilde{\psi}(\xi, \eta)$ we also have the class of solutions $a^{1-n}\tilde{\psi}\left(\frac{\xi}{a}, \frac{\eta}{a^n}\right)$. An immediate question will be if the application $\tilde{\psi}(\xi, \eta) \rightarrow a^{1-n}\tilde{\psi}\left(\frac{\xi}{a}, \frac{\eta}{a^n}\right)$ has any fixed points, i.e., if this correspondence, by a suitable choice of the constants a and b , can lead to such functions which satisfy the equality

$$\tilde{\psi}(\xi, \eta) = a^{1-n}\tilde{\psi}\left(\frac{\xi}{a}, \frac{\eta}{a^n}\right).$$

It is shown that the necessary form of the functions $\tilde{\psi}(\xi, \eta)$ to fulfil the above requirement is $\tilde{\psi}(\xi, \eta) \equiv \xi^{1-n}f\left(\frac{\eta}{\xi^n}\right)$ for any rational n [52].

On the other hand the fulfilment of the condition

$$\lim_{\eta \rightarrow \infty} \frac{\partial \tilde{\psi}}{\partial \eta} = \lim_{\eta \rightarrow \infty} \xi^{1-2n} f' \left(\frac{\eta}{\xi^n} \right) = 1,$$

implies a compulsory choice for n , namely $n = \frac{1}{2}$.

Therefore we intend to look for those solutions (of the boundary layer system) which are of the form $\tilde{\psi} = \xi^{\frac{1}{2}} f(\theta)$ with $\xi > 0$ and $\theta = \frac{\eta}{\xi^{\frac{3}{2}}}$. In the language of the function f , the Prandtl equation becomes a nonlinear ordinary differential equation

$$2f'''(\theta) + f(\theta)f''(\theta) = 0 \quad (3.2)$$

with the boundary conditions

$$f(0) = f'(0) = 0, \quad f'(\infty) = 1$$

H. Weyl formulated a successive approximations procedure which proves the existence and the uniqueness of the solution of the above equation. This solution has been exactly calculated but it presents some inconvenience. Thus for ξ small, \tilde{v} becomes infinite which could be avoided by choosing a suitable system of coordinates. At the same time, in $V(0)$ the Reynolds number $R_x = \frac{v_x}{\nu}$ (there is no reference length associated to the problem) becomes small, irrespective of how small is ν , which contradicts the basic hypothesis that the Reynolds number is always very large. In spite of all these shortcomings, which cannot be avoided in the boundary layer theory, the obtained solutions agree very well with experience at all the points outside of $V(0)$.

Figure 3.2 points out the shape of the longitudinal velocity profile $\tilde{u} = f'(\theta)$ which comes from the Weyl solution. Experience confirms these results, showing that this velocity profile tends to stabilize.

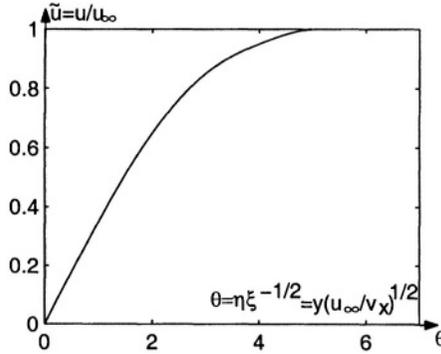


Figure 3.2. The profile of the longitudinal velocities

Before ending this section we try to give a definition of the boundary layer “thickness”, even if this concept is not very precisely stated. One accepts an understanding that the thickness corresponding to the abscissa x is that y for which $u = 0,99U_\infty$. Therefore, it corresponds to the value θ which satisfies $1 - f'(\theta) = 0.01$, which means this value should be approximately $\theta = 5$. Consequently we have

$$\delta = 5\sqrt{\frac{\nu x}{U_\infty}} = 5\sqrt{\frac{\nu}{U_\infty}},$$

that is the thickness grows together with \sqrt{x} and hence the shape of the boundary layer “border” has a parabolic shape.

The aim of this book is not to overview the analytical or “practical” methods for solving the boundary layer equations. There is a large variety of such methods but most of them are valid only in particular cases or they are not sufficiently rigorous concerning the approximations made. In fact this last remark involves many of the papers on the boundary layer theory, the practical applications imposing a “rush” for effective solutions which are not always correct from the mathematical point of view.

In what follows we will focus on a probabilistic algorithm which allows modification of the fluid flow governed by the Euler equations, in the vicinity of the boundary, in order to simulate the boundary layer effects and implicitly to get new approximations, in the same vicinity, for the solutions of the Navier–Stokes system.

8.3 Probabilistic Algorithm for the Prandtl Equations

In what follows we will describe a random procedure (due to A. I. Chorin) based on a distribution of vortex sheets that allows construction of a practical numerical algorithm for approaching the boundary layer equations.

Let us consider, first, the heat equation for an infinite rod, namely

$$v_t = \nu v_{xx}, \quad -\infty < x < \infty, t \geq 0,$$

where $v = v(x, t)$ represents the temperature in the rod and ν its conductivity.

Accepting that, at the initial moment, $v(x, 0) = \delta(x)$, $\delta(x)$ being the Dirac distribution, then the (distributional) solution of this equation is the fundamental (Green) solution given by

$$G(x, t) = \frac{1}{\sqrt{4\pi\nu t}} \exp\left(-\frac{x^2}{4\nu t}\right).$$

This solution could be probabilistically interpreted in two ways:

1) Fix the time t and place N particles, of mass $\frac{1}{N}$, at the origin $x = 0$. Suppose that these particles “jump” so that the associated random variables follow the Gaussian distribution with mean zero and variance $2\nu t$. Thus, the probability that such a particle will “land” between x and $x + dx$ is the Gauss probability density function multiplied by dx (the length of the landing interval), precisely $\frac{1}{\sqrt{4\pi\nu t}} \exp\left(-\frac{x^2}{4\nu t}\right) dx$.

If we repeat this with a very large number of particles (provided that their total mass is unity), then, according to the central limit theorem, the probability density function of the arithmetic average of the associated independent Gaussian random variables when their number increases indefinitely, converges to the probability density function of the individual Gaussian distribution considered above;

2) Let us split up the time interval $[0, t]$ into l subintervals, each of them with length $\Delta t = \frac{t}{l}$, and consider the following procedure in a step by step manner.

Again let us place the N particles of mass $\frac{1}{N}$, at the origin, but now at $t = 0$ too. Suppose that these particles will undergo a random walk, more precisely, the position of the i^{th} particle at the moment $m\Delta t$ ($i = 1, \dots, N$; $m = 1, \dots, l$) is

$$x_i^{m+1} = x_i^m + \eta_i^m, \quad (x_i^0 = 0)$$

where η_i^m are independent Gaussian random variables, each of them with mean 0 and variance $2\nu\Delta t$. The final displacement of the i^{th} particle

is the sum of its displacements and it has, obviously, a Gaussian distribution with mean zero and variance $l \times 2\nu\Delta t = 2\nu t$. Automatically the probability density function associated to one particle (its random variable) at time t , has the same structure as above and methods 1) and 2) are equivalent.

Let us recall now the same heat equation but with the initial condition $v(x, 0) = f(x)$. We know that the solution of this problem is

$$v(x, t) = \int_{-\infty}^{\infty} G(x, x', t) f(x') dx'$$

where

$$G(x, x', t) = \frac{1}{\sqrt{4\pi\nu t}} \exp\left[-\frac{(x-x')^2}{4\nu t}\right].$$

But this solution has also a probabilistic interpretation. More precisely, let us consider the N particles, starting at a random initial position x_i^0 , $i = \overline{1, N}$, and let us assign to each of them the mass $\frac{f(x_i^0)}{N}$. If we let the particles perform a random walk (as in the method 2), keeping their mass constant, then, after l steps, the expected distribution of mass for the N particles, at a real position, is given by the above $v(x, t)$ solution.

If the heat equation is considered only on the half-line $x \geq 0$, with boundary condition $v(0, t) = 0$, then the Green function for this problem is

$$G^*(x, x', t) = G(x, x', t) - G(x, -x', t)$$

with

$$G(x, x', t) = \frac{1}{\sqrt{4\pi\nu t}} \exp\left[-\frac{(x-x')^2}{4\nu t}\right].$$

As

$$G^*(0, x', t) = 0, \quad G^*(x, x', 0) = \delta(x - x')$$

and

$$\partial_t G^*(x, x', t) = \nu \partial_x^2 G^*(x, x', t),$$

the solution of the heat problem

$$v_t = \nu v_{xx}, x \leq 0, t > 0,$$

$$v(x, 0) = f(x), v(0, t) = 0,$$

is

$$v(x, t) = \int_{-\infty}^{\infty} G^*(x, x', t) f(x') dx'.$$

The probabilistic interpretation of the last result is obtained as above, by starting with N particles of mass $\frac{f(x')}{N}$ at x' and N of mass $-\frac{f(x')}{N}$ at $-x'$, and letting them all (random) walk (by, for instance, the method 2).

Random walk methods will now be applied to vortex sheets. For the sake of simplicity, let us consider the plane fluid flow in the upper half plane $y \geq 0$ and suppose that the boundary $y = 0$ (the infinite flat plate) is rigid and at rest while the free-stream velocity of magnitude U is parallel to the real axis. Let us seek that solution of the Navier–Stokes system which is parallel to the flat plate and independent on x , that is $u = u(y, t)$, $v = 0$, the pressure constancy being also ensured such that $\text{grad} p = 0$.

Obviously the appropriate Euler system solution is $(U, 0)$.

Since the Navier–Stokes equations require the boundary conditions

$$u(0, t) = 0, u(\infty, t) = U$$

and thus

$$u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} \equiv 0,$$

the Navier–Stokes system reduces to

$$\frac{\partial u}{\partial t} = \frac{1}{R} \frac{\partial^2 u}{\partial y^2}$$

or by introducing the nondimensional variables $y' = \frac{y}{L}$ and $t' = \frac{t}{T}$, to

$$\frac{\partial u}{\partial t'} = \frac{L^2}{RT} \frac{\partial^2 u}{\partial y'^2}.$$

If $\frac{L^2}{T} = 1$, then the nondimensionalized form of this equation is the same as that of the above equation and it will be the same with the boundary conditions. Accepting that the nondimensionalized equation

with appropriate boundary conditions has a unique solution, this solution must satisfy $u\left(\frac{y}{L}, \frac{t}{T}\right) = u(y, t)$ if $L^2 = T$. Picking $T = t$, $L = \sqrt{t}$, that is $u\left(\frac{y}{\sqrt{t}}, 1\right) = u(y, t)$, we can state that u depends on y and t only through the combination $\frac{y}{\sqrt{t}}$. Set $\eta = \frac{y}{2\sqrt{\nu t}}$, $\nu = \frac{1}{R}$ and $u(y, t) = Uf(\eta)$. Then the initial equation becomes the following ordinary differential equation (in the function f), with appropriate boundary conditions, more precisely

$$f''(\eta) + 2\eta f'(\eta) = 0, f(\infty) = 1, f(0) = 0.$$

But the unique solution of this bilocal problem is

$$u = \frac{2U}{\sqrt{\pi}} \int_0^\eta e^{-s^2} ds$$

where we have used the well-known result $\int_0^\infty e^{-s^2} ds = \frac{\sqrt{\pi}}{2}$.

This solution shows that there is a significant deviation from the Euler equation solution in a region near the wall (the boundary layer) whose “thickness” is proportional to $\frac{\sqrt{t}}{\sqrt{R}}$ and thus, for fixed time, the boundary layer decreases as $\frac{1}{\sqrt{R}}$.

Correspondingly, the vorticity of the flow is

$$\xi = -\frac{U}{\sqrt{\pi\nu t}} \exp\left(-\frac{y^2}{\nu t}\right),$$

satisfying the equation $\xi_t = \frac{1}{R}\xi_{yy}$.

Unfortunately the boundary conditions for vorticity are not explicit and they should be determined from the adherence conditions on the boundary.

To reconstruct this solution using random walks method, we first define a vortex sheet of strength ξ as a fluid flow parallel to the real axis Ox where the component u “jumps” by the amount ξ when y crosses a parallel line with Ox , say $y = y_0$, i.e., $u(y_0^+) - u(y_0^-) = -\xi$.

As $t \rightarrow 0^+$, the solution tends to the constant value U , for $y > 0$, while it vanishes ($u = 0$) for $y = 0$. In other words, when $t \rightarrow 0^+$ the solution approaches a vortex sheet on Ox with strength $-U$.

Let us replace this vortex sheet by N “small” vortex sheets, each of strength $-\frac{2U}{N}$. Accept that each of these smaller vortex sheets undergoes a random walk in the y direction defined by

$$y_i^{m+1} = y_i^m + \eta_i^m, (y_i^0 = 0)$$

where η_i^m are Gaussian random variables with mean zero and variance $2\nu\Delta t$ whilst $\Delta t = \frac{t}{l}$, $m = 1, 2, \dots, l$.

We state that for large N the distribution of vorticity is constructed this way and from it the function

$$u(y, t) = U + \int_y^\infty \xi(y, t)$$

satisfies the heat equation ($u_t = \frac{1}{R}u_{yy}$ and $\xi_t = \frac{1}{R}u_{yy}$). This is clear from the random walk method developed above for the heat equation. What requires additional explanation is why u satisfies the no-slip condition on the boundary. If we remark that on the average, half of the vortex sheets are above Ox and half below, we can write

$$u(0, t) = U + \int_0^\infty \xi(y, t) dt$$

or, in a discrete version,

$$u(0, t) = U + \sum_{i=1}^{N/2} \xi_i .$$

But the strength of the i^{th} vortex sheet is $\xi_i = -\frac{2U}{N}$ and therefore $u(0, t) = 0$.

The random walk method based on vortex sheets will now be extended to the solution of the Prandtl equation (in an unsteady regime) for the half-infinite flat plate (the positive real semiaxis).

The associated fluid flow (boundary layer) will be approximated at $t = 0$ by a set of N vortex sheets of finite width h , corresponding to the coordinates $x \in [x_i - \frac{h}{2}, x_i + \frac{h}{2}]$ and $y = y_i$, of strength ξ_i . To displace these vortex sheets we split up the time interval $[0, t]$ in l parts of duration $\Delta t = \frac{t}{l}$ and we advance in time (from t to $t + \Delta t$) following the algorithm:

- (i) the vortices move according to a discrete approximation of the ideal (Euler) flow;
- (ii) vorticity is added by placing new vortex sheets on the boundary so that the resultant flow satisfies the adherence condition on the boundary;
- (iii) the vortex sheets undergo a random walk as that described in the previous flat plate example to approximate the solution of the heat equation $\xi_t = \nu\xi_{yy}$, and to preserve the boundary conditions on boundary $u = v = 0$;

(iv) time is advanced by the step Δt and the procedure restarts until time t is reached.

Obviously, the number of vortex sheets will increase in time, which corresponds to the fact that vorticity is created in the boundary layer.

Let explain now step (i). It is known that the velocity component u satisfies

$$u(x, y) = u(\infty) - \int_{\infty}^y \frac{\partial u}{\partial y} dy = u(\infty) - \int_{\infty}^y \xi dy$$

or, in a discrete version, the component u of the velocity of the i^{th} vortex due to the vortex sheets, is given by

$$u(x_i, y_i) = u(\infty) - \sum \xi_j.$$

This sum is extended over all the vortex sheets such that $y_j > y_i$ and $|x_i - x_j| < \frac{h}{2}$, that is for all the vortex sheets whose “shadow” on the Ox axis contains the point (x_i, y_i) . On the other hand the incompressibility and the boundary condition $v(x, 0) = 0$ lead to

$$v(x, y, t) = v(x, 0, t) - \int_0^y u_x(x, s, t) ds = -\frac{\partial}{\partial x} \int_0^y u(x, s, t) ds.$$

This last relation determines v in terms of u^j and a corresponding (discrete) approximate evaluation could be

$$v(x, y, t) = \frac{1}{h} \left[\int_0^y u\left(x + \frac{h}{2}, s, t\right) ds - \int_0^y u\left(x - \frac{h}{2}, s, t\right) ds \right].$$

But a more useful approximation is obtained by rewriting the above relation in terms of the vortex strengths ξ_j , precisely

$$v_i(x_i, y_i, t) = \frac{1}{h} (I_+^i - I_-^i)$$

where

⁹Obviously, due to this relation, if $u(\infty)$ is prescribed we will not be allowed to prescribe $v(\infty)$ too.

$$I_+^i(x_i, y_i) = \sum_+ y_j^* \xi_j, \quad y_i^* = \min(y_j, y_i),$$

and

$$I_-^i(x_i, y_i) = \sum_- y_j^* \xi_j.$$

Here \sum_+ means the sum over all $j \neq i$ for which $|x_j - (x_i + \frac{h}{2})| < \frac{h}{2}$ and \sum_- means the sum over all j which satisfy $|x_j - (x_i - \frac{h}{2})| < \frac{h}{2}$.

We could summarize all this by saying that in step (i) of the above algorithm, the i^{th} vortex sheet is moved by

$$x_i^{m+1} = x_i^m + u_i \Delta t,$$

$$y_i^{m+1} = y_i^m + v_i \Delta t$$

where u_i and v_i are given by the respective above expressions.

The new velocity field is now determined by the same vortex sheet but considered at their new positions. This new velocity field satisfies $v = 0$ on the real axis (by construction) and also $u(\infty) = U$. Concerning the condition $u = 0$ on the real axis at the beginning of the procedure, it needs not remain so.

The aim of the second step (ii) is just to correct the boundary conditions. This may be done as follows: divide the real axis into segments of length h and, supposing that at the center P_l of one of these segments $u = u_l \neq 0$, we place at P_l one or more vortex sheets with the same sum of strengths $2u_l$, which will guarantee that, on average, $u = 0$ on the Ox axis in the new flow.

In step (iii) we add a random y component to positions (x_i, y_i) of the existing vortex sheets, precisely a Gaussian random variable η (with mean 0 and variance $2\nu\Delta t$), such that the new positions are given by

$$x_i^{m+1} = x_i^m + u_i \Delta t,$$

$$y_i^{m+1} = y_i^m + v_i \Delta t + \eta_i^m.$$

Intuitively, the vortex sheets move about in ideal flow together with a random y -component, simulating viscous diffusion. Vortex sheets newly created (to observe the boundary conditions) diffuse out from the boundary by means of the same random component y and then get “swept” downstream by the main flow.

If there is a leading edge (such as the origin in the case of the half-infinite plate situated on the positive real semiaxis), the model will be forced to create more vortex sheets at this edge in order to satisfy the adherence condition (since they are immediately swept downstream by the flow with no replacement).

Regarding the length of the vortex sheet displacement, if in the Ox direction its average is proportional to Δt , in the Oy direction the “average” jump (displacement) will be proportional to $\sqrt{\Delta t}$.

Details about the use of this model on vortex sheets can be found in the papers [20] while some theoretical aspects are treated in [21].

8.4 Example

Let us consider, as a simple problem, a semiinfinite flat plate aligned with a uniform flow of constant velocity U and of constant physical properties, including density ρ [22]. The boundary layer equations are in this case simplified to

$$\begin{aligned} u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} &= \nu \frac{\partial^2 u}{\partial y^2}, \\ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0, \end{aligned} \tag{3.3}$$

where $\nu = \mu/\rho$ is the kinematic viscosity of the fluid. From these equations we could calculate the velocity components u, v . The model is valid for the thin laminar boundary layer within an incompressible fluid but also for a compressible fluid with a velocity much slower than the speed of sound.

At any point x on the plate we have three boundary conditions — two for the first equation and one for the second — namely the non-slip conditions at the surface and the uniform flow at far distances, that is

$$u|_{y=0} = 0, \quad v|_{y=0} = 0, \quad u|_{y=\infty} = U. \tag{3.4}$$

The differential equation and the boundary conditions for f (3.2) are therefore

$$\begin{aligned} f''' + \frac{1}{2} f f'' &= 0, \\ f(0) = f'(0) &= 0, \\ f'(\infty) &= 1, \end{aligned} \tag{3.5}$$

(the Blasius problem) and the velocity components become

$$\begin{aligned} u &= U f', \\ v &= \frac{1}{2} \left(\frac{\nu U}{x} \right)^{1/2} (\eta f' - f). \end{aligned} \tag{3.6}$$

This problem could be numerically solved. First, it is transformed into a system of three first order equations

$$\begin{aligned} f' &= p, & f(0) &= 0, \\ p' &= q, & p(0) &= 0, \\ q' &= -\frac{1}{2}fq, & q(0) &= q_0, \end{aligned}$$

with q_0 not known for the moment. It will be calculated by successive numerical integrations with a Runge–Kutta method such that $p(\infty) = 1$ is satisfied.

If we have f and its derivatives p and q , we could calculate the velocity components within the boundary layer from formulas (3.6).

Let us consider a numerical example with $U = 30\text{m/s}$ and the kinematic viscosity of the air (at sea level) $\nu = 1.49 \times 10^{-5}\text{m}^2/\text{s}$. The problem (3.5) is solved by the MATLAB program

```
for i=1:10 q(i)=i/10;
[t,x]=ode45(@edstrlim,[0,10],[0 0 q(i)]);
r(i)=x(length(x),2);
end;
plot(q,r);grid;xlabel('q');ylabel('r');
which uses the function subprogram edstrlim.m
function yprim=edstrlim(x,y);
yprim=zeros(3,1);
yprim(1)=y(2);
yprim(2)=y(3);
yprim(3)=-y(1)*y(3)/2;
```

The program chooses different values for q_0 and solves the corresponding Cauchy problem. The values of r representing f' for large values ($\eta = 10$) are taken and the value $q_0 = 0.3320572$ for which $f'(10) = 1$ is found (see for example Figure 3.3).

The corresponding solution $f'(\eta)$ is represented in Figure 3.4.

The structure of the boundary layer could be now obtained by representing the components of the velocities u respectively v from the formulas (3.6). We remark that the thickness of the boundary layer (defined as the height for which $u = 0.994U$ which occurs for $\eta = 5.2$) is of the form

$$d(x) = 5.2 \left(\frac{\nu x}{U} \right)^{1/2},$$

therefore it is represented by a parabola, see Figure 3.5.

We also remark that the boundary layer thickness is about 0.37cm and the Reynolds number corresponding to this distance is $R = Ux/\nu = 2.01 \times 10^6$; the Reynolds number must be large in order to ensure the

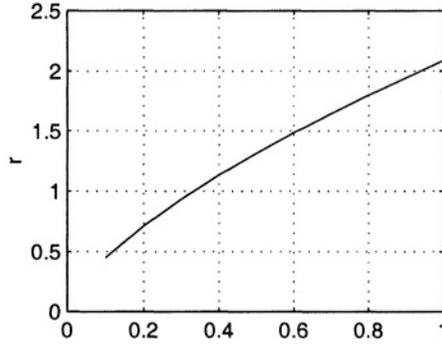


Figure 3.3. Choosing the initial condition q_0

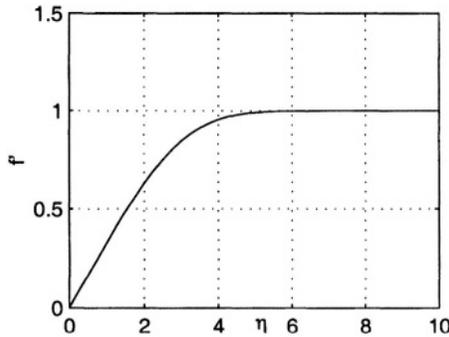


Figure 3.4. The solution of the Blasius problem. The graph of f'

validity of the boundary layer theory. Moreover, the shear stress is

$$\tau = \mu \frac{\partial u}{\partial y} = \mu U \left(\frac{U}{\nu x} \right)^{1/2} f'' ,$$

thus f'' describes the dimensionless shear stress in the boundary layer. Consequently, the particular value $f''(0) = q_0$ which is the value calculated in the program, is the dimensionless shear stress on the flat plate.

We could avoid the calculation of q_0 (which needs the successive solving of Cauchy problems on large intervals $[0, \eta]$) by using the following change of coordinates.

Let $\eta = kz$ where k is a constant that will be determined, and let g be a function associated to f through $f(\eta) = g(z)/k$. Then

$$\frac{d^n f}{d\eta^n} = \frac{1}{k^{n+1}} \frac{d^n g}{dz^n}$$

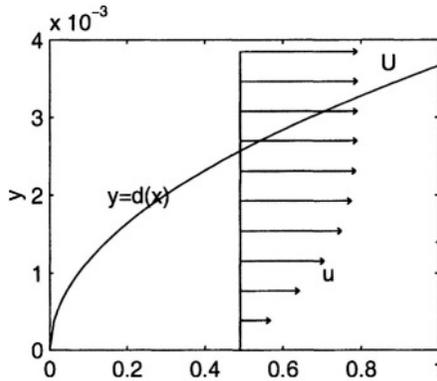


Figure 3.5. The boundary layer for a flat plate

so that the Blasius problem (3.5) becomes

$$\begin{aligned} g''' + \frac{1}{2}gg'' &= 0, \\ g(0) = g'(0) &= 0, \\ g'(\infty) &= k^2. \end{aligned}$$

But k appears only in the condition at infinity, therefore we may choose $g''(0) = 1$. By solving this single Cauchy problem, we obtain its solution $g(z)$, together with the derivatives $g'(z)$ and $g''(z)$, on a reasonably large interval for z . Taking the square root of g' we find the value of k at the end of that interval. Then, $f(\eta) = g(z)/k$ and $\eta = kz$.

Other procedure could be the use of the relation $f''(0) = k^{-3}$ after the calculation of k , and the solving of the Cauchy problem for f with these initial data.

The above problem may be complicated by injection or suction of fluid through the body surface resulting in a modification of the structure of the boundary layer and also of the heat transfer. If the injection of fluid is suitably distributed, the fluid flow remains self-similar, that is the equations describing the phenomenon and the boundary conditions may be transformed into a form with a single parameter as independent variable.

Such a case is when the velocity of the injected (or sucked) fluid is of the form

$$v_0 = C \left(\frac{\nu U}{x} \right)^{1/2}$$

where C is a constant. In this case the equation and the initial conditions of the problem (3.5) remain the same, excepting of $f(0) = -2C$ where C

positive or negative means injection, respectively suction of fluid. The results are shown in Figure 3.6.

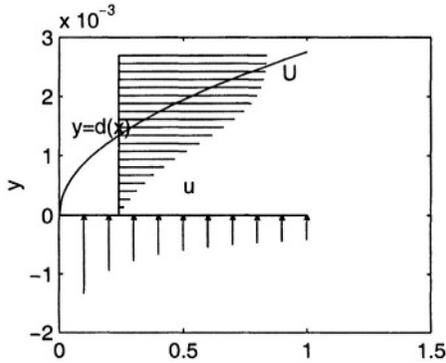


Figure 3.6. Boundary layer with injection of fluid

Now in the velocity profile within the boundary layer there exists an inflection point. At that point $u_{yy} = 0$ and this means an instability of the flow and a turbulence may develop in the boundary layer.

We remark that in the case of an injection or suction of fluid, we cannot apply the method of changing of variables to solve the Blasius problem. The constant k appears now at a boundary condition, not only at infinity and now $g''(0)$ cannot be arbitrarily chosen.

8.5 Dynamic Boundary Layer with Sliding on a Plane Plaue

We will now determine the characteristic values of the viscous boundary layer, disregarding the classical hypothesis of adherence to the wall [114].

Let us consider a semifinite plane plaue situated on the Ox axis, having the edge at O , attacked under a null angle by a viscous incompressible fluid stream. The flow is plane and we let Oxy be the plane of the flow. The fluid flow equations are

$$\rho \mathbf{v} \cdot \nabla u = -\frac{\partial p}{\partial x} + \mu \Delta u, \quad \rho \mathbf{v} \cdot \nabla v = -\frac{\partial p}{\partial y} + \mu \Delta v,$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0,$$

where $\mathbf{v} = \mathbf{v}(u, v)$, i.e.,

$$\rho \left(u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right) = \mu \frac{\partial^2 u}{\partial y^2}, \quad (3.7)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

according to the approximations of the boundary layer theory. Unlike the theory of the classical boundary layer, in which to these equations one associates the boundary conditions

$$u(x, 0) = 0, \quad v(x, 0) = 0, \quad u(x, \infty) = u_\infty,$$

in our case, the boundary conditions will be

$$u(x, 0) = L_1 \frac{\partial u}{\partial y}(x, 0), \quad v(x, 0) = 0, \quad u(x, \infty) = u_\infty, \quad (3.8)$$

the first signifying the fact that the fluid, in contact with the plaque, slides on its surface.

Taking v from the second equation (3.7) and replacing it into the first equation, we get

$$\rho \left[u \frac{\partial u}{\partial x} - \left(\int_0^y \frac{\partial u}{\partial x} dy \right) \frac{\partial u}{\partial y} \right] = \mu \frac{\partial^2 u}{\partial y^2}$$

and, by integration with respect to y from $y = 0$ to $y = \delta(x)$, we obtain

$$\rho \int_0^\delta u \frac{\partial u}{\partial x} dy - \rho \left[u \int_0^y \frac{\partial u}{\partial x} dy \right]_0^\delta + \rho \int_0^\delta u \frac{\partial u}{\partial x} dy = -\tau_w$$

where $\tau_w = \mu \left(\frac{\partial u}{\partial y} \right)_w$, thus leading to the integral relationship

$$\rho u_\infty^2 \frac{d}{dx} \int_0^\delta \frac{u}{u_\infty} \left(\frac{u}{u_\infty} - 1 \right) dy = -\tau_w. \quad (3.9)$$

We shall use this integral relationship by considering a velocity profile within the boundary layer of the shape

$$\frac{u}{u_\infty} \equiv \bar{u} = a_0 + a_1 \eta + a_2 \eta^2 + a_3 \eta^3 + a_4 \eta^4,$$

where

$$0 \leq y \leq \delta(x), \quad \frac{u}{u_\infty} \equiv \bar{u} = 1 \text{ for } y \geq \delta(x), \quad \eta \equiv \frac{y}{\delta(x)}.$$

The a_i coefficients can be determined by using the appropriate conditions

$$\bar{u} = L \frac{\partial \bar{u}}{\partial \eta}, \quad \frac{\partial^2 \bar{u}}{\partial \eta^2} = 0 \text{ for } \eta = 0,$$

$$\bar{u} = 1, \quad \frac{\partial \bar{u}}{\partial \eta} = 0, \quad \frac{\partial^2 \bar{u}}{\partial \eta^2} = 0 \text{ for } \eta = 1,$$

where $L = \frac{L_1}{\delta(x)}$.

Following the calculations, there appears the nondimensional profile of the horizontal component of the velocity, in the shape

$$\bar{u} = \frac{1}{1 + 2L} (2L + 2\eta - 2\eta^3 + \eta^4). \tag{3.10}$$

In Figures 3.7, respectively 3.8, we present the profile of the nondimensional velocity together with the influence of the L parameter on the velocity's profile.

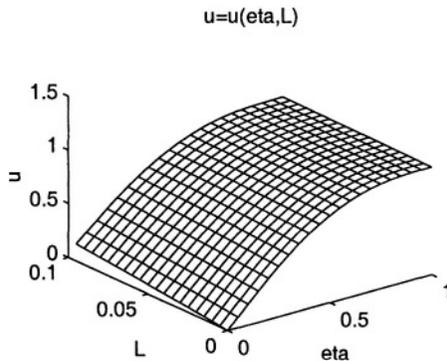


Figure 3.7. The profile of the nonmensional velocity

Now, one can also determine other characteristic values of the boundary layer. For instance, the local tension between two neighbor layers $\tau = \mu \left(\frac{\partial u}{\partial y} \right)$ has the expression

$$\tau = \frac{2\mu u_\infty}{\delta(1 + 2L)} (1 - 3\eta^2 + 2\eta^3)$$

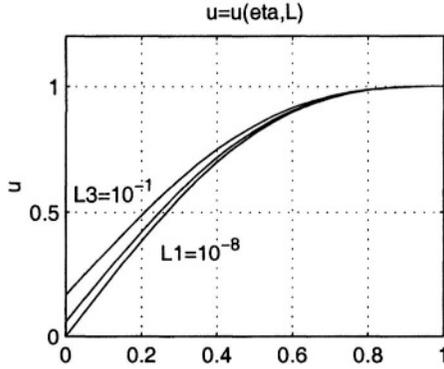


Figure 3.8. The influence of the L parameter

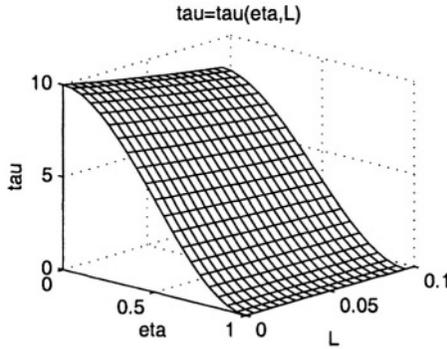


Figure 3.9. The local tension between two neighbor layers

and it is represented, within the section $x = const.$ in Figure 3.9.

The local stress on the plaque has the expression

$$\tau_w = \frac{2\mu u_\infty}{\delta(1 + 2L)}. \tag{3.11}$$

Replacing the velocity expression (3.10) and the local stress on the plaque (3.11) in the integral relationship (3.9), we get

$$\rho u_\infty^2 \frac{d}{dx} \left[\frac{\delta}{(1 + 2L)^2} \left(\frac{3L}{5} + \frac{37}{315} \right) \right] = \frac{2\mu u_\infty}{\delta(1 + 2L)},$$

respectively

$$\frac{d(\delta^2)}{dx} = (1 + 2L) \frac{315}{37 + 189L} \frac{4\mu}{\rho u_\infty},$$

from where, by integrating, it turns out that

$$\delta(x) = 2\sqrt{(1 + 2L) \frac{315}{37 + 189L} \frac{\mu}{\rho u_\infty^3} x},$$

due to $\delta(0) = 0$. From the relation (3.11) we get the expression of the local stress on the plaque

$$\tau_w(x) = \frac{1}{(1 + 2L)^{3/2}} \sqrt{\frac{37 + 189L}{315} \frac{\rho \mu u_\infty^3}{x}}$$

which is represented in Figure 3.10.

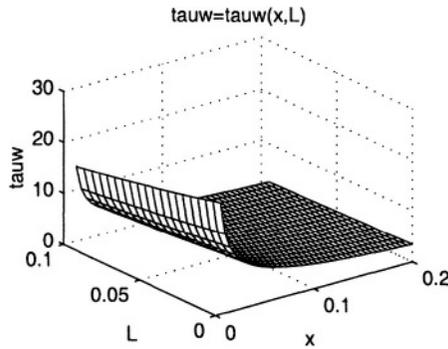


Figure 3.10. The local stress on the plaque

The influence of the abscissa x and of the L parameter on the thickness of the boundary layer is presented in Figure 3.11.

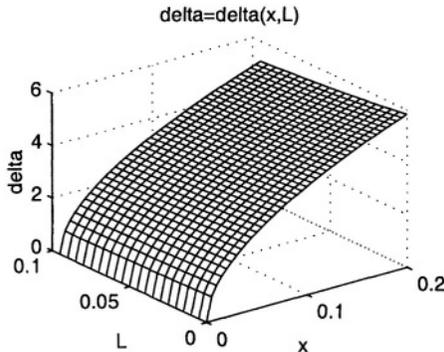


Figure 3.11. The thickness of the boundary layer

Chapter 4

INTRODUCTION TO NUMERICAL SOLUTIONS FOR ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS

1. Introduction

The equations describing the flow of fluids are ordinary or partial differential equations which combine the flow variables (the velocity components, the pressure, etc.) and their derivatives. But for most of these equations there are no analytical methods to find their solutions. Consequently, different *numerical methods* should be used, methods which allow us to produce approximative solutions by using computers. For more details on such methods which are also presented in this book, we refer to [4], [13], [18], [22], [43], [79], [100], [120], [121], [125], [128], [131], [145], [155].

The main quantitative feature that we deal with is the *accuracy* of a numerical method, i.e., its ability to approximate “as well as possible” the analytical solution of the given problem when the approximation tools become “fine enough”. The main qualitative feature taken into account is the *stability* of the method, i.e., its ability to not propagate and not accumulate errors from the previous calculations to the following ones.

The first step to numerically solve a given problem is its *numerical discretization*. This means that each component of the differential or partial differential equation is transformed into a “numerical analogue” which can be represented in the computer and then processed by a computer program, built on some algorithm.

The continuous form of these models could be represented as

$$Au = f.$$

Excepting some very simple cases, we can not determine the exact solutions of these equations and therefore we should find at least some

approximative solutions that describe well enough the physical phenomenon. These approximative solutions must be the elements u_h from a finite dimensional space, calculable by an acceptable effort from a finite system of equations of the type

$$A_h u_h = f_h.$$

Here h is a parameter supposed to tend towards zero, when the dimension of the system tends to infinity. The essential problem is the link between u_h and u . For its study, we need also a link between the finite dimensional space and the continuous space which allows finally the evaluation of the distance (deviation) between u_h and u , distance (deviation) that must become small for a small h (the convergence problem). For this, we need first a study which ensures that A_h becomes closer to A when $h \rightarrow 0$ (the consistency problem). Moreover, we need also a study which ensures that u_h belongs to a bounded set when $h \rightarrow 0$ (the stability problem).

For example, the finite differences method based upon the Taylor series, describes the derivatives of a function as the difference between its values at various points. In other words, the method replaces the derivative operators from A with combinations of some "translation" operators into A_h . If we know the values of the function u and its derivatives at the point x , we could approximate the values of u at the neighboring points $x + h$ or $x - h$ by

$$u(x + h) = u(x) + h \frac{du}{dx} + \frac{h^2}{2} \frac{d^2u}{dx^2} + \frac{h^3}{6} \frac{d^3u}{dx^3} + \dots$$

$$u(x - h) = u(x) - h \frac{du}{dx} + \frac{h^2}{2} \frac{d^2u}{dx^2} - \frac{h^3}{6} \frac{d^3u}{dx^3} + \dots$$

where h is small and the derivatives of u are calculated at x .

But if we know the values of u at $x - h$, x , $x + h$, by adding and subtracting the above formulas we can approximate the first and the second order derivatives of u at x , namely

$$\frac{du}{dx} = \frac{1}{h} [u(x + h) - u(x)] + O(h)$$

or

$$\frac{du}{dx} = \frac{1}{2h} [u(x + h) - u(x - h)] + O(h^2)$$

and

$$\frac{d^2u}{dx^2} = \frac{1}{h^2} [u(x - h) - 2u(x) + u(x + h)] + O(h^2)$$

where $O(h)$ or $O(h^2)$ represents the error order.

Combining these formulas into the given equation $Au - f = 0$, we get

$$A_h u_h - f_h \equiv \sum_k a_k u(x + kh) - \sum_k b_k f(x + kh) = 0$$

The above formulas and others deduced by various techniques, as we will see in the next sections, allow the replacing of every term from the given equation, and thus obtaining its numerical analogue. This can be performed by choosing a grid in the computational domain and replacing the derivatives at the grid points with finite differences, as above. Finally, we obtain a system from which we calculate the values of the unknown functions at the grid points, i.e., we calculate the numerical solution.

By this procedure, a differential or partial differential equation defined on the entire domain, that is at an infinite number of points, is transformed into a system with a finite number of equations which describes the relations between the values of the unknown solution at a finite number of points (belonging to the domain).

If u is the exact solution and u_h the numerical one, then $A_h u - f_h$ is called the *residue*. If $A_h u - f_h = O(h^p)$ when $h \rightarrow 0$, p is called the *truncation order*. The discretization procedure is *consistent* if the truncation error tends towards zero when $h \rightarrow 0$. But consistency is not sufficient to prove the convergence of u_h towards u . We have

$$u - u_h = (A_h)^{-1} (A_h u - f_h)$$

and thus a uniform boundedness of $(A_h)^{-1}$ into the considered functional space is also necessary, a property which is called the *stability* of the approximation scheme. It comes usually from the relation

$$\left\| (A_h)^{-1} u \right\| \leq K(u)$$

by applying the Banach–Steinhaus theorem [121].

There are other aspects that must be taken into account when we analyze a numerical method. Let us take an illustrative example, specifically

$$\begin{aligned} u_x - v u_{xx} &= 1, \\ u(0) &= 0, u(1) = 0, \end{aligned}$$

where the exact solution is

$$u = x - \frac{1 - e^{\frac{x}{v}}}{1 - e^{\frac{1}{v}}}.$$

Let us discretize this equation with centered finite differences

$$\frac{1}{2h} [u(x+h) - u(x-h)] - \frac{v}{h^2} [u(x-h) - 2u(x) + u(x+h)] = 1.$$

If u_j denotes the approximation of $u(jh)$, for $j = 0, 1, \dots, N$ where $h = 1/N$, we can calculate the exact discrete solution from the above equations

$$u_j = jh - \frac{1 - q^j}{1 - q^N}, \quad q = \frac{2 + P}{2 - P}$$

where $P = h/v$ is the Peclet number. For $P > 2$ we remark some important oscillations of the numerical solution in the vicinity of $x = 1$, see Figure 4.1. For $P < 2$ we have no oscillations.

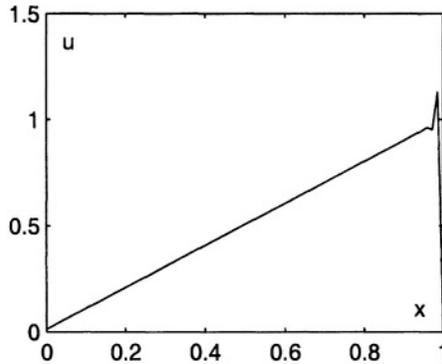


Figure 4.1. The spatial instability

This particular behaviour is called *spatial instability* of the numerical method and it is due to the dominant advective character of the equation in the case of a small coefficient v .

If we use another numerical scheme, for instance

$$\frac{1}{h} (u_j - u_{j-1}) - \frac{v}{h^2} (u_{j+1} - 2u_j + u_{j-1}) = 1,$$

the numerical solution is given through the same formula but with $q = 1 + 1/P$ and therefore the spatial instability does not interfere.

In the case of a time evolution, by discretization of the time derivative one can obtain explicit or implicit links between the values of the unknown function at different time instants. It is necessary to study the time stability of the envisaged numerical method.

The passing from a time level to another is numerically performed by multiplication by a complex factor — the so-called *amplification factor*. The errors appear, in magnitude — the *dissipative errors* — if the amplification factor is, in magnitude, less than 1, or in *phase*, if the numerical solution is advected along a different speed than the exact solution. If

the amplification factor is in magnitude larger than 1 the scheme is unstable. The phase errors are joined to the odd order derivatives which are present in the equation, while the dissipative ones are joined to the even order derivatives.

The discretization is often performed in two stages, using the *lines method*. First, a spatial discretization is performed, obtaining a system of time differential equations. To this system the specific methods are then applied. The distribution of the eigenvalues of the spatial operator from the discretized equation and the behaviour of the amplification factor have an important role for the study of the algorithm.

The schemes which are not of this form are the *space-time* schemes. Typical examples are the Lax–Wendroff (1960) and MacCormack (1969) methods, but from the 1980s they gradually were replaced by the lines methods. A reason for this is that the numerically steady solutions for the space-time schemes could depend on the considered time step-size.

In physical problems, the admissible values of some variables are limited to some intervals. On the other hand, some numerical methods allow the generation of spurious oscillations in the numerical solutions, violating the above requirement.

Numerical schemes with a higher accuracy and generating lower oscillations must be used. One of the properties characterizing such schemes is the reduction of the total variation of the numerical solution (TVD - Total Variation Diminishing) when marching in time, $TV(u^{n+1}) \leq TV(u^n)$ where $TV(u^n) = \sum_j |u_{j+1}^n - u_j^n|$.

A much used scheme is MUSCL (Monotonic Upstream Scheme for Conservation Laws), elaborated by Van Leer in 1983. For the construction of a nonoscillatory scheme it is important to reconstruct a local interpolant of the unknown function from a discrete set of values.

Harten and Osher (1987) found a criterion which allows the construction of schemes not-TVD but yet nonoscillatory. A reconstruction of degree k , $R(u, k)$ of the function u is *essentially nonoscillatory (ENO)* if $TV(R(u, k)) \leq TV(u) + O(h^r)$ for $r \leq k$. Of course, in the neighborhood of some singularities of the solution, the accuracy of these schemes is not so good and must be improved by the grid refinement. But this action could lead to stability problems which could be avoided by choosing of some spatial discretizations with better stability qualities.

In the sequel we will illustrate, by some simple examples, the main numerical methods for the basic types of problems of fluid dynamics. We remark that, taking into account the significance and the frequency of the appearance of these equations in practical problems, a lot of software

was elaborated, more or less comprehensive, more or less accessible, in order to solve numerically such problems.

Nowadays, the calculation of the values of some elementary or special functions is no more a problem; many optimized algorithms are implemented on all computing packages and the solving of linear systems of equations $Ax = b$ is very easy. The exact solving methods for such systems are now accessible in MATLAB by the command $x = A \setminus b$, which analyses the matrix A and chooses the optimal solving procedure. The frequently encountered case of a sparse matrix A is also considered; so we may solve large systems of thousands of equations within an acceptable computing time.

For very large systems, some iterative methods are also available (**gmres** - *Generalized Minimum Residual*, **pcg** - *Preconditioned Conjugate Gradients*, for instance). These iterative methods need, usually, the description of the matrix A or only the algorithm to calculate the matrix-vector product Au and they are particularly efficient. Of course, complex problems may lead to very large systems of algebraic equations whose solving is very difficult or even impossible with the already implemented methods. In these cases it is necessary to find and to programme specific algorithms taking into account the specific structure of the system.

Analogously, the numerically solving of the main problems for partial differential equations is facilitated by using the **(PDE)-Partial Differential Equations** toolbox of MATLAB which allows a complete treatment, from a description of the computational domain, to imposing of the initial and the boundary conditions, choice of the (constant or variable) coefficients of the equations, discretization of the domain by a suitable triangular mesh, implementation of the finite element method (including visualization of the solution), mesh refinement, etc.

Unfortunately, the increasing specificity of the problems reduces the flexibility of these packages. They are designed to solve standard problems, more and more complex, with few variations, for specific domains and taking into account only certain equations and phenomena. We remark, for instance, the industrial packages FLUENT or COSMOS, used to solve problems from fluid dynamics and heat transfer in 3D, which is in a continuous development. Other software, based on the finite element method, finite differences, finite volumes or spectral methods are FEATFLOW, SIMPLE, QUICK, PHOENICS, FLOTRAN, NSFLEX, FIDAP, FIRE, LISS, FASTEST, FEMLAB and many others, for educational or scientific purposes, accessible on INTERNET.

2. Discretization of a Simple Equation

In order to illustrate and compare some discretization methods, we apply them to a simple equation (the one-dimensional diffusion equation)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

2.1 Using the Finite Difference Method

We start by establishing the domain where the equation is studied. If, for example, we model the diffusion of a gas into a tube of length l , the spatial domain is the interval of the Ox axis associated to this length, i.e., $(0, l)$. The time domain begins at $t = 0$ and indefinitely extends to the positive direction of the time axis Ot , i.e., $(0, +\infty)$. Concluding, the equation domain is $\Omega = (0, l) \times (0, +\infty) \subset \mathbb{R}^2$.

Now we can choose the grid. We will construct a grid formed by the straight lines $x = x_j$, $j = 0, 1, \dots, n$ where $x_0 = 0$ and $x_n = l$, with the constant step size $\Delta x = x_{j+1} - x_j$ for all j in the Ox direction and the straight lines $t = t_k$, $k = 0, 1, \dots$ where $t_0 = 0$ with the constant step size $\Delta t = t_{k+1} - t_k$ for all k in the Ot direction. The nodes will be the intersection points of these straight lines, i.e., (x_j, t_k) , $j = 0, 1, \dots, n$, $k = 0, 1, \dots$

We are able now to discretize the equation by replacing the derivatives by finite differences. For example, if we denote by $u_j^k = u(x_j, t_k)$, we obtain for the node (x_j, t_k) ,

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} = \frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{\Delta x^2},$$

which could be reset in the form

$$u_j^{k+1} = \frac{\Delta t}{\Delta x^2} u_{j-1}^k + \left(1 - 2 \frac{\Delta t}{\Delta x^2}\right) u_j^k + \frac{\Delta t}{\Delta x^2} u_{j+1}^k.$$

Applying these formulas for any $j = 1, \dots, n - 1$, we see that from the known values for $k = 0$ (*the initial conditions*) we can calculate those for $k = 1$, then from these values we calculate those for $k = 2$ and so on. At each step, we must know the values u_0^k and u_n^k (from the *boundary conditions*) in order to complete the time level values. Such a procedure is called *explicit*. There are many such formulas, as we will see in a next chapter.

2.2 Using the Finite Element Method

We will choose the same grid as that for the finite difference method but for instance we will discretize the equation only with respect to the

time

$$\frac{u^{k+1} - u^k}{\Delta t} = \frac{\partial^2 u}{\partial x^2}$$

where $u^k = u(x, t_k)$.

Let us construct the variational (or weak) form of this equation, by multiplication with the known function v and by integration upon $(0, l)$

$$\int_0^l v \frac{u^{k+1} - u^k}{\Delta t} dx = \int_0^l v \frac{\partial^2 u}{\partial x^2} dx$$

which becomes, after an integration by parts,

$$\int_0^l v \frac{u^{k+1} - u^k}{\Delta t} dx = v \frac{\partial u}{\partial x} \Big|_0^l - \int_0^l \frac{\partial v}{\partial x} \frac{\partial u}{\partial x} dx.$$

Let us transform now this equation into its numerical analogue. We divide the spatial domain $(0, l)$ into elements, for example $(x_1, x_2) \cup (x_2, x_3)$, on each element we seek the unknown function under the form $u = \sum_{j=1}^2 N_j u_j$ where N_j are the shape functions and u_j are those corresponding to that element's nodal values. Choosing the multipliers v to be the shape functions on each element and considering the right-hand side of the variational equation at the same time instant t_k (the explicit procedure), we find

$$\int_{x_1}^{x_2} N_n \sum_{j=1}^2 \frac{N_j u_j^{k+1} - N_j u_j^k}{\Delta t} dx = v \frac{\partial u}{\partial x} \Big|_{x_1}^{x_2} - \int_0^l \frac{\partial N_n}{\partial x} \sum_{j=1}^2 \frac{\partial N_j u_j^k}{\partial x} dx,$$

$n = 1, 2$, for the first element and a similar equation for the second.

But the shape functions are simple, the above integrals can be exactly calculated, the integrated parts reciprocally reduce at the interior nodes and finally we obtain two equations for each element, having as unknowns the nodal values. In matrix form, these equations are, for each element

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

Assembling these elements, the local numbering 1 – 2 becomes a global numbering 1 – 2 – 3 and the above systems become

$$\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ u_3^{k+1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ 0 \end{pmatrix}$$

for the first element and

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & b_{11} & b_{12} \\ 0 & b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ u_3^{k+1} \end{pmatrix} = \begin{pmatrix} 0 \\ g_1 \\ g_2 \end{pmatrix}$$

for the second.

Combining these local systems into a global system, we get

$$\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} + b_{11} & b_{12} \\ 0 & b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ u_3^{k+1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 + g_1 \\ g_2 \end{pmatrix}.$$

Here we introduce the boundary conditions and then, by solving the system, we get the nodal values of the solution at the instant t_{k+1} from the values at the instant t_k (which appear on the right-hand side). We also remark, although for two elements it is not yet apparent, that the matrix of the system is a sparse matrix and thus the system could be solved by corresponding techniques.

2.3 Using the Finite Volume Method

At the first step we discretize in time the equation,

$$\frac{u^{k+1} - u^k}{\Delta t} = \frac{\partial^2 u}{\partial x^2}.$$

Then, at the time instant t_k , we divide the spatial domain $(0, l)$ into finite volumes (in our case they are intervals too) but having the reference point P at the center. Considering three such neighboring finite volumes, with centers at the points W and E (to West respectively to East of P), these volumes have their interior boundaries placed at the points w between W and P , respectively e between P and E . The discretization of the spatial derivative is now performed by the formula

$$\frac{\partial^2 u}{\partial x^2} \Big|_P = \frac{\frac{\partial u}{\partial x} \Big|_e - \frac{\partial u}{\partial x} \Big|_w}{x_e - x_w}$$

and then

$$\frac{\partial u}{\partial x} \Big|_e = \frac{u_E - u_P}{x_E - x_P}, \quad \frac{\partial u}{\partial x} \Big|_w = \frac{u_P - u_W}{x_P - x_W}.$$

Replacing into the above equation for every reference point E, P, W we obtain another system from which we can calculate u_E, u_P, u_W at the next time instant t_{k+1} . This step is performed as for the finite differences method, using the initial and boundary conditions. What is different in these two methods is the discretization procedure.

2.4 Comparison of the Discretization Techniques

The above presented methods have a common feature: they generate equations for the values of the unknown functions at a finite number of points in the computational domain.

But there are also several differences. The finite difference and the finite volume methods generate numerical equations at the reference point based on the values at neighboring points. The finite element method produces equations for each element independently of all other elements. Only when the equations are collected together and assembled into a global matrix are the interactions between elements taken into account.

The finite element method takes care of boundary conditions of Neumann type while the other two methods can easily apply to the Dirichlet conditions.

The finite difference method could be easily extended to multidimensional spatial domains if the chosen grid is regular (the cells must look cuboid, in a topological sense). The grid indexing is simple but some difficulties appear for the domain with a complex geometry.

For the finite element method there are no restrictions on the connection of the elements when the sides (or faces) of the elements are correctly aligned and have the same nodes for the neighboring elements. This flexibility allows us to model a very complex geometry.

The finite volume method could also use irregular grids like the grids for the finite element methods, but keeps the simplicity of writing the equations like that for the finite difference method. Of course, the presence of a complex geometry slows down the computational programs.

Another advantage of the finite element method is that of the specific mode to deduce the equations for each element which are then assembled. Therefore, the addition of new elements by refinement of the existing ones is not a major problem. For the other methods, the mesh refinement is a major task and could involve the rewriting of the program.

But for all the methods used for the discrete analogue of the initial equation, the obtained system of simultaneous equations must be solved. The time marching from one time level to another could lead to a blow-up of the numerical accumulated errors (the *numerical instability* of the computations). This instability must be counteracted by using suitable discretization procedures. On the other hand, when the spatial dimensions of the cells tend towards zero, the numerical solutions must tend towards the analytical solution of the problem (the *convergence* of the algorithm). The following chapters will detail these features.

3. The Cauchy Problem for Ordinary Differential Equations

The simplest problems for ordinary differential equations (ODE) are that for the first order equations

$$\frac{dy}{dt} = f(t, y) \quad (4.1)$$

where $y(x)$ is the unknown function. The geometric interpretation of such an equation is based on the idea that for a given function $y = y(t)$, its derivative $\frac{dy(t)}{dt}$ represents the slope of the tangent to its graph at the point t . If at any point (t, y) from \mathbb{R}^2 (or from the definition domain of the equation) we draw a vector of slope $f(t, y)$, we obtain a vector field and therefore the differential equation defines a family of curves (trajectories) which are tangent at every point (t, y) to the corresponding vector of the field.

For example, for the differential equation $\frac{dy}{dt} = t^2 y$ we obtain Figure 4.2 where the (trajectories) curves family mentioned is obvious. From

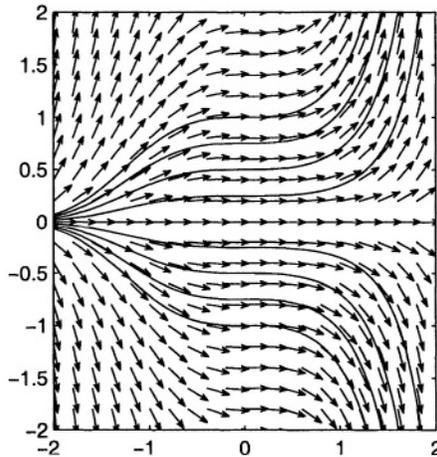


Figure 4.2. The flow field generated by the equation $\frac{dy}{dt} = t^2 y$

here arises also the notion of *flow field* generated by the differential equation, because the image is similar to the motion of the particles of some fluid flow.

It is “obvious” from the picture that we can choose a unique solution by choosing a point (t, y) on the respective curve, i.e., by imposing a condition of the form

$$y(t_0) = y_0 \quad (4.2)$$

called also a *Cauchy condition*. The two relations (4.1, 4.2) form a *Cauchy problem*.

There exists a natural trend to “a priori” suppose the existence and the uniqueness of the solution of a Cauchy problem since the differential equation models a real, physical, observable phenomenon. However, the real process and its mathematical model are two distinct entities. The model reflects only *partially* the phenomenon, therefore it is possible that some models have either no solutions or many solutions, some of which without physical relevance.

The aim of the existence and uniqueness theorems is to describe families of equations as large as possible for which the existence and the uniqueness of the Cauchy problem is ensured. For some difficult problems, often there are no explicit formulas for the solutions and implicitly numerical calculations must be used. In these cases it is important to know that a solution exists before investing time and computing effort to look for something that eventually could not be found.

Definition. A *solution* of the Cauchy problem (4.1, 4.2) is a differentiable function y of t , on an interval I which contains t_0 , which verifies

$$\frac{d}{dt}y(t) = f(t, y(t)), \forall t \in I$$

and

$$y(t_0) = y_0.$$

We remark that this definition could be weakened, by accepting the nondifferentiability of y on a “small enough” set of points $t \in I$.

In order to ensure the existence and the uniqueness we must impose some restraints on the function f , i.e., on the slopes of the trajectories generated by the differential equation. For example, the problem

$$\frac{dy}{dt} = 2\sqrt{y}, \quad y(0) = 0$$

has two solutions on $I = (0, +\infty)$, $z(t) = 0$ and $w(t) = t^2$. This may occur due to the rapid change of the slopes of the solutions near $t = 0$, generated by the function \sqrt{y} .

The usual requirements that ensure the existence and the uniqueness are the continuity of the function f with respect to t and the satisfaction of the Lipschitz condition

$$\exists K > 0 : |f(t, z) - f(t, w)| \leq K |z - w|, \forall t, z, w$$

with respect to the second argument of f . The proof of the existence theorem is based on the transformation of the given differential equation

into an integral equation

$$y(t) = y(t_0) + \int_{t_0}^t f(s, y(s)) ds$$

and on the fact that this Volterra type equation has a solution which could be found by a convergent process of successive iterations (Picard), namely

$$\begin{aligned} y_0(t) &= y_0, \\ y_1(t) &= y_0 + \int_{t_0}^t f(s, y_0(s)) ds, \\ y_2(t) &= y_0 + \int_{t_0}^t f(s, y_1(s)) ds, \\ &\vdots \\ y_n(t) &= y_0 + \int_{t_0}^t f(s, y_{n-1}(s)) ds, \\ &\vdots \end{aligned}$$

THEOREM 4.1. *Suppose that in $D = [t_0 - a, t_0 + a] \times [y_0 - b, y_0 + b]$ the function $f(t, y)$ is continuous with respect to t and verifies a Lipschitz condition*

$$\exists K > 0 : |f(t, z) - f(t, w)| \leq K |z - w|, \forall (t, z), (t, w) \in D.$$

Then there exists a unique solution of the Cauchy problem (4.1, 4.2), which can be extended until the boundary of D .

Let us recall the example of Figure (4.2)

$$\frac{dy}{dt} = t^2 y, \quad y(0) = 1$$

where the associated integral equation is

$$y(x) = 1 + \int_0^x s^2 y(s) ds.$$

The successive iterations are

$$y_0(x) = 1,$$

$$y_1(t) = 1 + \int_0^t s^2 ds = 1 + \frac{t^3}{3},$$

$$y_2(t) = 1 + \int_0^t s^2 \left(1 + \frac{s^3}{3}\right) ds = 1 + \frac{t^3}{3} + \frac{t^6}{18},$$

$$y_3(t) = 1 + \int_0^t s^2 \left(1 + \frac{s^3}{3} + \frac{s^6}{18}\right) ds = 1 + \frac{t^3}{3} + \frac{t^6}{18} + \frac{t^9}{162},$$

⋮

and we recognize the partial sums of the power series expansion of the exact solution $y = e^{t^3/3}$.

In many cases we can find such explicit solutions. But, also many important problems have no such representations of the solutions and we should use numerical approximation methods.

There are many such numerical methods. In simple cases, a simple method could be satisfactory but more “serious” problems could require the more elaborate methods.

A first problem to solve is to establish *what the numerical method calculates*. As an algorithm which runs a finite time interval gives only a finite number of outputs, we should determine what those values represent. They could be approximations of the coefficients of some series expansion (as for the previous example) or they could be approximations of the values of the solution at a finite number of points, previously or even chosen while running. Moreover, the numerical method should allow also some estimations of the approximation errors.

A second problem is to *calculate the next values from the previous ones*, for example to calculate $y(t+h)$ once given $y(t)$. This suggests the Taylor’s series finite expansion (Taylor’s formula)

$$y(t+h) = y(t) + h \frac{dy(t)}{dt} + \frac{h^2}{2!} \frac{d^2y(t)}{dt^2} + \dots + \frac{h^{n+1}}{(n+1)!} \frac{d^{n+1}y(\xi)}{dt^{n+1}}$$

where the last term is an error term and $\xi \in (t, t+h)$.

The simplest numerical method (Euler) derives from the above expansion by truncation after the linear term

$$y(t+h) = y(t) + h \frac{dy(t)}{dt}$$

which leads to the basic formula

$$y_n = y_{n-1} + hf(t_{n-1}, y_{n-1}) \quad (4.3)$$

where $y_n \approx y(nh)$, $t_n = nh$ and h is a chosen step size.

Assuming that the second derivative of the solution is bounded by M in magnitude, one can show that the step error is of order $O(h^2)$ and the total error on the interval (a, b) where $a = t_0 < t_1 < \dots < t_n = nh = b$ is bounded by $M \frac{b-a}{2} h$, i.e., it is of order $O(h)$.

We could obtain better methods, with errors of order $O(h^p)$, for $p > 1$, using the above integral representation

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s)) ds = y(t) + h \left(\frac{1}{h} \int_t^{t+h} f(s, y(s)) ds \right).$$

Here the last term in the parentheses represents an average slope of the solution on the interval $(t, t+h)$. A good numerical method should calculate, as accurately as possible, this average slope.

For example, the Euler method takes as average slope the solution value at t . Of course, a better value seems to be the slope considered at the midpoint of the interval $t + \frac{h}{2}$, i.e.,

$$y(t+h) = y(t) + hf \left(t + \frac{h}{2}, y(t + \frac{h}{2}) \right).$$

The problem here is the calculation of the solution $y(t + \frac{h}{2})$ which is, in fact, the same problem as that to be solved. But this value at the midpoint of the interval could be also approximated by an ‘‘Euler step’’, precisely

$$y(t + \frac{h}{2}) \approx y(t) + \frac{h}{2} f(t, y(t)),$$

and thus we obtain an algorithm of the form

| | |
|---|-------|
| $K_1 = f(t, y(t)),$ $K_2 = f(t + \frac{h}{2}, y(t) + \frac{h}{2} K_1),$ $y(t+h) = y(t) + hK_2.$ | (4.4) |
|---|-------|

By developing these expressions we obtain a coincidence with the Taylor development of the solution until the term in h^2 so that the step error of the above algorithm (Runge) is of order $O(h^3)$ while the total error on (a, b) is of order $O(h^2)$. The price paid for this is the twice evaluation of the function f at each step.

The general methods of such type, called *Runge-Kutta methods* consist of a sequence of stages, at each stage evaluating an approximative value of the slope of the exact solution. The final step advances the solution from t to $t + h$ by using a weighted sum of the above calculated slopes. This means

$$\begin{aligned}
 K_1 &= f(t, y(t)), \\
 K_2 &= f(t + c_2h, y(t) + ha_{2,1}K_1), \\
 K_3 &= f(t + c_3h, y(t) + ha_{3,1}K_1 + ha_{3,2}K_2), \\
 &\vdots \\
 K_s &= f(t + c_sh, y(t) + ha_{s,1}K_1 + ha_{s,2}K_2 + \dots + ha_{s,s-1}K_{s-1}), \\
 y(t + h) &= y(t) + h(b_1K_1 + b_2K_2 + \dots b_sK_s)
 \end{aligned}$$

(4.5)

where s is the number of stages. A particular method is characterized by the coefficients $a_{i,j}, b_i$ and c_i which could be given in a Butcher table — see Table 4.1.

Table 4.1. Runge-Kutta method

| | | | | | |
|-------|-----------|-----------|-----|-------------|-------|
| 0 | | | | | |
| c_2 | $a_{2,1}$ | | | | |
| c_3 | $a_{3,1}$ | $a_{3,2}$ | | | |
| ... | ... | ... | ... | | |
| c_s | $a_{s,1}$ | $a_{s,2}$ | ... | $a_{s,s-1}$ | |
| | b_1 | b_2 | ... | b_{s-1} | b_s |

For example, the above Runge method (4.4) has Table 4.2

Table 4.2. Runge method

| | |
|-----|-----|
| 0 | |
| 1/2 | 1/2 |
| | 0 1 |

These methods use a fixed step size h . By diminishing h the accuracy, but also the computing time, increases. It is possible to diminish the step size only where the approximative solution changes rapidly its values and we could use a larger step size in the regions with a slow variation of the solution. Consequently, the step size h should be modified while

calculating and in agreement with the solution's behaviour. This task could be performed, for instance, by running (in parallel) two different methods, one for the solution propagation and the other to estimate and to control the errors.

For example, the popular method RK4 with 4 stages of Kutta, Table 4.3, gives the approximation

Table 4.3. Kutta method

| | | | | |
|-----|-----|-----|-----|-----|
| 0 | | | | |
| 1/2 | 1/2 | | | |
| 1/2 | 0 | 1/2 | | |
| 1 | 0 | 0 | 1 | |
| | 1/6 | 2/6 | 2/6 | 1/6 |

$$y(t+h) = RK4(t,h) + Mh^5 + O(h^6)$$

where $y(t+h)$ is the exact solution and $RK4(t,h)$ is a step obtained by this method. The coincidence with the Taylor series, of the exact solution is until the order 4. This method could be coupled by a RK3 method, of order 3, Table 4.4, which gives a similar formula

Table 4.4. RK3 method

| | | | |
|-----|-----|-----|-----|
| 0 | | | |
| 1/3 | 1/3 | | |
| 2/3 | 0 | 2/3 | |
| | 1/4 | 0 | 3/4 |

$$y(t+h) = RK3(t,h) + Kh^4 + O(h^5).$$

By subtraction of the above two representations for $y(t+h)$ we get

$$0 = RK3(t,h) - RK4(t,h) + Kh^4 + O(h^5)$$

from where

$$Kh^4 = RK4(t,h) - RK3(t,h) + O(h^5).$$

Consequently, calculating RK4 we can give a good approximation of the error of RK3. But this parallel calculation requires new evaluations of the function f . Fehlberg has discovered that there exist some pairs of Runge-Kutta methods with different truncation orders while the main

lines of the respective tables are the same. So that, the step size h could be fitted using only one supplementary evaluation for the function f .

Such a pair is formed by the methods described in Tables 4.5, 4.6 with the truncation order 5, respectively 6, so that the accuracy of the method is of order 4, respectively 5. There are many other such pairs,

Table 4.5. RK5 method

| | | | | | |
|-----|--------|----------|--------|-------|------|
| 0 | | | | | |
| 2/9 | 2/9 | | | | |
| 1/3 | 1/12 | 1/4 | | | |
| 3/4 | 69/128 | -243/128 | 135/64 | | |
| 1 | -17/12 | 27/4 | -27/5 | 16/5 | |
| | 1/9 | 0 | 9/20 | 16/45 | 1/12 |

Table 4.6. RK6 method

| | | | | | | |
|-----|--------|----------|--------|--------|-------|------|
| 0 | | | | | | |
| 2/9 | 2/9 | | | | | |
| 1/3 | 1/12 | 1/4 | | | | |
| 3/4 | 69/128 | -243/128 | 135/64 | | | |
| 1 | -17/12 | 27/4 | -27/5 | 16/5 | | |
| 5/6 | 65/432 | -5/16 | 13/16 | 4/27 | 5/144 | |
| | 47/450 | 0 | 12/25 | 32/225 | 1/30 | 6/25 |

implemented in the usual computing packages.

The above presented methods are also applicable (in the vector form) for the first order systems of differential equations, namely

$$\begin{aligned} \frac{dy_1}{dt} &= f_1(t, y_1, \dots, y_n), y_1(0) = y_{01}, \\ &\dots \\ \frac{dy_n}{dt} &= f_n(t, y_1, \dots, y_n), y_n(0) = y_{0n}. \end{aligned}$$

Therefore, the higher order differential equations

$$\frac{d^n y}{dt^n} + f\left(t, y, \frac{dy}{dt}, \dots, \frac{d^{n-1}y}{dt^{n-1}}\right) = 0,$$

$$y(0) = y_0, y'(0) = y'_0, \dots, y^{(n-1)}(0) = y_0^{(n-1)},$$

which, by the change of variable and function

$$t = y_0, y = y_1, y' = y_2, \dots, y^{(n-1)} = y_n,$$

is reduced to a system of the form

$$\begin{aligned}y_0' &= 1, \\y_1' &= y_2, \\y_2' &= y_3, \\&\dots \\y_{n-1}' &= y_n, \\y_n' &= f(y_0, y_1, y_2, \dots, y_n)\end{aligned}$$

can also use the above methods.

For example, the problem

$$\begin{aligned}\frac{d^2y}{dt^2} + \cos t \frac{dy}{dt} + y &= 0, \\y(0) = 0, y'(0) &= 1\end{aligned}$$

reduces to the system

$$\begin{aligned}y_0' &= 1, \\y_1' &= y_2, \\y_2' &= -y_1 - y_2 \cos y_0, \\y_0(0) = 0, y_1(0) = 0, y_2(0) &= 1\end{aligned}$$

which is of the form

$$\begin{aligned}Y' &= F(Y), \\Y(0) &= Y_0.\end{aligned}$$

The numerical integration of this problem by MATLAB requires a subprogram which describes the system

```
function yp=funct(t,y)
yp=zeros(2,1);
yp(1)=y(2);
yp(2)=-y(1)-y(2)*cos(t);
```

saved as `funct.m`, while the main program

```
[t,y]=ode45(@funct,[0,50],[0,1]);
plot(t,y(:,1));pause;plot(y(:,1),y(:,2));
```

performs the integration of the system on the interval $[0,50]$ with the given initial conditions and plots the solution $y(t)$ and the phase portrait (i.e., the curve y' as function of y , parametrized by t).

For the approximating Runge–Kutta methods, an essential fact is that they are *one-step* methods. This means that the approximative solution at a next time level $t+h$ is calculated from the solution at the given time level t only. But after performing several such steps, we could also use

the *multi-step* methods which use the information from more previous time levels.

The most used multi-step procedures are the Adams–Bashforth (AB2 and AB3) methods,

$$Y^{n+1} = Y^n + h \left(\frac{3}{2}F^n - \frac{1}{2}F^{n-1} \right),$$

$$Y^{n+1} = Y^n + h \left(\frac{23}{12}F^n - \frac{16}{12}F^{n-1} + \frac{5}{12}F^{n-2} \right)$$

and Adams–Moulton (Crank–Nicolson and AM3) methods

$$Y^{n+1} = Y^n + \frac{h}{2} (F^{n+1} + F^n),$$

$$Y^{n+1} = Y^n + \frac{h}{12} (5F^{n+1} + 8F^n - F^{n-1})$$

where $Y^n \approx Y(nh)$ and $F^n = F(Y^n)$.

3.1 Examples

In order to present some very simple examples of the motion of a body, we will follow Chow [22], taking into account also the forces exerted by the surrounding fluid that leads to systems of differential equations.

3.1.1 Falling of a Spherical Body

Let us consider a spherical body, of mass m and diameter d , located at $t = 0$ at the origin of the Oz axis, which is chosen in the direction of the gravitational acceleration. The initial velocity of the body is v'_0 and it moves under the action of the gravitational force mg along the Oz axis. At the moment t the body is at the distance $z(t)$ from the origin and it has the velocity $v(t)$, all these functions satisfying the differential system

$$\frac{dz}{dt} = v(t), \tag{4.6}$$

$$\frac{dv}{dt} = \frac{1}{A} [B - Cv |v| c_d(v)],$$

where $A = 1 + \frac{\bar{\rho}}{2}$, $B = (1 - \bar{\rho})g$, $C = \frac{3\bar{\rho}}{4d}$ and $\bar{\rho} = \frac{\rho_f}{\rho}$, ρ_f being the mass density of the surrounding fluid while ρ is the density of the body.

Here c_d is the (dimensionless) *drag coefficient* which expresses the influence of the viscosity of the fluid. It depends on the shape of the

body and the Reynolds number R and, generally, it is difficult to find it analytically so that some appropriate experiments are used for this purpose. If the fluid has the kinematic viscosity ν , the experimental expression for c_d as a function of the Reynolds number $R = \frac{vd}{\nu}$ (in the case of a smooth sphere) could be approximated by

$$c_d(R) = \begin{cases} \frac{24}{R}, & R \leq 1, \\ \frac{24}{R^{0.646}}, & 1 < R \leq 400, \\ 0.5, & 400 < R \leq 3 \times 10^5, \\ 0.000366R^{0.4275}, & 3 \times 10^5 < R \leq 2 \times 10^6, \\ 0.18, & R > 2 \times 10^6. \end{cases}$$

The particular values for a steel sphere dropping in air (under atmospheric conditions at sea level), are $\rho = 8000\text{kg/m}^3$, $\rho_f = 1.22\text{kg/m}^3$, $\nu = 1.49 \times 10^{-5}\text{m}^2/\text{s}$, $g = 9,8\text{m/s}^2$. Obviously, in vacuum, without any surrounding fluid, $\rho_f = \bar{\rho} = 0$ and the differential system becomes

$$\frac{dz}{dt} = v(t), \quad \frac{dv}{dt} = g$$

with the solution $v(t) = v_0 + gt$, $z(t) = z_0 + v_0t + \frac{1}{2}gt^2$, where z_0 and v_0 are respectively the initial position and velocity.

Now we have a mathematical model of the phenomenon, represented by the system (4.6) together with the initial conditions, so that we are able to perform various numerical experiments. The numerical results are confirmed by physical experiments if we are placed in the domain of the model's validity. The MATLAB programs are:

a) program of function type, computing the coefficient c_d , saved as `drag.m`

```
function cd=drag(Re)
if Re==0 cd=0;
elseif Re>=0 & Re<=1 cd = 24/Re;
elseif Re>1 & Re<=400 cd=24/Re^0.646;
elseif Re>400 &Re<=3.e5 cd=0.5;
elseif Re>3.e5 &Re<=2.e6 cd=3.66e-4*Re^0.4275;
else cd=0.18;
end;
```

b) program of function type describing the system (4.6), saved as `ecdif11.m`

```

function yprim=ecdif11(x,y)
global RO D ROF NU;
yprim=zeros(2,1);
g=9.81;robar=ROF/RO;
a=1+robar/2;b=(1-robar)*g;c=3*robar/4/D;
r=abs(y(2))*D/NU; cd=drag(r);
yprim(1)=y(2);
yprim(2)=(b-c*y(2)*abs(y(2))*cd)/a;
c) the main program, saved as freefall.m
function [t,x]=freefall(ro,d,rof,nu,Tf,z0,v0)
global RO D ROF NU;
RO=ro;D=d;ROF=rof;NU=nu;
[t,x]=ode45(@ecdif11,[0,Tf],[z0 v0]);
plot(t,x(:,2),'.','MarkerSize',12);
xlabel('t(s)');ylabel('v(m/s)');

```

and called up with particular values of the parameters.

The results of numerical simulations with different values of the diameters of the spheres are represented in Figure 4.3 where the time variation of the velocities for some particular diameters are shown.

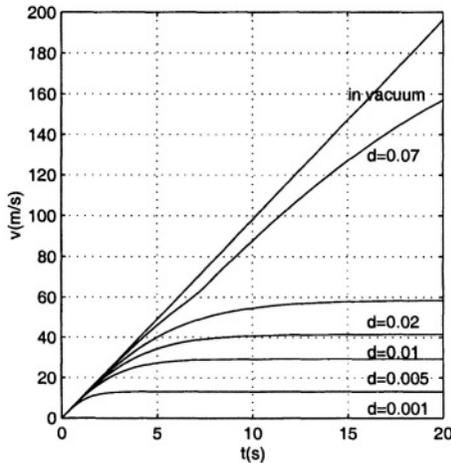


Figure 4.3. Velocities of steel spheres falling in air (for particular diameters)

We remark that after some time the bodies reach a final constant velocity which increases with the diameter of the sphere. For a large sphere, the effect of the viscosity becomes negligible in comparison with body inertia, so that the sphere would behave as if it were moving in

a vacuum. In this case, the velocity increases indefinitely with time without a terminal constant velocity.

The terminal velocity for a particular fluid and diameter d could be calculated by taking to zero the right-hand side of the velocity equation from (4.6), i.e., $v|v|c_d(v) = \frac{B}{C}$. If we plot the values of the expression $v|v|c_d(v)$ and if we remark that for $d = 0.01$ (for instance), under the above conditions, we have $B/C = 857.5741$, then the calculated terminal velocity will be $v_{st} = 41.41m/s$.

Moreover, we remark from the same equation of the velocity (4.6) that if $v > v_{st}$, then the right-hand side of the equation is negative, so the velocity v diminishes and, conversely, if $v < v_{st}$ the right-hand side is positive so that the velocity v increases. This means that v_{st} is a *steady stable solution* of the system (4.6).

We must remark that the above model for the numerical experiments is suitable only for subsonic velocities (for supersonic velocities the effect of the shock waves must be taken into account). Also, if the displacement of the body is large, the variation of the air density is significant and it must be used in the model.

The reader could perform many numerical experiments, for example with a ping-pong ball (with a density supposed to be equal to that of the air) and of diameter $d = 0.036m$, in water, where $\rho_f = 1000kg/m^3$ and $\nu = 1 \times 10^{-6}m^2/s$ while $\rho = 1.22kg/m^3$ or with a glass sphere with $\rho = 2500kg/m^3$, etc.

3.1.2 Ballistic Problem

Let us study now the translation motion of a body through a fluid in the Oxy plane, where the Oy axis is in the opposite direction to that of the gravitational force. The body has a velocity of components (u, v) and the fluid has a velocity of components (u_f, v_f) which depend on the position and time. Assuming a spherical body of diameter d and mass m , the governing equations (which take also into account the specific fluid dynamic forces) are

$$\begin{aligned} \frac{d^2x}{dt^2} &= \frac{3\bar{\rho} c_d(u_f - u)w_r}{4d \left(1 + \frac{1}{2}\bar{\rho}\right)}, \\ \frac{d^2y}{dt^2} &= \frac{-(1 - \bar{\rho})g + \frac{3\bar{\rho}}{4d}c_d(v_f - v)w_r}{1 + \frac{1}{2}\bar{\rho}}, \end{aligned} \tag{4.7}$$

where $w_r = \sqrt{(u_f - u)^2 + (v_f - v)^2}$. We will consider as an example a steel sphere of diameter $d = 0.3m$ moving in the air, starting from the initial position $(0,0)$ with an initial velocity $800m/s$ which makes an

angle (elevation) θ_0 with the horizontal Ox direction. The motion in a vacuum is obtained for $\rho_f = 0$. Moreover, for large initial velocities, the variable density of the air at a higher altitude, must be considered by using, for instance, the function $\rho_f = 1.22e^{-0.000118y} \text{ kg/m}^3$.

The MATLAB subprogram describing the differential system is:

```
function yprim=ecdif14(x,y)
global theta0 cod;
yprim=zeros(4,1); ro=8000;
if cod==1 rof=1.22; else
rof=1.22*exp(-0.000118*y(2));end;
g=9.8;robar=rof/ro;nu=0.0000149; d=0.3;
a=1+robar/2;b=(1-robar)*g;c=3*robar/4/d;
uf=-10;vf=0;wr=sqrt((uf-y(3))^2+(vf-y(4))^2);
cd=0.4;
yprim(1)=y(3);
yprim(2)=y(4);
yprim(3)=c*cd*(uf-y(3))*wr/a;
yprim(4)=(-b+c*cd*(vf-y(4))*wr)/a;
if y(2)<0 yprim(1)=0;yprim(2)=0;end;
```

where an opposite horizontal wind was considered, i.e., $u = -10 \text{ m/s}$ and for simplicity, the drag coefficient was taken as $c_d = 0.4$ (corresponding to the postcalculated Reynolds number, which now depends also on the Mach number). The computation is stopped if the projectile reaches its initial height $y = 0$. The main program, saved as `p14.m`, is the following

```
global theta0 cod;
w0=800; theta0=theta0*pi/180;
cod=1; [t,x]=ode45(@ecdif14,[0 100],...
[0 0 w0*cos(theta0) w0*sin(theta0)]');
plot(x(:,1),x(:,2));xlabel('x');ylabel('y');
axis([0 4000 0 3500]); grid;hold on;
cod=2; [t,x]=ode45(@ecdif14,[0 100],...
[0 0 w0*cos(theta0) w0*sin(theta0)]');
plot(x(:,1),x(:,2),'.');xlabel('x');ylabel('y');
axis([0 4000 0 3500]); grid;
title('rhof variable: ... rhof constant: ---');
hold off;
```

and it is called by the command

```
global theta0 cod;theta0=60;p14;
```

The results are shown in Figure 4.4. We remark the changes in the range depending on the density of the air. Any elevations and wind velocities may be tested and compared with the motion in the vacuum. The program is also useful for other problems, for instance to determine

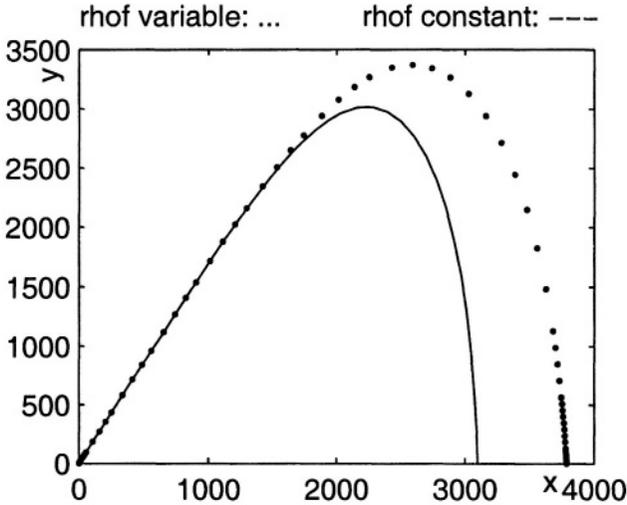


Figure 4.4. The motion of a projectile

the elevation such that the maximum range is reached, for certain given conditions. In this case the suitable drag coefficient must be taken into account, by using the subprogram `drag.m`.

3.1.3 Shock Waves in Viscous Fluids

In a real fluid flow, the velocity and the pressure vary smoothly through a thin shock region instead of jumping, as described in the inviscid theory. Let us study now numerically the structure of a shock in the presence of the viscosity, for a simplified problem.

Suppose the shock propagates at a constant supersonic velocity u_1 along the negative direction of the Ox axis. Let the coordinate system move at the shock wave velocity, so that it becomes steady with respect to this frame. Let us use the subscripts $_1$ and $_2$ for the far upstream, respectively for the far downstream, given quantities.

For a steady one-dimensional flow the continuity, motion and energy equation become respectively

$$\begin{aligned}\frac{d}{dx}(\rho u) &= 0, \\ \rho u \frac{du}{dx} &= -\frac{dp}{dx} + \frac{d}{dx} \left(\mu' \frac{du}{dx} \right), \\ \rho u \frac{d}{dx} \left(c_p T + \frac{u^2}{2} \right) &= \frac{d}{dx} \left(u \mu' \frac{du}{dx} + k \frac{dT}{dx} \right),\end{aligned}$$

where $\mu' = 2\mu + \lambda$, while μ and λ are the viscosity coefficients of the fluid, c_p is the constant-pressure specific heat and k is the thermal conductivity. Integrating with respect to x on an interval containing the shock, we get

$$\begin{aligned}\rho u &= \rho_1 u_1 = m, \\ \mu' \frac{du}{dx} - mu - p &= -mu_1 - p_1, \\ u \mu' \frac{du}{dx} + k \frac{dT}{dx} - m \left(c_p T + \frac{u^2}{2} \right) &= -m \left(c_p T_1 + \frac{u_1^2}{2} \right),\end{aligned}$$

where m is the mass flux through the shock. The left sides of the above equations become, far downstream (where the velocity and the temperature are uniform),

$$\begin{aligned}\rho_2 u_2 &= \rho_1 u_1, \\ m(u_1 - u_2) &= p_2 - p_1, \\ c_p T_2 + \frac{u_2^2}{2} &= c_p T_1 + \frac{u_1^2}{2},\end{aligned}$$

which represent the laws of conservation of mass, momentum and energy across the shock.

The effective integration of the above equations may be generally performed only by numerical methods, after some simplifications. Let us replace the pressure in the state equation (the Clapeyron relation)

$$p = \rho R T = m R \frac{T}{u}$$

where R is the gas constant. Let us replace $\mu' \frac{du}{dx}$ from the obtained equation into the energy equation. Using the dimensionless variables

$$U = \frac{m}{m u_1 + p_1} u, \quad T' = \frac{m^2 R}{(m u_1 + p_1)^2 T},$$

from the relation $\frac{c_p}{R} = \frac{\gamma}{\gamma-1}$, we get the new formulations of the momentum and energy equations, that is

$$\frac{dU}{dx} = \frac{m}{\mu'} \left(U + \frac{T'}{U} - 1 \right),$$

$$\frac{dT'}{dx} = \frac{\gamma-1}{\gamma} \frac{m}{\mu} \text{Pr} \left(\frac{1}{\gamma-1} T' - \frac{1}{2} U^2 + U - \frac{\alpha}{2} \right),$$

where $\text{Pr} = \mu c_p / k$ is the *Prandtl number* and α is the dimensionless parameter

$$\alpha = \frac{2m^2 \left(c_p T_1 + \frac{u_1^2}{2} \right)}{(m u_1 + p_1)^2}.$$

Consider now a simpler case, of a monoatomic gas, so that $\lambda = -\frac{2}{3}\mu$ and $\mu' = \frac{4}{3}\mu, \gamma = \frac{5}{3}$. Finally, we get the equations

$$\frac{dU}{dx} = \frac{3m}{4\mu} \left(U + \frac{T'}{U} - 1 \right),$$

$$\frac{dT'}{dx} = \frac{m}{5\mu} \text{Pr} (3T' - U^2 + 2U - \alpha).$$

The boundary conditions at the end of the shock are

$$\left. \frac{dU}{dx} \right|_{x=\pm\infty} = 0, \quad \left. \frac{dT'}{dx} \right|_{x=\pm\infty} = 0,$$

and the use of these conditions for the above equations yields to an algebraic system for U and T' with the solutions

$$U = \frac{5 \pm \varepsilon}{8}, \quad T' = \frac{15 - \varepsilon^2 \mp 2\varepsilon}{64}$$

where $\varepsilon = \sqrt{25 - 16\alpha}$ characterizes the shock strength. The upper and lower signs give the upstream, respectively the downstream, conditions.

Now we rewrite the above system by introducing the new variables w and t through the relations

$$U = \frac{5 + \varepsilon w}{8}, \quad T' = \frac{15 - \varepsilon^2 + 2\varepsilon t}{64}$$

and thus we obtain the “shock equations”

$$\frac{dw}{dx} = \frac{3m}{4\mu} \frac{2(t+w) - \varepsilon(1-w^2)}{5 + \varepsilon w},$$

$$\frac{dt}{dx} = \frac{m \text{Pr}}{10\mu} [6(t+w) + \varepsilon(1-w^2)].$$
(4.8)

The steady solutions of this system, obtained for $x \rightarrow \pm\infty$, could be deduced by solving the system

$$\frac{3m}{4\mu} \frac{2(t+w) - \varepsilon(1-w^2)}{5 + \varepsilon w} = 0,$$

$$\frac{m \text{Pr}}{10\mu} [6(t+w) + \varepsilon(1-w^2)] = 0,$$

which leads to

$$P_1 : w = 1, t = -1, \quad P_2 : w = -1, t = 1$$

where P_1 represents the upstream and P_2 the downstream conditions. Computing the Jacobian of the left side functions at the two points $P_{1,2}$ for the particular data $\text{Pr} = \frac{2}{3}$ and $\varepsilon = 1.77$, we find that at P_1 there are two real positive eigenvalues, so it is an unstable node, while at P_2 there are one positive $\lambda_1 = 0.6837$ and one negative $\lambda_2 = -0.6412$ eigenvalue, so it is a saddle point. In this case, the heteroclinic trajectory joining the two steady points must be numerically calculated from P_2 towards P_1 , i.e., downstream towards upstream, in the decreasing of x direction. This trajectory is a stable manifold for P_2 and it is tangent at P_2 to the linear stable subspace generated by the eigenvector of the Jacobian corresponding to the negative eigenvalue $\mathbf{v} = (-0.8534, 0.5213)$.

The calculation could be even more simplified by dividing the equations (4.8), thus obtaining

$$\frac{dt}{dw} = \frac{2}{15} \text{Pr}(5 + \varepsilon w) \frac{6(t+w) + \varepsilon(1-w^2)}{2(t+w) - \varepsilon(1-w^2)}, \tag{4.9}$$

i.e., a unique differential equation which will be integrated from $w = -1$ towards $w = 1$ with the Cauchy condition $\left. \frac{dt}{dw} \right|_{w=-1} = -\frac{0.5213}{0.8534} = -0.6109$ for our particular case. Of course, we do not start exactly from the critical point but from a neighboring (towards the stable manifold direction) point $w = -1 + 0.001, t = 1 - 0.6109 \times 0.001$.

The numerical results could be compared with the experimental (wind tunnel) ones. We will introduce the dimensionaless distance

$$X = \frac{\frac{2}{5} \text{Pr}}{1 + 2 \text{Pr}} \frac{m\varepsilon}{\mu^*} x$$

where the reference viscosity coefficient μ^* is to be evaluated at the temperature $T^* = 3T_0/4$, T_0 being the constant upstream temperature of the fluid. Finally, we have

$$\frac{dX}{dw} = \frac{\frac{8}{15} \varepsilon \text{Pr}}{1 + 2 \text{Pr}} [0.076(15 - \varepsilon^2 + 2\varepsilon t)]^{0.647} \frac{5 + \varepsilon w}{2(t+w) - \varepsilon(1-w^2)}.$$

This equation will be joined with the equation (4.9), together with the Cauchy condition $X = 3.30$ for $w = -0.999$, deduced from the experiments and which determines the X coordinates. The results are shown in Figure 4.5.

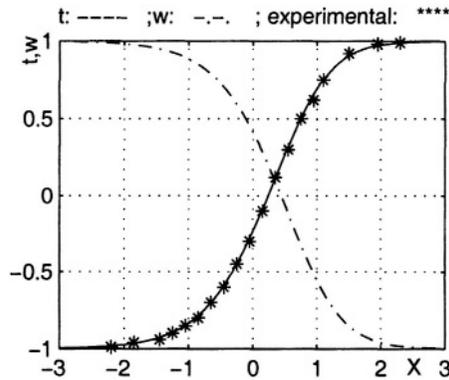


Figure 4.5. Shock waves in viscous fluids

The MATLAB program is:

```
[t,x]=ode45(@edsoc,-1+0.001,1,[1-0.001*0.6109,3.30]);
plot(x(:,2),x(:,1),'-',x(:,2),t,'-');
axis([-3 3 -1 1]);grid;hold on;
plot([-2.2 -1.85 -1.45 -1.25 -1.05 -0.85 -0.65...
-0.45 -0.25 -0.05 0.15 0.35 0.55 0.75 0.95 1.1...
1.5 1.95 2.30],[-0.99 -0.96 -0.94 -0.90 -0.85...
-0.80 -0.70 -0.60 -0.45 -0.30 -0.10 0.12 0.30...
0.50 0.62 0.75 0.92 0.98 0.99],'*');hold off;
xlabel('X');ylabel('t,w');
title('t: ---- ;w: -.-. ; experimental: ****');
which uses the function subprogram edsoc.m
function yprim=edsoc(t,y)
yprim=zeros(2,1);e=1.77;pr=2/3;
yprim(1)=2/15*pr*(5+e*t)*(6*(y(1)+t)+e*(1-t^2))...
/(2*(y(1)+t)-e*(1-t^2));
yprim(2)=8/15*e*pr/(1+2*pr)*((0.076*(15-e^2+...
2*e*y(1))^0.647*(5+e*t))/(2*(y(1)+t)-e*(1-t^2));
```

We remark an excellent agreement between the numerical simulation and physical experiment results concerning the structure of the shock wave. The reference length x/X in this particular case is $0.0013m$ so that the shock interval is of length $0.68cm$. See [22] for more details.

4. Partial Differential Equations

4.1 Classification of Partial Differential Equations

Different phenomena are governed by partial differential equations of different structures and types. For example, the inviscid compressible fluid flow (in a subsonic regime) around a body could be described, by linearization, with the equation

$$(1 - M^2)\Phi_{xx} + \Phi_{yy} = 0,$$

where Φ is the velocities potential and $M < 1$ is the *Mach number* (the ratio of the fluid velocity and the sound speed). In this case, the perturbation generated by the presence of the body propagates in all directions.

In the supersonic case, for $M > 1$, the above equation changes its type, the two coefficients being now of different sign. Physically, the fluid in its motion goes beyond the perturbations produced in front of the body and thus a perturbation region appears only behind the body, bounded by two straight lines — the characteristics of the partial differential equation. On the characteristics, the first derivatives of the components of the velocity are different from one side to another, due to the fact that the perturbations exist only at one side so that the second order derivatives of the velocity potential Φ are not defined on these lines.

The type of a second order partial differential equation is induced by the existence (reality) of these characteristics. Suppose that the equation of Φ is

$$A\Phi_{xx} + 2B\Phi_{xy} + C\Phi_{yy} = D \quad (4.10)$$

where A, B, C, D could be functions of $x, y, \Phi, \Phi_x, \Phi_y$ (Monge equation). The variations of the velocity components Φ_x, Φ_y passing from (x, y) to $(x + dx, y + dy)$ are given by

$$\begin{aligned} dx\Phi_{xx} + dy\Phi_{xy} &= d\Phi_x, \\ dx\Phi_{xy} + dy\Phi_{yy} &= d\Phi_y. \end{aligned}$$

Let us now consider the above three relations as a system having as unknowns the second order derivatives of Φ , taking into account the fact that along the characteristics these derivatives are not defined. Therefore the determinant of the system must vanish

$$\begin{vmatrix} A & B & C \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} = 0,$$

i.e., we have the differential relation

$$A \left(\frac{dy}{dx} \right)^2 - 2B \frac{dy}{dx} + C = 0.$$

Consequently, on the characteristics we can write

$$\frac{dy}{dx} = \frac{B \pm \sqrt{B^2 - AC}}{A}.$$

There are three different cases.

a) If $B^2 - AC > 0$, then through every point (x, y) from the computational domain, two characteristics pass (like the case of the supersonic flow) and the equation (4.10) is called *of hyperbolic type*. For example, the equations describing oscillations, particularly the wave equation, are of this type;

b) If $B^2 - AC < 0$, then there are no real characteristics. These equations are of *elliptic type*, like the equation for the subsonic flow case or the Laplace or the Poisson equations;

c) If $B^2 - AC = 0$, there exists through every point of the computational domain only one real characteristic and the equation is of *parabolic type*. The equations describing diffusion or dissipation phenomena are of this type.

We remark that these types of equations describe not only different types of phenomena but also their solutions are of different types and can be numerically found by using different techniques.

In the case of systems of partial differential equations we have a similar situation. Let

$$\begin{aligned} a_1 u_x + b_1 u_y + c_1 v_x + d_1 v_y &= f_1, \\ a_2 u_x + b_2 u_y + c_2 v_x + d_2 v_y &= f_2 \end{aligned} \quad (4.11)$$

be such a system, where a_1, \dots, d_2 and f_1, f_2 are functions of x, y, u, v .

Being placed at a point in the Oxy plane, let us seek the directions along which the derivatives of u and v are not determined — the so-called *characteristic lines*. If we add to the above system (4.11) the equations

$$\begin{aligned} u_x dx + u_y dy &= du, \\ v_x dx + v_y dy &= dv, \end{aligned} \quad (4.12)$$

we see that u_x, u_y, v_x, v_y could be undetermined only if the determinant

$$A = \begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{vmatrix}$$

is zero. Therefore

$$ady^2 + bdx dy + cdx^2 = 0$$

where

$$\begin{aligned} a &= a_1c_2 - a_2c_1, \\ b &= -(a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1), \\ c &= b_1d_2 - b_2d_1, \end{aligned}$$

or, in other form,

$$\frac{dy}{dx} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

The above equations give the directions of the characteristic lines through the current point (x, y) . As in the case of a single equation, we have three situations:

a) $b^2 - 4ac > 0$, the system is *hyperbolic* and we have two characteristic curves through (x, y) ,

b) $b^2 - 4ac = 0$, the system is *parabolic* and we have a single characteristic curve through the given point

and

c) $b^2 - 4ac < 0$, the system is *elliptic* and we have no real characteristic lines through that point.

We remark that in the hyperbolic case, if we try to solve the above system with respect to the derivatives of u and v (by Cramer's rule, for instance) we are led to an undetermination only when the respective numerators are also zero. So that we obtain the equations

$$\begin{vmatrix} f_1 & b_1 & c_1 & d_1 \\ f_2 & b_2 & c_2 & d_2 \\ du & dy & 0 & 0 \\ dv & 0 & dx & dy \end{vmatrix} = 0, \text{ etc.}$$

which are, in fact, *differential equations* for the variables u and v . These equations are valid only on the characteristic lines and the integration of the system reduces, in fact, to the integration of these differential equations.

4.2 The Behaviour of Different Types of PDE

a) *Hyperbolic equations.* In this case the information from the point P of the computational domain influences only the region between the characteristics through P , see Figure 4.6.

The value of the solution at P is influenced only by the values of the data on the interval (a, b) between the characteristics through P . The inviscid steady supersonic fluids and the inviscid compressible subsonic unsteady fluids are described by such type of equations. For the unsteady

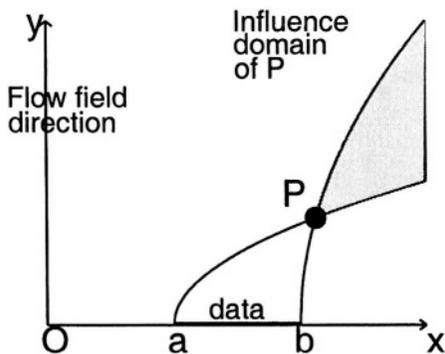


Figure 4.6. The influence domain for the hyperbolic case

case the role of the Oy axis is taken by the time axis and its direction is also a flow field direction.

b) *Parabolic equations.* The value of the solution at the point P from the plane Oxy influences the whole region of the plane to one side of the characteristic through P , see Figure 4.7.

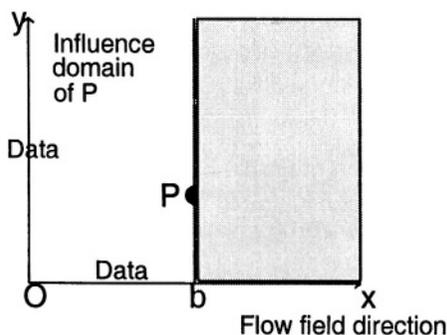


Figure 4.7. The influence domain for the parabolic case

If the axes Ox, Oy are the boundaries of the computational domain, the solution of the equation at P depends on the values of the data on the semiaxis Oy and on the semiaxis Ox from O to b . This solution could be calculated starting from the data and marching in the flow field direction (here the Ox direction). Some reduced forms of the Navier–Stokes equations (for example the Stokes system) and the boundary layer problems are of such a type.

c) *Elliptic equations.* The information from P influences the entire computational domain. The value of the solution at P depends on the data on the entire boundary $Obcd$, see Figure 4.8.

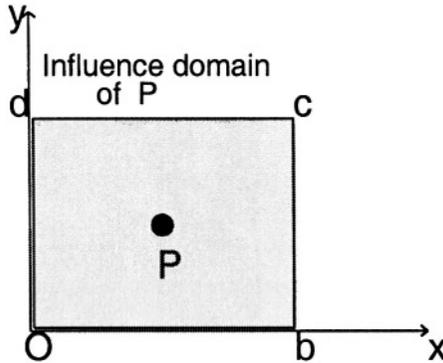


Figure 4.8. The influence domain for the elliptic case

What is specific for this case is the fact that the solution at P must be calculated simultaneously with the solutions at all the points from the computational domain. This is a different procedure than that for the parabolic and hyperbolic cases where the information marches from the data of the problem in the flow field direction to the solution at other points. Based on this fact, the elliptic problems are also-called *equilibrium problems*.

The subsonic steady inviscid and the incompressible fluid flows are governed by equations of this type. On the boundary we could have *Dirichlet* type conditions, when the values of u, v are given or *Neumann* conditions, when the values of the derivatives $\frac{\partial u}{\partial x}, \dots$ are given. Of course, mixed conditions are also used.

d) The same problem may lead to equations which are of different types in different regions. For example, the supersonic motion of a blunt body through the atmosphere (or, the same thing, the supersonic air flow past that body) shows a region with supersonic velocity, with $M > 1$ and, in front of the body, a region with a local subsonic velocity, with $M < 1$ so we are in a transonic case. In the first case the fluid flow is described by a hyperbolic equation and in the second case by an elliptic equation, see Figure 4.9.

The method of a simultaneous treatment of the two regions requires that, starting with the given initial conditions, one marches in time considering the unsteady equations which determine the fluid flow. After

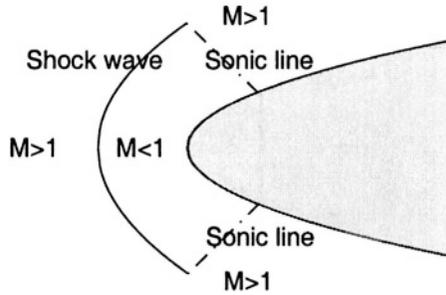


Figure 4.9. The transonic case

a long time, the solution approaches the steady state which describes the fluid flow into both regions, the super and subsonic regions.

We also remark that if we try to solve a problem with wrong or incomplete initial and boundary conditions, the numerical solutions could be obtained but these are spurious solutions, without physical relevance. A problem is *well-posed* in the Hadamard sense if its solution exists, it is unique and it depends continuously on data. It is important to know this fact before taking the numerical approach on the respective problem.

4.3 Burgers' Equation

We shall now consider, following [42], the nonlinear equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = 0$$

written in the conservative form which could be rewritten into the non-conservative form

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0. \quad (4.13)$$

These two forms are equivalent in the continuous approach but of different behavior in the discrete (by finite differences) approach. We remark the analogy between the nonconservative form and the linear advection equation, but now the advection velocity is no longer constant, depending on the solution u . The initial shape

$$u(x, 0) = u_0(x) \quad (4.14)$$

distorts at the next time levels. More precisely, the points where u is greater are moving faster in a direction given by the sign of u .

4.3.1 Classical and Weak Solutions

If we choose a point on a curve $x = x(t)$ of the plane (x, t) and we calculate the total derivative of u on it, we find

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + p \frac{\partial u}{\partial x}$$

where $p = \frac{dx}{dt}$.

We remark that the derivative $\frac{du}{dt}$ vanishes in the direction of slope u if and only if u is a solution of Burgers' equation. If we consider the family of straight lines indexed by a parameter ξ ,

$$x = a(\xi)t + b(\xi),$$

and impose the condition $u = a(\xi)$, then $u(x, t)$ is constant on each straight line. These straight lines are, in fact, the characteristic curves of the equation.

The solution of the Cauchy problem (4.13)+(4.14) can be given as follows: Through the point $(\xi, 0)$ of the Ox axis passes a single straight line of slope $u_0(\xi)$, of equation

$$x = u_0(\xi)t + \xi. \quad (4.15)$$

On this characteristic line the solution u is of a constant value, the value at the point ξ of the Ox axis,

$$u = u_0(\xi). \quad (4.16)$$

The equations (4.15)+(4.16) constitute a parametric representation of the solution $u(x, t)$ of the Cauchy problem. Theoretically, from the equation (4.15) we obtain ξ as a function of x and t and replacing it into the equation (4.16) we obtain the analytical form of the solution $u(x, t)$.

We remark that for the linear advection equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0$$

the characteristic curves were the parallel straight lines $x - ct = \text{const.}$ For Burgers' equation the characteristic curves are straight lines too but, generally, they are nonparallel; the slopes depend on the value of the solution at the considered point. This is an effect of the nonlinearity of the equation.

Let us consider three examples of different initial conditions in order to point out this phenomenon.

Example 1.

$$u_0(\xi) = \begin{cases} 0, & \xi \leq 0 \\ \xi, & \xi > 0. \end{cases}$$

As above, the parametric form of the solution is

$$\begin{cases} x = \xi, & u = 0, & \xi \leq 0, t > 0 \\ x = \xi t + \xi, & u = \xi, & \xi > 0, t > 0 \end{cases}$$

from where, by eliminating ξ , we obtain

$$u(x, t) = \begin{cases} 0, & x \leq 0, t > 0 \\ \frac{x}{t+1}, & x > 0, t > 0. \end{cases}$$

This is a continuous, piecewise derivable solution and its regularity is similar to the regularity of the initial profile. The derivative discontinuity moves on the characteristic curve $x = 0$.

Example 2.

$$u_0(\xi) = \begin{cases} 0, & \xi < 0 \\ 1, & \xi > 0 \end{cases}$$

Here u_0 has a discontinuity at the origin and let us consider for u_0 at this point, all the values α between 0 and 1. As above, the parametric form of the solution is

$$\begin{cases} x = \xi, & u = 0, & \xi < 0, t > 0 \\ x = t + \xi, & u = 1, & \xi > 0, t > 0 \\ x = \alpha t, & u = \alpha, & \xi = 0, t > 0, \alpha \in [0, 1] \end{cases}$$

Eliminating the parameters ξ and α from the above equations, we obtain

$$u(x, t) = \begin{cases} 0, & x < 0, t > 0 \\ 1, & x > t > 0 \\ \frac{x}{t}, & x \in [0, t], t > 0 \end{cases}$$

In this case, the initial shape is discontinuous at the origin. From this point we have, in the plane (x, t) , a set $x = \alpha t$ of characteristic curves and the Cauchy problem solution is still continuous in the halfplane $t > 0$.

Example 3.

$$u_0(\xi) = \begin{cases} 0, & \xi < 0 \\ -\xi^2, & \xi > 0 \end{cases}$$

If in the previous cases u_0 was a monotonically increasing function, now u_0 is a monotonically decreasing function. The characteristic slopes decrease, because

$$\frac{dx}{dt} = u_0(\xi) = -\xi^2, \xi > 0.$$

Consequently, the characteristics intersect in the halfplane $t > 0$. But, on each characteristic, u is of the constant value coming from the Ox

axis and therefore at the intersection point of the characteristics u must take different values. This is possible only if we accept *discontinuous solutions* of the equation. These solutions appear although the initial profile was a continuous differentiable function.

Such discontinuities appear in the physical phenomenon described by the Burgers equation. In gas dynamics, for example, they are called shocks or shock waves. For their mathematical characterization we need the notion of *weak solution*, which allows the discontinuities, see section 1.3.5. The *shock condition* becomes

$$\left. \frac{dx}{dt} \right|_{\Sigma} = \frac{u_1 + u_2}{2}, \quad (4.17)$$

that is the slope of the shock is the average of the values on its sides.

Example 4. Let us consider now the initial profile

$$u_0(\xi) = \begin{cases} 1, & \xi < 0 \\ 0, & \xi > 0. \end{cases}$$

The solution is (in parametric form)

$$\begin{cases} x = t + \xi, & u = 1, & \xi < 0, t > 0 \\ x = \xi, & u = 0, & \xi > 0, t > 0. \end{cases}$$

But u_0 decreases, so the characteristics intersect themselves and a shock appears, beginning, in this case, from the origin. Its slope is $\frac{dx}{dt} = \frac{1+0}{2}$ so the shock's equation is $x = \frac{t}{2}$.

The solution of the Cauchy problem is therefore

$$u(x, t) = \begin{cases} 1, & x < \frac{t}{2}, t > 0 \\ 0, & x > \frac{t}{2}, t > 0 \end{cases}$$

and we remark that there is a discontinuity at $x = \frac{t}{2}$.

The extension of the notion of solution allows significant physical results even in the case of decreasing initial shapes. Conversely, the uniqueness of the solution is lost.

If we resume Example 2, for which

$$u_0(\xi) = \begin{cases} 0, & \xi < 0 \\ 1, & \xi > 0 \end{cases}$$

we easily remark that together with the continuous solution

$$u(x, t) = \begin{cases} 0, & x < 0, t > 0 \\ 1, & x > t > 0 \\ \frac{x}{t}, & x \in [0, t], t > 0 \end{cases}$$

we also have a discontinuous solution

$$u(x, t) = \begin{cases} 0, & x < \frac{t}{2}, t > 0 \\ 1, & x > \frac{t}{2}, t > 0 \end{cases}$$

which verifies the equation on subdomains, together with the initial condition and the shock condition. But this discontinuous solution does not verify the entropy condition (see again section 3.5, Chapter 1) and, of course, it has no physical significance and must be eliminated.

The following theorem can be proved:

THEOREM 4.2. *If the initial profile is a bounded and measurable function, then the Cauchy problem for the Burgers equation has a unique entropy solution.*

We conclude:

- a) the elliptic or parabolic equations cannot allow shocks,
- b) the linear hyperbolic equations allow shocks only if these exist in the initial or boundary conditions,
- c) the nonlinear hyperbolic equations allow shocks, even without discontinuities in the problem's data.

4.3.2 Burgers' Equation with Dissipative Term

Let us now consider the equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = v \frac{\partial^2 u}{\partial x^2} \quad (4.18)$$

where v is a positive constant. This is a parabolic equation, and it may be considered as derived from the diffusion equation with a convective term $\frac{\partial}{\partial x} \left(\frac{u^2}{2} \right)$ or derived from the Burgers equation with a dissipative term $v \frac{\partial^2 u}{\partial x^2}$. Generally, v is considered small, so we have in fact a singularly perturbed problem. This equation is often used for testing numerical methods because it is a model of Navier–Stokes equations.

Looking for stationary solutions of this equation we consider the differential equation

$$\frac{d}{dx} \left(\frac{u^2}{2} \right) = v \frac{d^2 u}{dx^2}$$

and we obtain, by integration,

$$\frac{u^2}{2} = v \frac{du}{dx} \pm \frac{C^2}{2}.$$

Choosing the + sign and $C > 0$ we get the differential equation

$$dx = \frac{v}{C} \left(\frac{du}{u - C} - \frac{du}{u + C} \right)$$

which yields

$$u(x) = C \frac{1 + Ke^{\frac{Cx}{v}}}{1 - Ke^{\frac{Cx}{v}}}.$$

In this last form of the solution we consider $K < 0$ in order to focus on the solutions defined on \mathbb{R} . For $K = -1$ these solutions are $u(x) = -C \tanh\left(\frac{Cx}{2v}\right)$, i.e., they are decaying functions from C to $-C$, and their slope at the origin tends to $-\infty$ as $v \searrow 0$. At the limit we obtain a shock (a discontinuity verifying the entropy condition). We have

THEOREM 4.3.

- a) *The problem (4.18) + (4.14) has a unique regular solution for $t > 0$;*
 b) *This solution tends, as $v \searrow 0$, to the weak solution of the problem (4.13) + (4.14) verifying the entropy condition $u(x - 0, t) \geq u(x + 0, t)$ for all $x \in \mathbb{R}$ and $t > 0$.*

4.4 Stokes' Problem

A very important and much studied example, which introduces the difficulties of the Navier–Stokes system is the Stokes problem, which means

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} &= \frac{1}{R} \Delta u, & \text{in } \Omega, \\ \frac{\partial v}{\partial t} + \frac{\partial p}{\partial y} &= \frac{1}{R} \Delta v, & \text{in } \Omega, \\ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0, & \text{in } \Omega, \\ (u, v)|_{\partial\Omega} &= (u_{fr}, v_{fr}), & \text{on } \partial\Omega, \\ (u, v)|_{t=0} &= (u_0, v_0), & \text{in } \Omega, \end{aligned} \tag{4.19}$$

where (u, v) are the components of the velocity flow, p is the pressure and R the Reynolds number. We remark the lack of a boundary condition for the pressure and the presence of the equation $\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$ at every time instant, see also section 3, Chapter 3.

4.4.1 Direct Solving

We will present here, following [126], a very important direct method of Glowinski and Pironneau to solve the Stokes system. Let us consider

the problem

$$\begin{aligned} (-\nabla^2 + \gamma)\mathbf{u} + \nabla p &= \mathbf{g}, \\ \nabla \cdot \mathbf{u} &= 0, \\ \mathbf{u}|_S &= \mathbf{b}, \end{aligned} \quad (4.20)$$

which defines the Stokes problem on the tridimensional domain V and where $S = \partial V$, coming from the temporal discretization of the linearized incompressible equations.

The particularity of the method is the introduction, besides the Poisson equation $-\nabla^2 p = -\nabla \cdot \mathbf{g}$ for the pressure, of another Poisson equation for a scalar unknown ψ ,

$$-\nabla^2 \psi = \nabla \cdot \mathbf{u}, \psi|_S = 0. \quad (4.21)$$

By applying to that equation the operator $(-\nabla^2 + \gamma)$ we remark that ψ is a solution of the fourth order elliptic equation

$$(-\nabla^2 + \gamma)\nabla^2 \psi = 0.$$

It means that we may ensure $\nabla \cdot \mathbf{u} = 0$ if the solution of the equation (4.21) is $\psi = 0$. But the solution ψ of the fourth order equation will be $\psi = 0$ if $\psi|_S = 0$ and $\frac{\partial \psi}{\partial n}\Big|_S = 0$.

Consequently, the equation and the conditions of the Stokes problem (4.20) will be fulfilled by the solutions p and \mathbf{u} of the system

$$\begin{aligned} -\nabla^2 p &= -\nabla \cdot \mathbf{g}, \\ (-\nabla^2 + \gamma)\mathbf{u} &= -\nabla p + \mathbf{g}, \mathbf{u}|_S = \mathbf{b}, \\ -\nabla^2 \psi &= \nabla \cdot \mathbf{u}, \psi|_S = 0, \end{aligned} \quad (4.22)$$

if the auxiliary unknown ψ verifies also the Neumann condition $\frac{\partial \psi}{\partial n}\Big|_S = 0$.

We remark that the last condition is a substitute for the non-existent boundary condition for the pressure. In order to determine the boundary condition for p which ensures the fulfillment of the incompressibility equation $\nabla \cdot \mathbf{u} = 0$, we will consider the system

$$\begin{aligned} -\nabla^2 p_\lambda &= -\nabla \cdot \mathbf{g}, p_\lambda|_S = \lambda, \\ (-\nabla^2 + \gamma)\mathbf{u}_\lambda &= -\nabla p_\lambda + \mathbf{g}, \mathbf{u}_\lambda|_S = \mathbf{b}, \\ -\nabla^2 \psi_\lambda &= \nabla \cdot \mathbf{u}_\lambda, \psi_\lambda|_S = 0, \end{aligned}$$

and we calculate λ for which $\left. \frac{\partial \psi_\lambda}{\partial n} \right|_S = 0$. Here λ is an unknown defined on the surface S and which is supposed to be of null average in order to fix the indetermination (up to an additive constant) of p .

The condition $\left. \frac{\partial \psi_\lambda}{\partial n} \right|_S = 0$ is next rewritten in a variational (integral) form

$$-\oint \frac{\partial \psi_\lambda}{\partial n} \mu dS = 0 \tag{4.23}$$

for every function μ defined and of null average on S .

By using the Green formula for ∇^2 (which transforms the surface integral into a volume integral), the equations of the system (4.22) and the similar equations for a system for μ , the integral from (4.23) may be written as

$$-\oint \frac{\partial \psi_\lambda}{\partial n} \mu dS = \int (\gamma \mathbf{u}_\lambda \cdot \mathbf{u}_\mu + \nabla \times \mathbf{u}_\lambda \cdot \nabla \times \mathbf{u}_\mu + \nabla \cdot \mathbf{u}_\lambda \nabla \cdot \mathbf{u}_\mu) dV$$

which shows the symmetry of that integral. However, it is useless for calculations because of the necessity to record the values of \mathbf{u}_λ for each function λ . It is more workable to use the decomposition of the solution (p, \mathbf{u}, ψ) into

$$\begin{pmatrix} p(x) \\ \mathbf{u}(x) \\ \psi(x) \end{pmatrix} = \begin{pmatrix} p^0(x) \\ \mathbf{u}^0(x) \\ \psi^0(x) \end{pmatrix} + \oint \begin{pmatrix} p'(x; \sigma') \\ \mathbf{u}'(x; \sigma') \\ \psi'(x; \sigma') \end{pmatrix} \lambda(\sigma') dS(\sigma')$$

where p', \mathbf{u}', ψ' are solutions, for every $\sigma' \in S \setminus \sigma^*$ (σ^* being an arbitrary fixed point on S), of the three elliptic problems

$$\begin{aligned} -\nabla^2 p' &= 0, \quad p'|_S = \delta^{(2)}(s - \sigma') - \delta^{(2)}(s - \sigma^*), \\ (-\nabla^2 + \gamma) \mathbf{u}' &= -\nabla p', \quad \mathbf{u}'|_S = 0, \\ -\nabla^2 \psi' &= \nabla \cdot \mathbf{u}', \quad \psi'|_S = 0, \end{aligned}$$

and $p^0, \mathbf{u}^0, \psi^0$ are solutions of the problems

$$\begin{aligned} -\nabla^2 p^0 &= -\nabla \cdot \mathbf{g}, \quad p^0|_S = 0, \\ (-\nabla^2 + \gamma) \mathbf{u}^0 &= -\nabla p^0 + \mathbf{g}, \quad \mathbf{u}^0|_S = \mathbf{b}, \\ -\nabla^2 \psi^0 &= \nabla \cdot \mathbf{u}^0, \quad \psi^0|_S = 0. \end{aligned}$$

Here $\delta^{(2)}$ is the Dirac function on S for a tridimensional domain V .

Instead of the functions μ defined only on S , we will introduce the auxiliary scalar functions $w(x; \sigma')$ defined on V by

$$\begin{aligned} w(x; \sigma') &\text{ arbitrary on } V, \\ w(x; \sigma')|_S &= \delta^{(2)}(s - \sigma') - \delta^{(2)}(s - \sigma^*). \end{aligned}$$

With these functions, the problem (4.23) will be transformed into the linear problem

$$\tilde{A}\lambda = \beta$$

where

$$\begin{aligned} \tilde{A}(\sigma, \sigma') &= - \int (\nabla\psi' + \mathbf{u}') \cdot \nabla w \, dV, \\ \beta(\sigma) &= \int (\nabla\psi^0 + \mathbf{u}^0) \cdot \nabla w \, dV. \end{aligned}$$

Practically, the functions $w(x; \sigma)$ may be taken nonvanishing only at a sharp region in the neighborhood of the boundary S , which leads to a more efficient evaluation of the integrals. We remark also that the linear operator \tilde{A} is a symmetrical one, so the algebraic system of equations may be iteratively solved by the efficient conjugate gradient method.

4.5 The Navier–Stokes System

Let us consider a bidimensional domain Ω (the extension to tridimensional domains is immediate) and the Navier–Stokes system written in the form

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \nabla p &= -(\mathbf{u} \cdot \nabla) \mathbf{u} + \nu \nabla^2 \mathbf{u}, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0, & \text{in } \Omega, \\ \mathbf{u}|_{\partial\Omega} &= \mathbf{u}_f. \end{aligned} \quad (4.24)$$

Almost all the numerical procedures to solve a system of this form use the *fractional step method*. The velocity \mathbf{u} is advanced in time by an approximation of the first equation, obtaining an “intermediate” velocity. It is then used in an elliptic equation which imposes the incompressibility condition and determines the pressure at the end of the time step.

We can remark that the usual methods are (time convergence) of second order for the velocity but only of first order for the pressure. In the sequel we will describe a particular numerical method and we will show how one can obtain a complete second order (in time) accuracy.

As in Chapter 3, the basic theorem is that of Ladyzhenskaya, as a particular case of the orthogonal decomposition results of Hodge.

THEOREM 4.4. *Every vectorial field v defined on the domain Ω allows a unique orthogonal decomposition $\mathbf{v} = \mathbf{w} + \nabla\Phi$, where \mathbf{w} is a solenoidal field with a zero normal component to the boundary $\partial\Omega = S$.*

If we return to the system (4.24), we remark that the first equation is such a decomposition and it may be rewritten as

$$\frac{\partial \mathbf{u}}{\partial t} = \mathcal{P} [- (\mathbf{u} \cdot \nabla) \mathbf{u} + \nu \nabla^2 \mathbf{u}]$$

where \mathcal{P} is an operator which projects a vectorial field on the space of the solenoidal vector fields, with suitable boundary conditions.

By half-discretization in time, the equations (4.24) become

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \nabla p^{n+\frac{1}{2}} &= - [(\mathbf{u} \cdot \nabla) \mathbf{u}]^{n+\frac{1}{2}} + \frac{\nu}{2} \nabla^2 (\mathbf{u}^{n+1} + \mathbf{u}^n), \quad (4.25) \\ \nabla \cdot \mathbf{u}^{n+1} &= 0, \\ \mathbf{u}^{n+1}|_{\partial\Omega} &= \mathbf{u}_{fr}^{n+1}. \end{aligned}$$

Here $[(\mathbf{u} \cdot \nabla) \mathbf{u}]^{n+\frac{1}{2}}$ represents a second order approximation at the time level $t^{n+1/2}$, which is usually explicitly calculated.

The above half-discretized problem is solved by a fractional step procedure. From the first equation we determine an “intermediate” velocity \mathbf{u}^* , which is then projected on the space of divergence free vectorial fields, obtaining \mathbf{u}^{n+1} . A typical algorithm is of the form

Step I. We solve for \mathbf{u}^* ,

$$\begin{aligned} \frac{\mathbf{u}^* - \mathbf{u}^n}{\Delta t} + \nabla q &= - [(\mathbf{u} \cdot \nabla) \mathbf{u}]^{n+\frac{1}{2}} + \frac{\nu}{2} \nabla^2 (\mathbf{u}^* + \mathbf{u}^n), \quad (4.26) \\ B(\mathbf{u}^*) &= 0, \end{aligned}$$

where q is an approximation of $p^{n+\frac{1}{2}}$ and $B(\mathbf{u}^*)$ is a boundary condition for \mathbf{u}^* , which can be specified depending on the particular method.

Step II. We project \mathbf{u}^* on the solenoidal fields space

$$\begin{aligned} \mathbf{u}^* &= \mathbf{u}^{n+1} + \Delta t \nabla \Phi^{n+1}, \quad (4.27) \\ \nabla \cdot \mathbf{u}^{n+1} &= 0, \end{aligned}$$

with boundary value conditions consistent with $B(\mathbf{u}^*) = 0$ and $\mathbf{u}^{n+1}|_{\partial\Omega} = \mathbf{u}_{fr}^{n+1}$.

Step III. We update the pressure

$$p^{n+\frac{1}{2}} = q + L (\Phi^{n+1})$$

where L represents the dependence of $p^{n+\frac{1}{2}}$ with respect to Φ^{n+1} .

In the sequel, we pass to the next time level. Such type of methods are called *projection methods*. Particular methods should be pointed out

1. by the approximation of the pressure q ,
2. by choosing the boundary condition $B(u^*)$,
3. by choosing the function $L(\Phi^{n+1})$.

These three approaches must be correlated in order to obtain a second order accuracy of the method. For instance, the boundary condition for u^* must be consistent with the first equation (4.27) but the function Φ^{n+1} is not yet calculated at this instant and should be approximated, depending on the choice of q . Similarly, replacing the first equation of (4.27) into the first equation of (4.26), by eliminating u^* and comparing with the first equation from (4.25) we obtain an update for the pressure

$$p^{n+\frac{1}{2}} = q + \Phi^{n+1} - \frac{\nu \Delta t}{2} \nabla^2 \Phi^{n+1}.$$

This update must be taken into account in order to obtain a second order accuracy for the pressure, on the boundary too, and in order to eliminate the spurious modes for the pressure.

The choice of the boundary conditions may be better understood by referring to an alternative formulation of the Navier–Stokes equations. Let there be new variables \mathbf{m} and χ , connected with the flow velocity by the relationship

$$\mathbf{m} = \mathbf{u} + \nabla \chi \tag{4.28}$$

and so that \mathbf{u} and p obey the Navier–Stokes equations. For instance, we require that \mathbf{m} verify on Ω ,

$$\begin{aligned} \frac{\partial \mathbf{m}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} &= \nu \nabla^2 \mathbf{m}, \\ \mathbf{u}|_{\partial\Omega} &= \mathbf{u}_{fr}, \end{aligned} \tag{4.29}$$

where

$$\mathbf{u} = \mathcal{P}(\mathbf{m}). \tag{4.30}$$

The equations (4.28,4.29,4.30) constitute an equivalent formulation of the Navier–Stokes equations, where the pressure was eliminated. It could be calculated, if it is necessary, from the relationship

$$p = \frac{\partial \chi}{\partial t} - \nu \nabla^2 \chi \tag{4.31}$$

obtained by comparison of the first equation (4.29) with the first equation (4.24). It is easy to remark that even the boundary conditions are given for \mathbf{u} , the equation (4.28) shows that there is a coupling of the boundary conditions for \mathbf{m} and $\nabla \chi$.

The time half-discretized form for the above equations is

$$\frac{\mathbf{m}^{n+1} - \mathbf{m}^n}{\Delta t} = - [(\mathbf{u} \cdot \nabla) \mathbf{u}]^{n+\frac{1}{2}} + \frac{\nu}{2} \nabla^2 (\mathbf{m}^{n+1} + \mathbf{m}^n), \quad (4.32)$$

$$\mathbf{u}^{n+1} = \mathbf{m}^{n+1} - \nabla \chi^{n+1}.$$

If it is necessary, the pressure may be computed from the second order approximation of the equation (4.31)

$$p^{n+\frac{1}{2}} = \frac{\chi^{n+1} - \chi^n}{\Delta t} - \frac{\nu}{2} \nabla^2 (\chi^{n+1} + \chi^n).$$

The numerical calculation of the projection \mathcal{P} is made usually by solving a Poisson equation. Let \mathbf{w} be a given vectorial field which must be decomposed into $\mathbf{w} = \mathbf{v} + \nabla \Phi$, where \mathbf{v} is of free divergence and satisfies $\mathbf{v}|_{\partial\Omega} = \mathbf{v}_{fr}$, where $\int_{\partial\Omega} \mathbf{v}_{fr} dS = 0$. In order to find \mathbf{v} we have

$$\mathbf{v} = \mathcal{P}(\mathbf{w}) = \mathbf{w} - \nabla \Phi$$

where

$$\nabla^2 \Phi = \nabla \cdot \mathbf{w},$$

$$\mathbf{n} \cdot \nabla \Phi|_{\partial\Omega} = \mathbf{n} \cdot (\mathbf{w}|_{\partial\Omega} - \mathbf{v}_{fr}).$$

We remark that, for the thus defined projection, \mathbf{v} always automatically satisfies the boundary condition in the normal direction to the boundary $\mathbf{n} \cdot \mathbf{v}|_{\partial\Omega} = \mathbf{n} \cdot \mathbf{v}_{fr}$ but in the tangential direction to the boundary we will have $\tau \cdot \mathbf{v}|_{\partial\Omega} = \tau \cdot \mathbf{v}_{fr}$ only if \mathbf{w} is so that

$$\tau \cdot \mathbf{w}|_{\partial\Omega} = \tau \cdot (\mathbf{v}_{fr} + \nabla \Phi|_{\partial\Omega}).$$

This fact must be taken into account at the choice of the boundary conditions for the equations (4.26) and (4.32) where the projection of the solution must verify both the normal and the tangential boundary conditions.

With regard to the above facts, we will describe two projection methods of second order accuracy and without spurious pressure modes.

The first method, which is similar to that proposed by Liu in 1997, may be written as

$$\frac{\mathbf{m}^{n+1} - \mathbf{m}^n}{\Delta t} = - [(\mathbf{u} \cdot \nabla_h) \mathbf{u}]^{n+\frac{1}{2}} + \frac{\nu}{2} \nabla_h^2 (\mathbf{m}^{n+1} + \mathbf{m}^n),$$

$$\mathbf{n} \cdot \mathbf{m}^{n+1}|_{\partial\Omega} = \mathbf{n} \cdot \mathbf{u}_{fr}^{n+1},$$

$$\tau \cdot \mathbf{m}^{n+1}|_{\partial\Omega} = \tau \cdot \mathbf{u}_{fr}^{n+1} + \tau \cdot \nabla_h (2\chi^n - \chi^{n-1})|_{\partial\Omega}.$$

The velocity at the end of the time step is

$$\mathbf{u}^{n+1} = \mathbf{m}^{n+1} - \nabla_h \chi^{n+1}$$

where χ^{n+1} is the solution of the problem

$$\begin{aligned} \nabla_h^2 \chi^{n+1} &= \nabla_h \cdot \mathbf{m}^{n+1}, & in & \quad \Omega, \\ \mathbf{n} \cdot \nabla_h \chi^{n+1} |_{\partial\Omega} &= 0. \end{aligned}$$

If it is necessary, the pressure may be calculated from the relationship

$$p^{n+\frac{1}{2}} = \frac{\chi^{n+1} - \chi^n}{\Delta t} - \frac{\nu}{2} \nabla_h^2 (\chi^{n+1} + \chi^n).$$

In the above relations, the index h means the centered differences discretization, of second order accuracy. The term $[(\mathbf{u} \cdot \nabla_h) \mathbf{u}]^{n+\frac{1}{2}}$ is calculated by centered differences in space and second order extrapolation in time.

The second method, similar to those proposed by Kim and Moin in 1985, is

$$\begin{aligned} \frac{\mathbf{u}^* - \mathbf{u}^n}{\Delta t} &= - [(\mathbf{u} \cdot \nabla_h) \mathbf{u}]^{n+\frac{1}{2}} + \frac{\nu}{2} \nabla^2 (\mathbf{u}^* + \mathbf{u}^n), \\ \mathbf{n} \cdot \mathbf{u}^* |_{\partial\Omega} &= \mathbf{n} \cdot \mathbf{u}_{fr}^{n+1}, \\ \tau \cdot \mathbf{u}^* |_{\partial\Omega} &= \tau \cdot (\mathbf{u}_{fr}^{n+1} + \Delta t \nabla_h \Phi^n) |_{\partial\Omega}. \end{aligned}$$

Then

$$\mathbf{u}^{n+1} = \mathbf{u}^* - \Delta t \nabla_h \Phi^{n+1}$$

where Φ^{n+1} is the solution of the problem

$$\begin{aligned} \Delta t \nabla_h^2 \Phi^{n+1} &= \nabla_h \cdot \mathbf{u}^*, & in & \quad \Omega, \\ \mathbf{n} \cdot \nabla_h \Phi^{n+1} |_{\partial\Omega} &= 0. \end{aligned}$$

If it is necessary, the pressure may be calculated from the relationship

$$\nabla_h p^{n+\frac{1}{2}} = \nabla_h \Phi^{n+1} - \frac{\nu \Delta t}{2} \nabla_h \nabla_h^2 \Phi^{n+1}.$$

In the numerical calculations, we remark that the time extrapolation, where it intercedes, does not perform at the first time step. Here one may use an iterative procedure. For instance, in the case of the first method,

$$\begin{aligned} \frac{\mathbf{m}^{1,k} - \mathbf{m}^0}{\Delta t} &= - [(\mathbf{u} \cdot \nabla_h) \mathbf{u}]^{\frac{1}{2},k} + \frac{\nu}{2} \nabla_h^2 (\mathbf{m}^{1,k} + \mathbf{m}^0), \\ \mathbf{n} \cdot \mathbf{m}^{1,k} |_{\partial\Omega} &= \mathbf{n} \cdot \mathbf{u}_{fr}^1, \\ \tau \cdot \mathbf{m}^{1,k} |_{\partial\Omega} &= \tau \cdot \mathbf{u}_{fr}^1 + \tau \cdot \nabla_h \chi^{1,k-1} |_{\partial\Omega}, \end{aligned}$$

followed by

$$\begin{aligned}\nabla_h^2 \chi^{1,k} &= \nabla_h \cdot \mathbf{m}^{1,k}, \\ \mathbf{n} \cdot \nabla_h \chi^{1,k} |_{\partial\Omega} &= 0,\end{aligned}$$

where the iterations start with $\chi^{1,0} = \chi^0$. The advection term is reset at each iteration taking the average of the derivatives of \mathbf{u}^0 and $\mathbf{u}^{1,k}$.

Another remark is that, in the same relations, we calculate finite differences at neighboring points to the boundary ones, for instance $\nabla_h \cdot \mathbf{u}^*$. The necessary unknown values at the boundary are calculated by quadratic extrapolation from the first three inside values.

More comments on the considered methods may be found in [12].

We retain the idea that, generally, the numerical solution of the Navier–Stokes system is obtained by the following general scheme:

First, we perform a half-discretization in time, by one of the known procedures from the differential equations — backward or forward Euler, Crank-Nicolson or θ -scheme — and we obtain a sequence of steady (generalized) Navier–Stokes systems, with given boundary conditions, in the form:

Being given \mathbf{u}^n and time step size $k = t_{n+1} - t_n$, let us find $\mathbf{u} = \mathbf{u}^{n+1}$ and $p = p^{n+1}$ such that

$$\begin{aligned}\frac{\mathbf{u} - \mathbf{u}^n}{k} + \theta [-\nu \nabla^2 \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}] + \nabla p &= \mathbf{g}^{n+1}, \\ \nabla \cdot \mathbf{u} &= 0\end{aligned}$$

with the right-hand side

$$\mathbf{g}^{n+1} = \theta \mathbf{f}^{n+1} + (1 - \theta) \mathbf{f}^n - (1 - \theta) [-\nu \nabla^2 \mathbf{u}^n + \mathbf{u}^n \cdot \nabla \mathbf{u}^n].$$

This problem may be stated in the compact form

$$\begin{aligned}[I + \theta k N(\mathbf{u})] \mathbf{u} + k \nabla p &= [I - \theta_1 k N(\mathbf{u}^n)] \mathbf{u}^n + \theta_2 k \mathbf{f}^{n+1} + \theta_3 k \mathbf{f}^n, \quad (4.33) \\ \nabla \cdot \mathbf{u} &= 0,\end{aligned}$$

where we have used the notation $N(\mathbf{v})\mathbf{u} = -\nu \nabla^2 \mathbf{u} + \mathbf{v} \cdot \nabla \mathbf{u}$.

Second, we perform the spatial discretization by the finite element method (FEATFLOW, FLUENT), finite difference (SIMPLE, QUICK), finite volume, spectral methods. Some commercial or scientific packages are PHOENICS, FLOTRAN, NSFLEX, FIDAP, FIRE, LISS, FASTEST. By denoting again \mathbf{u} , respectively p , the discrete values of the corresponding functions, the discrete version of the problem (4.33) is:

Being given \mathbf{u}^n and the time step size k , let us find $\mathbf{u} = \mathbf{u}^{n+1}$ and $p = p^{n+1}$ such that

$$\begin{aligned} S\mathbf{u} + \mathbf{k}Bp &= \mathbf{g}, \\ B^T\mathbf{u} &= \mathbf{0}, \end{aligned}$$

where

$$\begin{aligned} S\mathbf{u} &= [M + \theta k N(\mathbf{u})]\mathbf{u}, \\ \mathbf{g} &= [M - \theta_1 k N(\mathbf{u}^n)]\mathbf{u}^n + \theta_2 k \mathbf{f}^{n+1} + \theta_3 k \mathbf{f}^n. \end{aligned}$$

Here M is the mass matrix, B is the gradient matrix and $-B^T$ the transpose of the divergence matrix. The problem becomes a nonlinear algebraic system, which may be usually iteratively solved.

Particular choices lead to particular algorithms, completed by procedures to describe the complex geometries domains, convergence tests, local refinement of the meshes, etc.

4.5.1 Projection-Diffusion Method

We will present now, following [7], [147], [148], a so-called “projection-diffusion algorithm”, elaborated by a French group led by G. Labrosse, to solve the Navier–Stokes unsteady system. This algorithm uses no auxiliary temporal schemes to decouple the velocity field and the pressure.

Let us consider the system

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} - \nu \nabla^2 \mathbf{u} + \nabla p &= f \text{ in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0 \text{ in } \Omega, \\ \mathbf{u} &= \mathbf{u}_0 \text{ on } \partial\Omega, \end{aligned}$$

where f contains, besides some sources, the advective contribution of $(\mathbf{u} \cdot \nabla)\mathbf{u}$. We assume here $\Omega = (-1, 1) \times (-1, 1)$. The projection-diffusion method is suggested by the physical process to instantaneous adaptation of the pressure field on the whole domain, keeping both the solenoidality of u and of the acceleration $\mathbf{u}^* = \frac{\partial \mathbf{u}}{\partial t} - \nu \nabla^2 \mathbf{u}$. The method consists in solving, at each instant, of the problems.

1. The pressure calculation from the system

$$\begin{aligned} \mathbf{u}^* + \nabla p &= f \text{ in } \Omega_x, \Omega_y, \\ \nabla \cdot \mathbf{u}^* &= 0 \text{ in } \bar{\Omega}, \\ \mathbf{u}^* \cdot \mathbf{n} &= \left[\frac{\partial \mathbf{u}}{\partial t} - \nu \nabla^2 \mathbf{u} \right] \cdot \mathbf{n} \text{ on } \partial\Omega \end{aligned}$$

where $\Omega_x = (-1, 1) \times [-1, 1]$, $\Omega_y = [-1, 1] \times (-1, 1)$ are the computational domains for the \mathbf{u} components from the first equation u , respectively v . Here n is the normal unit to $\partial\Omega$.

2. The calculation of the velocities field \mathbf{u} at the next time instant, from the problem

$$\begin{aligned}\frac{\partial \mathbf{u}}{\partial t} - \nu \nabla^2 \mathbf{u} &= \mathbf{u}^* \text{ in } \Omega, \\ \mathbf{u}|_{\partial\Omega} &= \mathbf{u}_0,\end{aligned}$$

implicitly solved in the spectral space. So, at every step we directly solve a Poisson type problem for each dependent variable (the velocities and the pressure).

Chapter 5

FINITE-DIFFERENCE METHODS

1. Boundary Value Problems for Ordinary Differential Equations

Some types of problems from fluid dynamics lead to boundary value problems for differential equations of the form

$$\frac{d^2y}{dx^2} + A(x)\frac{dy}{dx} + B(x)y = C(x), x \in [x_{\min}, x_{\max}], \quad (5.1)$$

$$y(x_{\min}) = y_m, \quad y(x_{\max}) = y_M.$$

The first step to approximately solve these problems by finite differences is to construct the grid

$$x_{\min} = x_0, \dots, x_j = jh, \dots, x_{N+1} = x_{\max}, j = 0, \dots, N + 1$$

with the step size $\Delta x = h = \frac{1}{N+1}$. The values of y evaluated at these points x_j will be denoted by y_j . We will evaluate also the derivatives of y at the same points x_j using the values of y at the neighboring grid points. From the Taylor expansion we have, for a small h ,

$$y_{j+m} \equiv y(x_j + mh) = y_j + mhy'_j + \frac{(mh)^2}{2!}y''_j + \frac{(mh)^3}{3!}y'''_j + \dots .$$

Therefore,

$$y_{j-1} = y_j - hy'_j + \frac{h^2}{2}y''_j - \frac{h^3}{6}y'''_j + \dots ,$$

$$y_{j+1} = y_j + hy'_j + \frac{h^2}{2}y''_j + \frac{h^3}{6}y'''_j + \dots$$

and, consequently,

$$y'_j = \frac{y_{j+1} - y_j}{h} - \frac{h}{2}y''_j + \dots .$$

This represents an approximating formula by *forward finite differences*. Analogously, we obtain

$$y'_j = \frac{y_j - y_{j-1}}{h} + \frac{h}{2}y''_j + \dots$$

which represents the approximating formula by *backward finite differences* and

$$y'_j = \frac{y_{j+1} - y_{j-1}}{2h} - \frac{h^2}{6}y'''_j + \dots$$

that by *centered finite differences*.

The approximation errors are of order $O(h)$ for the first two formulas and $O(h^2)$ for the last formula. But using also other values for m different from $+1$ and -1 (and considering more points in the grid) some formulas of higher accuracy order can be obtained.

The second derivative is similarly approximated,

$$y''_j = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} + \frac{h^2}{12}y''''_j + \dots$$

By replacing these formulas into the differential equation, we get

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} + A_j \frac{y_{j+1} - y_{j-1}}{2h} + B_j y_j = C_j$$

where by A_j, B_j, C_j we understand their values at x_j . Arranging the terms, we have the system

$$\left(1 - \frac{h}{2}A_j\right)y_{j-1} + (h^2B_j - 2)y_j + \left(1 + \frac{h}{2}A_j\right)y_{j+1} = h^2C_j, j = 1, \dots, N$$

which represents the requirement to verify the equation at the *interior grid points*.

The boundary conditions become

$$y_0 = y_m, y_{N+1} = y_M$$

which are the known values that pass to the right-hand side. Finally, the following tridiagonal system for y_1, \dots, y_N is obtained

$$M \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \tilde{C}_1 \\ h^2C_2 \\ \vdots \\ \tilde{C}_N \end{pmatrix},$$

where

$$M = \begin{pmatrix} h^2 B_1 - 2 & 1 + \frac{h}{2} A_1 & & & \\ 1 - \frac{h}{2} A_2 & h^2 B_2 - 2 & & \ddots & \\ & & \ddots & \ddots & 1 + \frac{h}{2} A_{N-1} \\ & & & 1 - \frac{h}{2} A_N & h^2 B_N - 2 \end{pmatrix},$$

$$\tilde{C}_1 = h^2 C_1 - \left(1 - \frac{h}{2} A_1\right) y_m, \quad \tilde{C}_N = h^2 C_N - \left(1 + \frac{h}{2} A_N\right) y_M.$$

By solving this system, using sparse matrices techniques, we get the approximative values of the solution y at the interior grid points. Similarly one could approach the systems of differential equations.

1.1 Supersonic Flow Past a Circular Cylindrical Airfoil

Let us consider the plane, steady, irrotational, inviscid, supersonic fluid flow past a symmetrical circular arcs airfoil, at zero angle of attack, see Figure 5.1. In a Cartesian reference frame Ozy , the equation of the

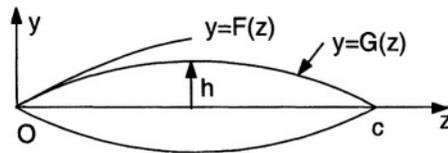


Figure 5.1.

upper side is

$$y = G(z) = -\frac{15}{16} + \sqrt{\left(\frac{17}{16}\right)^2 - \left(z - \frac{1}{2}\right)^2}$$

with

$$G_z = -\frac{z - \frac{1}{2}}{\sqrt{\left(\frac{17}{16}\right)^2 - \left(z - \frac{1}{2}\right)^2}},$$

the geometry of the profile being also characterized by the ratio $\frac{h}{c} = 8$, where h is the “arrow” of the profile and c is its “chord”. We suppose the free stream Mach number to be $M = 2.5$. For details, we refer to the monograph of M. Holt [64], page 69.

To numerically solve this problem we consider the B V L R (Babenko–Voskresenki–Liubimov–Rusanov) method. Let the equations of the given flow (the fluid is supposed to be compressible barotrop)

$$u \frac{\partial u}{\partial z} + v \frac{\partial u}{\partial y} + \frac{1}{\rho} \frac{\partial p}{\partial z} = 0,$$

$$u \frac{\partial v}{\partial z} + v \frac{\partial v}{\partial y} + \frac{1}{\rho} \frac{\partial p}{\partial y} = 0,$$

$$\frac{\partial(\rho u)}{\partial z} + \frac{\partial(\rho v)}{\partial y} = 0,$$

where $\rho = \rho(p)$. This system is equivalent to the matrix equation

$$A' \frac{\partial X}{\partial z} + B' \frac{\partial X}{\partial y} = 0$$

where

$$A' = \begin{pmatrix} u & 0 & \frac{1}{\rho} & 0 \\ 0 & u & 0 & 0 \\ \rho c^2 & 0 & u & 0 \\ \rho & 0 & 0 & u \end{pmatrix}, \quad B' = \begin{pmatrix} v & 0 & 0 & 0 \\ 0 & v & \frac{1}{\rho} & 0 \\ 0 & \rho c^2 & v & 0 \\ 0 & \rho & 0 & v \end{pmatrix}, \quad X = \begin{pmatrix} u \\ v \\ p \\ \rho \end{pmatrix}$$

(here the penultimate equation is a consequence of the last equation, of the Bernoulli integral and of the state equation).

By changing the variables $x = z$ and $\xi = \xi(x, y) = \frac{y - G(z)}{F(z) - G(z)}$, where the function $\xi(x, y)$ was chosen so that $\xi = 0$ on the wall and $\xi = 1$ along the shock wave, the above matrix equation could be rewritten

$$A \frac{\partial X}{\partial x} + B \frac{\partial X}{\partial \xi} = 0$$

where

$$A = A', \quad B = \xi_z A' + \xi_y B',$$

$$\xi = \frac{y - G(z)}{F(z) - G(z)}, \quad \xi_z = -\frac{G_z + \xi(F_z - G_z)}{F - G}, \quad \xi_y = \frac{1}{F - G}.$$

Obviously, to this equation considered for $x > x_0$ (given) and $0 \leq \xi \leq 1$, one attaches both the slip condition on the wall

$$u G_x - v = 0$$

and the boundary conditions (the jump conditions) on the shock, written in the form

$$\rho V_v = \rho_\infty V_{v\infty}, \text{ where } V_v = \frac{u F_x - v}{(1 + F_x^2)^{1/2}},$$

$$p + \rho_\infty V_{v\infty} V_v = p_\infty + \rho_\infty V_{v\infty}^2,$$

$$h + \frac{1}{2} V_v^2 = h_\infty + \frac{1}{2} V_{v\infty}^2, \text{ where } h = \frac{\gamma}{\gamma - 1} \cdot \frac{p}{\rho},$$

$$u + v F_x = u_\infty + v_\infty F_x.$$

Let us now consider a rectangular mesh, with step sizes $\Delta x = \tau$, $\Delta \xi = \frac{1}{M} = h_1$, with mesh nodes of coordinates $x^n = x_0 + n\tau$, $\xi_m = mh_1$ (M, n, m being integers). Let us denote the value of a mesh function f at the node (x^n, ξ_m) by $f(x^n, \xi_m) = f_m^n$.

We will deduce the system of differential equations attached to the above equations. We will use centered differences, with correction terms in the x direction (artificial viscosity), leading to an order2 of accuracy system which may be written in symbolic form

$$a_{m+1/2}^n X_{m+1}^{n+1} + b_{m+1/2}^n X_m^{n+1} = f_{m+1/2}. \tag{5.2}$$

We remark that this system represents $4M$ scalar equations attached to the points of the same “layer” (i.e., having the same index n). To these equations we add the slip-conditions on the wall and the four shock conditions. In the language of finite differences, these equations may be rewritten in the form

$$G_x^{n+1} u_0^{n+1} - v_0^{n+1} = 0,$$

$$(\rho V_v)_M^{n+1} = (\rho V_v)_\infty^{n+1},$$

$$[p + (\rho_\infty V_{v\infty}) V_v]_M^{n+1} = [p + \rho V_v^2]_\infty^{n+1},$$

$$h(p_M^{n+1}, \rho_M^{n+1}) + \frac{1}{2} (V_{vM}^{n+1})^2 = h_\infty + \frac{1}{2} (V_{v\infty}^{n+1})^2,$$

$$u_M^{n+1} + (F_x)^{n+1} v_M^{n+1} = u_\infty + (F_x)^{n+1} v_\infty,$$

where

$$V_{vM}^{n+1} = \frac{u_M^{n+1}(F_x)^{n+1} - v_M^{n+1}}{\left\{1 + [(F_x)^{n+1}]^2\right\}^{1/2}}.$$

In other words, the system contains on each “layer” $4M + 5$ equations with $4M + 5$ unknowns: the values of (u, v, p, ρ) for every $m = 0, \dots, M - 1$ and the values of (u, v, p, ρ) on the shock wave ($m = M$), together with the shock wave equation $F = F(x)$. The location of the shock wave is determined, finally, by the immediate formula

$$F^{n+1} = F^n + \frac{\tau}{2}(F_x^{n+1} + F_x^n).$$

This system may be iteratively solved by the “double sweep” method. Precisely, at the beginning of each iteration cycle, we use the last evaluation of X_m^{n+1} (at the step m) to compute the coefficients $a_{m+1/2}^n, b_{m+1/2}^n, f_{m+1/2}^n$ which depend effectively on X . In the sequel, we consider the system (5.2) as a linear system with the unknown X_{m+1}^{n+1} (from the step $m + 1$) with the known previously computed coefficients.

These iterations will be continued until the difference between the initial and final values for X becomes sufficiently small.

In order to effectively solve the proposed system by the “double sweep” method, we remark that along the airfoil profile (its upper side) the slip condition may be written

$$\mu_0 X_0 = g_0$$

where

$$\mu_0 = \frac{1}{(1 + G_z^2)^{1/2}} (G_z, -1, 0, 0)$$

and $g_0 = 0$.

By forward “sweep” this condition will be transferred, step by step, from the wall ($m = 0$) to the shock ($m = M$). At a certain point (at an intermediate step) we will establish a relationship of the type $\mu_m X_m = g_m$ with the recurrence formulas

$$\mu_{m+1} = \omega_{m+1} \mu_m (b^{-1} a)_{m+1/2},$$

$$g_{m+1} = \omega_{m+1} [\mu_m (b^{-1} f)_{m+1/2} - g_m]$$

where ω_{m+1} is a normalizing factor that makes $\|\mu_{m+1}\| = 1$. So, at every step μ_m and g_m are computed. For $m = M$ one comes on the shock wave where, again, $\mu_M X_M = g_M$. This equation together with

the four boundary conditions, written above on the shock wave, give a system of five equations which allows the determination of the five unknown functions (X_M and F). The effective solution of this system may be found in the paper “Three Dimensional Flow of Ideal Gases Around Smooth Bodies”, NASA TT F-380, of the authors of the B V L R method.

In order to perform now the reverse “sweep”, i.e., the successive determination of $\tilde{X}_{M-1}, \tilde{X}_{M-2}, \dots, \tilde{X}_0$, starting from the shock wave, we must get, by using the difference system (5.2) and the equation $\mu_m X_m = g_m$, a relationship of the form $X_m = c_m X_{m+1} + d_m$, where $\|c_m\| < 1$, the necessary condition for stability, which is feasible. Details on such a scheme may be found in A.N. Liubimov, V.V. Rusanov [86].

The computations will be continued until the difference between the forward values X_m and the reverse values \tilde{X}_m will be smaller than an “a priori” given number, i.e., until the computation stabilizes at a given approximation. The method provides a sufficiently accurate computation of the supersonic flow, the location of the shock wave being better represented than in the Prandtl–Meyer model.

2. Discretization of the Partial Differential Equations

Let $x = x_i, i = 1, \dots, m$ and $y = y_j, j = 1, \dots, n$ be a grid on the computational domain, with the nodes (x_i, y_j) and the step-sizes $\Delta x, \Delta y$ for the two directions, step-sizes of which could be different.

The *finite differences method* replaces the derivatives from the partial differential equation by finite differences, thus resulting an algebraic systems. The basic tool is the Taylor development in the neighborhood of the current point.

For example, if u is the horizontal component of the velocity, then at the point $P_{i,j}$ where $x = x_i$ and $y = y_j$, we have the value $u_{i,j}$ while $u_{i+1,j}$ at $P_{i+1,j}$ has the expression

$$u_{i+1,j} = u_{i,j} + \left(\frac{\partial u}{\partial x}\right)_{i,j} \Delta x + \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \frac{(\Delta x)^2}{2} + \left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} \frac{(\Delta x)^3}{6} + \dots \tag{5.3}$$

The exact value of $u_{i+1,j}$ could be obtained by taking into account all the terms of the series (if the series is convergent). Practically, the series is “truncated” by neglecting the high order terms and considering very small step sizes Δx . So that, we have

$$u_{i+1,j} \approx u_{i,j} + \left(\frac{\partial u}{\partial x}\right)_{i,j} \Delta x + \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \frac{(\Delta x)^2}{2}$$

with a second order accuracy or

$$u_{i+1,j} \approx u_{i,j} + \left(\frac{\partial u}{\partial x} \right)_{i,j} \Delta x$$

with a first order accuracy.

From these relations one could evaluate

$$\left(\frac{\partial u}{\partial x} \right)_{i,j} = \frac{u_{i+1,j} - u_{i,j}}{\Delta x} + O(\Delta x) \quad (5.4)$$

which approximates the first derivative by a *forward finite difference*.

Like the previous one-dimensional case, we have also

$$u_{i-1,j} = u_{i,j} - \left(\frac{\partial u}{\partial x} \right)_{i,j} \Delta x + \left(\frac{\partial^2 u}{\partial x^2} \right)_{i,j} \frac{(\Delta x)^2}{2} - \left(\frac{\partial^3 u}{\partial x^3} \right)_{i,j} \frac{(\Delta x)^3}{6} + \dots \quad (5.5)$$

from which

$$\left(\frac{\partial u}{\partial x} \right)_{i,j} = \frac{u_{i,j} - u_{i-1,j}}{\Delta x} + O(\Delta x), \quad (5.6)$$

that is the approximation of the derivative by a *backward finite difference*.

By subtraction of the formulas (5.3) and (5.5) we get

$$\left(\frac{\partial u}{\partial x} \right)_{i,j} = \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} + O(\Delta x^2)$$

i.e., the approximation by *centered differences*.

If we add the same formulas we obtain

$$\left(\frac{\partial^2 u}{\partial x^2} \right)_{i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} + O(\Delta x^2)$$

which is an approximation of the second order derivative.

Obviously, there exist similar formulas for the derivatives with respect to y :

$$\begin{aligned} \left(\frac{\partial u}{\partial y} \right)_{i,j} &= \frac{u_{i,j+1} - u_{i,j}}{\Delta y} + O(\Delta y), \\ \left(\frac{\partial u}{\partial y} \right)_{i,j} &= \frac{u_{i,j} - u_{i,j-1}}{\Delta y} + O(\Delta y), \\ \left(\frac{\partial u}{\partial y} \right)_{i,j} &= \frac{u_{i,j+1} - u_{i,j-1}}{2\Delta y} + O(\Delta y^2), \end{aligned}$$

$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta y^2} + O(\Delta y^2).$$

We also remark that

$$\begin{aligned} \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} &= \left(\frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x}\right)\right)_{i,j} \approx \frac{\left(\frac{\partial u}{\partial x}\right)_{i+1,j} - \left(\frac{\partial u}{\partial x}\right)_{i,j}}{\Delta x} \\ &= \frac{\frac{u_{i+1,j} - u_{i,j}}{\Delta x} - \frac{u_{i,j} - u_{i-1,j}}{\Delta x}}{\Delta x} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2}, \end{aligned}$$

i.e., forward and backward finite differences are used simultaneously. Thus, we could similarly generate different formulas for other kinds of derivatives. For instance,

$$\begin{aligned} \left(\frac{\partial^2 u}{\partial x \partial y}\right)_{i,j} &= \left(\frac{\partial}{\partial x} \left(\frac{\partial u}{\partial y}\right)\right)_{i,j} \approx \frac{\left(\frac{\partial u}{\partial y}\right)_{i+1,j} - \left(\frac{\partial u}{\partial y}\right)_{i-1,j}}{2\Delta x} \\ &= \frac{\frac{u_{i+1,j+1} - u_{i+1,j-1}}{2\Delta y} - \frac{u_{i-1,j+1} - u_{i-1,j-1}}{2\Delta y}}{2\Delta x} \end{aligned}$$

from which

$$\left(\frac{\partial^2 u}{\partial x \partial y}\right)_{i,j} = \frac{u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1}}{4\Delta x \Delta y} + O(\Delta x^2, \Delta y^2).$$

An important problem is how to approximate the derivatives at the boundary grid points, for example, how to approximate $\frac{\partial u}{\partial y}$ at the boundary node 1 from Figure 5.2.

Using one of the previous formulas, we have

$$\left(\frac{\partial u}{\partial y}\right)_1 = \frac{u_2 - u_1}{\Delta y} + O(\Delta y).$$

A more precise formula could give

$$\left(\frac{\partial u}{\partial y}\right)_1 = \frac{u_2 - u_0}{2\Delta y} + O(\Delta y^2)$$

but u_0 is unknown outside of the computational domain. The boundary condition

$$\left(\frac{\partial u}{\partial y}\right)_1 = 0$$

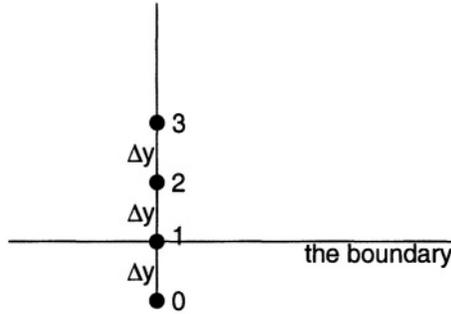


Figure 5.2. The approximation of the derivative at a boundary node

could be imposed by choosing $u_0 = u_2$ but we cannot calculate the derivative with this formula.

Suppose that in the neighborhood of the boundary, u is of the form

$$u(y) = a + by + cy^2. \tag{5.7}$$

Then

$$\begin{aligned} u_1 &= a, \\ u_2 &= a + b\Delta y + c\Delta y^2, \\ u_3 &= a + 2b\Delta y + 4c\Delta y^2, \end{aligned}$$

thus, having u_1, u_2, u_3 one could calculate a, b, c . But, on the other hand,

$$\left(\frac{\partial u}{\partial y}\right)_1 = (b + 2cy)_1 = b,$$

therefore

$$\left(\frac{\partial u}{\partial y}\right)_1 = \frac{-3u_1 + 4u_2 - u_3}{2\Delta y}.$$

Concerning the accuracy, we have

$$u_2 = u_1 + \left(\frac{\partial u}{\partial y}\right)_1 \Delta y + \left(\frac{\partial^2 u}{\partial y^2}\right)_1 \frac{(\Delta y)^2}{2} + \left(\frac{\partial^3 u}{\partial y^3}\right)_1 \frac{(\Delta y)^3}{6} + \dots \tag{5.8}$$

Comparing the formulas (5.7) with (5.8) we find

$$u_2 = a + b\Delta y + c\Delta y^2 + O(\Delta y^3)$$

with errors which affect u_1, u_2, u_3 . Dividing by Δy we obtain

$$\left(\frac{\partial u}{\partial y}\right)_1 = \frac{-3u_1 + 4u_2 - u_3}{2\Delta y} + O(\Delta y^2).$$

Such type of formulas are called *one-sided finite differences*. More details can be found in [124].

3. The Linear Advection Equation

The linear advection equation is

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0, x \in (0, 1), t > 0$$

where c is a constant that physically represents the advection velocity. It is easy to verify that the general solution of this equation is

$$u(x, t) = F(x - ct)$$

where F is an arbitrary, differentiable, single-valued function which represents, in fact, the shape of the solution u at $t = 0$. This profile is translated along the Ox -direction at the velocity c at the next time moments.

This equation is commonly used as an example and a test equation for many numerical methods.

3.1 Discretization of the Linear Advection Equation

The first step in the numerical treatment of this problem is the discretization. In this section we will study different types of discretization by finite differences following [79].

We define a spatial grid of $N + 2$ points, with a constant step size h ,

$$x_j = \left(j - \frac{1}{2}\right) h, \quad j = 0, 1, \dots, N + 1$$

where N of them lie within the computing interval $(0, 1)$. The solution u will be approximated at the points x_1, \dots, x_N while x_0, x_{N+1} will be used for describing the boundary conditions.

So, if u is fixed outside the computing interval, these boundary conditions are discretized by

$$u_0^k = u_0^0, \quad u_{N+1}^k = u_{N+1}^0.$$

In the case of periodic boundary conditions, we have

$$u_0^k = u_N^k, \quad u_{N+1}^k = u_1^k$$

while in the case of homogeneous Neumann conditions we have

$$u_0^k = u_1^k, \quad u_{N+1}^k = u_N^k.$$

Here a temporal grid is defined on $(0, \infty)$, with constant step size Δt ,

$$t_k = k\Delta t, \quad k = 0, 1, \dots$$

and the approximations of the solution u on the grid (x_j, t_k) are denoted by u_j^k ,

$$u_j^k \approx u(x_j, t_k).$$

We shall study different discretizations of the advection equation, obtained by various discretizations of partial derivatives.

3.1.1 Forward-Time and Centered-Space Scheme

We shall use the forward difference for the temporal derivative and the centered difference for the spatial derivative. So, we obtain a discrete form of a first order accurate in time and second order accurate in space, equation

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} + O(\Delta t) + c \left(\frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} + O(\Delta x^2) \right) = 0$$

or, by neglecting the “small” terms,

$$u_j^{k+1} = u_j^k - \frac{c\Delta t}{2\Delta x} (u_{j+1}^k - u_{j-1}^k). \quad (5.9)$$

First, let us analyze the stability of the scheme. We shall use the *von Neumann method*, based on the study of the behavior of a single Fourier mode

$$u(x, 0) = e^{inx}$$

in the approximation process.

The exact solution corresponding to this initial condition is

$$u_{ex}(x, t) = e^{in(x-ct)}.$$

If we are looking for solutions of the approximating equation (5.9) of the form

$$u_j^k = e^{in(x_j - c^* t_k)}$$

advected at the velocity c^* , then

$$u_{j\pm 1}^k = e^{\pm in\Delta x} u_j^k$$

and

$$u_j^{k\pm 1} = e^{\mp inc^* \Delta t} u_j^k.$$

Substituting in the equation (5.9) we find

$$e^{-inc^* \Delta t} = 1 - \frac{c\Delta t}{2\Delta x} (e^{in\Delta x} - e^{-in\Delta x})$$

or

$$e^{-inc^* \Delta t} = 1 - i \frac{c \Delta t}{\Delta x} \sin(n \Delta x).$$

But the above equation implies that the amplification factor $e^{-inc^* \Delta t}$ of the numerical solution passing from the moment t_k to the moment t_{k+1} has a magnitude greater than 1 (we note that c^* may be complex). Consequently, the numerical solution

$$u_j^k \equiv e^{in(x_j - c^* t_k)} = e^{inx_j} \left(e^{-inc^* \Delta t} \right)^k$$

is growing when $t \rightarrow \infty$ and this scheme is *unconditionally unstable*, that means useless.

This example shows that not any discretization gives valid numerical solutions.

3.1.2 Centered-Time and Centered-Space Scheme

The discretization of both derivatives, in space and in time, by centered differences, leads to

$$\frac{u_j^{k+1} - u_j^{k-1}}{2\Delta t} + O(\Delta t^2) + c \left(\frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} + O(\Delta x^2) \right) = 0$$

or

$$u_j^{k+1} = u_j^{k-1} - \frac{c \Delta t}{\Delta x} \left(u_{j+1}^k - u_{j-1}^k \right) \quad (5.10)$$

which is second order accurate.

Let us study the stability of this scheme. As in the previous section, we obtain

$$e^{-inc^* \Delta t} = e^{inc^* \Delta t} - \frac{c \Delta t}{\Delta x} \left(e^{in \Delta x} - e^{-in \Delta x} \right) \quad (5.11)$$

and consequently,

$$\sin(nc^* \Delta t) = \frac{c \Delta t}{\Delta x} \sin(n \Delta x). \quad (5.12)$$

This implies that c^* is real and now, as the left-hand-side has a magnitude less than 1, the above equality is satisfied for every n only if

$$\left| \frac{c \Delta t}{\Delta x} \right| \leq 1.$$

The factor $C = \frac{c \Delta t}{\Delta x}$, which can be considered a “nondimensional velocity”, is called the *Courant number*. The above condition is in fact a restriction on the time step size when the space step size is fixed, and it is called the *CFL (Courant–Friedrichs–Lewy) condition*.

If the CFL-condition is not satisfied then, denoting the amplification factor by

$$\varepsilon = e^{-inc^* \Delta t}$$

and by

$$\tilde{C} = \frac{c\Delta t}{\Delta x} \sin(n\Delta x),$$

the equation (5.11) becomes

$$\varepsilon^2 + 2i\tilde{C}\varepsilon - 1 = 0$$

such that, consequently

$$\varepsilon = -i\tilde{C} \pm \sqrt{1 - \tilde{C}^2}.$$

If $|\tilde{C}| \leq 1$ it is obvious that $|\varepsilon| = 1$ and the scheme is stable. But if $\tilde{C} > 1$, then

$$\varepsilon = i \left(\pm \sqrt{\tilde{C}^2 - 1} - \tilde{C} \right)$$

and the solutions of the difference equation (5.10) are combinations of two elementary solutions: one is oscillating and decaying but the other is oscillating and growing. This growing solution swamps the other and yields instability.

Let us study the accuracy of this scheme, supposing the CFL-condition satisfied. The equation (5.12) gives us the advection velocity of the numerical solution

$$c^* = \frac{1}{n\Delta t} \arcsin \left(\frac{c\Delta t}{\Delta x} \sin(n\Delta x) \right)$$

which may be put in the form

$$C^* = \frac{1}{n\Delta x} \arcsin (C \sin(n\Delta x))$$

where $C^* = \frac{c^*\Delta t}{\Delta x}$ and $C = \frac{c\Delta t}{\Delta x}$.

It should be noted that c^* may coincide with c (for all n) only for very particular spatial and temporal step sizes Δx and Δt . Such a case is $C = 1$, that is $\Delta x = c\Delta t$, which is situated on the stability limit.

If we decrease the step size Δt in order to increase the accuracy and to maintain the stability ($C < 1$), the result is a translation velocity of the numerical solution lower than the exact velocity. This fact is obvious if we plot c^* with respect to c or C^* with respect to C .

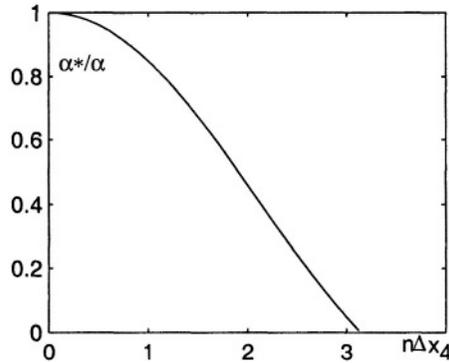


Figure 5.3. Numerical velocity with respect to the wave number

Moreover, the advection velocity of the numerical solution c^* depends on the wave number n . If we represent C^*/C with respect to $n\Delta x$ for a fixed C , for example $C = 1/4$, we obtain Figure 5.3.

We remark that if $n\Delta x = \pi$, then $c^* = 0$ so a wave with the wave number

$$n = \frac{\pi}{\Delta x}$$

never advects. This happens for waves of wavelength

$$\lambda = \frac{2\pi}{n} = 2\Delta x.$$

Longer waves spread numerically faster than shorter waves and the larger the wavelength the better is the numerical velocity.

But the initial profile of the unknown function u may be represented as the sum of a Fourier series and each term of the series is a wave with a specific numerical velocity. Consequently, the initial shape cannot be preserved by numerical advection with this scheme.

Even if the stability is ensured by imposing the CFL-condition, even if we have an acceptable accuracy when the initial profile is a superposition of waves with wavelength greater than the grid step Δx , there are other facts that make the above method difficult to use.

We remark that the equation (5.10) allows the computation of u at the time level $k + 1$ from its values at the time levels k and $k - 1$. But, at the first step, we know only the time level $k = 0$. The necessary values for the next time level may be computed, for example, using the method from the previous section. We suppose that the errors coming from this single step by the unstable method are small relative to other errors of the present method.

Another feature, much more serious, is that our scheme may generate two numerical solutions of the same problem. The value u_j^{k+1} is computed from the values $u_{j\pm 1}^k$ and u_j^{k-1} but ignoring the values u_j^k . If we mark on the grid (x_j, t_k) the points which are under the influence of u_j^k we see that these points are completely independent from those under the influence of $u_{j\pm 1}^k$ or u_j^{k+1} (like the white and black squares on a chess board). So, what we compute are *two uncoupled solutions*, that may be of different behavior and producing spurious numerical oscillations.

Of course, we may diminish this phenomenon by recoupling the partial solutions. For example, such a way which ensures the circulation of the information between the two types of grid points is to substitute the computed values u_j^k by the modified values

$$\bar{u}_j^k = u_j^k + \tau(u_j^{k+1} + \bar{u}_j^{k-1} - 2u_j^k)$$

where $\tau \in (0.01, 0.05)$. There are many types of such filters but their use leads to unnatural algorithms.

3.1.3 **Backward-Time and Centered-Space Scheme**

Let us consider now the following discretization of the linear advection equation

$$\frac{u_j^k - u_j^{k-1}}{\Delta t} + O(\Delta t) + c \left(\frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} + O(\Delta x^2) \right) = 0$$

or

$$u_j^k + \frac{C}{2} (u_{j+1}^k - u_{j-1}^k) = u_j^{k-1}. \tag{5.13}$$

This is an implicit scheme. The solution at the next time level is computed from the present time level by solving a tridiagonal system of equations.

Now, if we study the stability by the von Neumann method, replacing the wave $e^{in(x_j - c^*t_k)}$ in the previous equation, we obtain

$$1 + iC \sin(n\Delta x) = e^{inc^* \Delta t}.$$

The magnitude of the left-hand side is greater than 1, resulting thus in a complex c^* . So, the right-hand side modulus is greater than 1 and the amplification factor

$$e^{-inc^* \Delta t} = \frac{1}{e^{inc^* \Delta t}}$$

has magnitude less than 1. The scheme is then unconditionally stable but it does not preserve the amplitude of the waves.

3.1.4 Crank–Nicolson Scheme

Using the average of forward and backward schemes, we obtain

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} = \frac{c}{2} \left(\frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} + \frac{u_{j+1}^{k+1} - u_{j-1}^{k+1}}{2\Delta x} \right) = 0.$$

If we study the stability as in the previous sections, we have

$$\left(1 + \frac{iC}{2} \sin(n\Delta x) \right) e^{-inc^*\Delta t} = \left(1 - \frac{iC}{2} \sin(n\Delta x) \right).$$

The terms in brackets have the same magnitude, thus resulting in a unitary amplification factor $e^{-inc^*\Delta t}$. The implicit scheme is then unconditionally stable but, as in the previous sections, this scheme does not preserve the shape of the waves: the numerical velocity c^* depends on the wave number n . Particularly, the waves with the wave length $2\Delta x$, for which $n\Delta x = \pi$, yield $c^* = 0$.

3.1.5 Upstream Schemes

We have remarked above that the use of the centered-differences schemes for the spatial derivative does not yield good algorithms. Taking into account the fact that the partial differential equation advects the values of the solution from left to right (downstream), it is natural to use for the spatial discretization a finite difference that uses the known value (from left, upstream) and not the unknown value (from right, downstream) from the spatial grid point x_j .

Then we discretize the spatial derivative by a backward finite difference, using the upstream values of u . For $c > 0$ we obtain

$$u_j^{k+1} - u_j^k + C \left(u_j^k - u_{j-1}^k \right) = 0. \quad (5.14)$$

We firstly remark that this scheme is of first order of accuracy and we need only an upstream boundary condition, so we must specify only the value u_0^k .

The stability study, as in the previous sections, yields

$$e^{-inc^*\Delta t} = 1 - C + Ce^{-in\Delta x}.$$

We see that, generally, $c^* \neq c$. Moreover,

$$\begin{aligned} \left| e^{-inc^*\Delta t} \right|^2 &= (1 - C + Ce^{-in\Delta x}) (1 - C + Ce^{in\Delta x}) \\ &= 1 + 2C(1 - C) [\cos(n\Delta x) - 1]. \end{aligned}$$

It follows that the magnitude of the amplification factor is less than 1 for $0 \leq C \leq 1$ and greater than 1, conversely. So, for $C \in [0, 1)$ the numerical solution is stable but it decreases with time while the exact solution does not.

But, if $C = 1$, from the above relations it follows that $|e^{-inc^*\Delta t}| = 1$ so the numerical solution does not diminish and, more, $c^* = \frac{\Delta x}{\Delta t} = c$. In this particular case the numerical solution is “perfect”.

Even in the case $C \in [0, 1)$, when $c^* \neq c$ and it depends on the wave number n , we have no spurious maxima or minima, due to the numerical diffusion, manifested by a decreasing amplitude of the initial shape. Moreover, each step profile at the initial state is rounded.

Due to the conservation of the maxima and minima of the initial state, even not exactly in position or magnitude, we can say that this scheme is *monotony preserving*.

3.2 Numerical Dispersion and Numerical Diffusion

It is the moment to explain the reason of the numerical difficulties encountered at the above schemes. It should be recalled that we were trying to solve numerically the equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0$$

by discretizing the partial derivatives and neglecting the “small” terms (i.e., of order of some powers of Δt or Δx). But from the generic development in Taylor series

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} = \frac{1}{\Delta x} \left[\Delta x f'(x) + \frac{\Delta x^2}{2!} f''(x) + \frac{\Delta x^3}{3!} f'''(x) + \dots \right]$$

we remark that the neglected terms link to the high order derivatives of u with respect to x and t . This means that the exact equation we try to solve by simple discretizations becomes

$$\begin{aligned} \frac{\partial u}{\partial t} + C_1 \Delta t \frac{\partial^2 u}{\partial t^2} + C_2 \Delta t^2 \frac{\partial^3 u}{\partial t^3} + \dots \\ + c \frac{\partial u}{\partial x} + D_1 \Delta x \frac{\partial^2 u}{\partial x^2} + D_2 \Delta x^2 \frac{\partial^3 u}{\partial x^3} + \dots = 0. \end{aligned}$$

If we use a centered finite difference for the spatial derivative and if we suppose Δt sufficiently small such that the error comes only from Δx

and Δx^2 , we will have

$$\begin{aligned} & \frac{u_j^{k+1} - u_j^k}{\Delta t} + c \left(\frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} \right) \\ &= \frac{\partial u}{\partial t}(x_j, t_k) + c \left(\frac{\partial u}{\partial x}(x_j, t_k) + \frac{\Delta x^2}{6} \frac{\partial^3 u}{\partial x^3}(x_j, t_k) \right). \end{aligned}$$

So, in fact, the numerical solution approximates the solution of a (new) equation of the form

$$\frac{\partial u}{\partial t} + c \left(\frac{\partial u}{\partial x} + \frac{\Delta x^2}{6} \frac{\partial^3 u}{\partial x^3} \right) = 0. \quad (5.15)$$

If we replace here the test wave

$$u(x, t) = e^{in(x - c^*t)}$$

we obtain

$$-inc^*u + c \left(inu - i \frac{\Delta x^2}{6} n^3 u \right) = 0,$$

from which

$$c^* = c \left(1 - \frac{\Delta x^2 n^2}{6} \right).$$

Concluding, the numerical solution, which approximates in fact the solution of the equation (5.15), is advected by a velocity c^* slower than by the exact velocity c and this velocity depends on the wave number n . This is the origin of the *numerical dispersion* that we encountered in the above schemes and it is generated by the presence of odd derivatives into the considered equation.

Let us now take the scheme where the spatial derivative is approximated by a backward finite difference (for $c > 0$), where we also neglected the terms of order Δx^2 ,

$$\begin{aligned} & \frac{u_j^{k+1} - u_j^k}{\Delta t} + c \left(\frac{u_j^k - u_{j-1}^k}{\Delta x} \right) \\ &= \frac{\partial u}{\partial t}(x_j, t_k) + c \left(\frac{\partial u}{\partial x}(x_j, t_k) - \frac{\Delta x}{2} \frac{\partial^2 u}{\partial x^2}(x_j, t_k) \right). \end{aligned}$$

In this case the equation to solve is, in fact,

$$\frac{\partial u}{\partial t} + c \left(\frac{\partial u}{\partial x} - \frac{\Delta x}{2} \frac{\partial^2 u}{\partial x^2} \right) = 0.$$

The term which contains the second spatial derivative of u represents a diffusion and it smoothes the initial state. This term is not a physical one but it is the effect of the discretization of the spatial derivative and thus it is the origin of the *numerical diffusion* encountered in some schemes. This phenomenon is generated by the presence of even derivatives into the equation.

Also, by replacing the test solution

$$u(x, t) = e^{in(x-c^*t)}$$

into the above equation, we find

$$-inc^*u + c \left(inu + \frac{\Delta x}{2} n^2 u \right) = 0$$

and thus

$$c^* = c \left(1 - \frac{in\Delta x}{2} \right).$$

So,

$$u(x, t) = e^{in(x-ct)} e^{-n^2 c \frac{\Delta x}{2} t}$$

and the numerical solution moves with the same velocity as the exact solution while its amplitude decays to zero.

3.3 **Lax, Lax–Wendroff and MacCormack Methods**

There are many other discretization methods. For example, in the equation

$$u_t + cu_x = 0,$$

we can replace the spatial derivative by the centered finite difference and the temporal derivative by the formula

$$\frac{u_j^{k+1} - \frac{1}{2} (u_{j+1}^k + u_{j-1}^k)}{\Delta t} + c \frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} = 0$$

obtaining thus the *Lax method*

$$u_j^{k+1} = \frac{u_{j+1}^k - u_{j-1}^k}{2} - c \frac{\Delta t}{\Delta x} \frac{u_{j+1}^k - u_{j-1}^k}{2}.$$

In this case, by considering a perturbation

$$\varepsilon_m(x, t) = e^{at} e^{ik_m x}$$

the amplification factor becomes

$$e^{a\Delta t} = \cos(k_m \Delta x) - iC \sin(k_m \Delta x)$$

where $C = c \frac{\Delta t}{\Delta x}$. The stability condition $|e^{a\Delta t}| \leq 1$ yields $C \leq 1$ so in this case the CFL condition is valid too.

Consequently, for stability, the Courant number C must be less than 1 while, for accuracy, it is necessary that C be close to 1.

Lax-Wendroff method. Let there be a flow parameter value g_j^k at the point x_j and at the moment t_k . The value at the next time (moment) should be

$$g_j(t_k + \Delta t) = g_j(t_k) + \left(\frac{\partial g}{\partial t}\right)_j^k \Delta t + \left(\frac{\partial^2 g}{\partial t^2}\right)_j^k \frac{\Delta t^2}{2} + \dots$$

From the equations of the phenomenon we can directly compute $\left(\frac{\partial g}{\partial t}\right)_j^k$ and by derivation of the equations with respect to t we can also compute $\left(\frac{\partial^2 g}{\partial t^2}\right)_j^k$. This is a method of second order accuracy.

For example, for the linear advection equation we have

$$u_j^{k+1} = u_j^k + \frac{\partial u_j^k}{\partial t} \Delta t + \frac{\partial^2 u_j^k}{\partial t^2} \frac{\Delta t^2}{2} + \dots$$

We can substitute

$$-c \frac{\partial u_j^k}{\partial x} = \frac{\partial u_j^k}{\partial t}, \quad c^2 \frac{\partial^2 u_j^k}{\partial x^2} = \frac{\partial^2 u_j^k}{\partial t^2}$$

from the equation, which yields the scheme

$$u_j^{k+1} = u_j^k - c\Delta t \frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} + c^2 \frac{\Delta t^2}{2} \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{\Delta x^2}.$$

MacCormack method. This two-step method is easier to apply. The first step is a predictor one

$$u_j^{\overline{k+1}} = u_j^k - c\Delta t \frac{u_{j+1}^k - u_{j-1}^k}{\Delta x},$$

while the second step is the corrector

$$u_j^{k+1} = \frac{1}{2} \left(u_j^k + u_j^{\overline{k+1}} - c\Delta t \frac{u_j^{\overline{k+1}} - u_{j-1}^{\overline{k+1}}}{\Delta x} \right).$$

The accuracy is the same as for the Lax–Wendroff method but we do not need the use of second order derivatives.

Both methods are explicit, so the stability imposes constraints on the time step. If a wave propagates through a fluid with velocity u and the sound velocity is c , the stability constraint is

$$\Delta t = C \frac{\Delta x}{u + c}, C \leq 1.$$

Physically, this means that the time step must not exceed the necessary time to propagate the wave from a grid point to the next one. It should be better that C be closer to 1, but in the case of many grid points this can not always be achieved. We remark that the time step Δt may be variable during the integration process.

3.3.1 Fluid Flow Through a Nozzle

Let us illustrate the MacCormack method for the nonlinear problem of a fluid flow through a nozzle (a tube of variable section, larger at the ends and straightened at the interior), following the paper of J.D. Anderson Jr. [5]. The fluid comes from a reservoir where the flowfield variables are supposed to be constant.

The equations governing the phenomenon are the one-dimensional conservation equations, i.e.,

– the continuity equation

$$A \frac{\partial \rho}{\partial t} + \frac{\partial(\rho u A)}{\partial x} = 0,$$

– the momentum equation

$$\rho \frac{\partial u}{\partial t} + \rho u \frac{\partial u}{\partial x} = -\frac{\partial p}{\partial x},$$

– the energy equation

$$\rho \frac{\partial e}{\partial t} + \rho u \frac{\partial e}{\partial x} = -p \frac{\partial u}{\partial x} - p u \frac{\partial (\ln A)}{\partial x},$$

– the state equation (Clapeyron)

$$p = \rho R T,$$

where $A = A(x)$ is the cross-sectional area, as a function of the distance x along the nozzle.

If ρ_0 is the density in the reservoir, a_0 is the speed of the sound at the temperature T_0 of the reservoir, L is the length of the nozzle,

A^* the minimal area of the section of the nozzle, the above equations are nondimensionalized by

$$u' = \frac{u}{a_0}, T' = \frac{T}{T_0}, x' = \frac{x}{L}, t' = \frac{t}{L/a_0},$$

$$A' = \frac{A}{A^*}, Z = \ln \rho', V = \ln u', \Phi = \ln T'.$$

So, we get the system

$$\frac{\partial Z}{\partial t'} = -u' \left[\frac{\partial (\ln A')}{\partial x'} + \frac{\partial V}{\partial x'} + \frac{\partial Z}{\partial x'} \right],$$

$$\frac{\partial V}{\partial t'} = - \left(\frac{T'}{\gamma_0 u'} \right) \left(\frac{\partial \Phi}{\partial x'} + \frac{\partial Z}{\partial x'} \right) - u' \frac{\partial V}{\partial x'},$$

$$\frac{\partial \Phi}{\partial t'} = -(\gamma_0 - 1)u' \frac{\partial V}{\partial x'} - u' \frac{\partial \Phi}{\partial x'} - (\gamma_0 - 1)u' \frac{\partial (\ln A')}{\partial x'},$$

where the last equation corresponds to the calorically and thermally perfect gas (of constant γ) and consequently $e = \frac{RT}{\gamma - 1}$. In the calculations we take $A'(x') = 1 + 2.2(x' - 1.5)^2$, $x' \in [0, 3]$ and $\gamma_0 = 1.4$.

As boundary conditions at the first computing node from the left (in the reservoir), we have

$$Z_0 = 0, V_0 = 0.1, \Phi_0 = 0.$$

As initial conditions we take linear distributions for ρ' between 1 in the reservoir and 0.1 at the exit of the nozzle, at $x' = 3$, for u' between 0.1 and 1 and for T' between 1 and 0.1. During the evolution, V_0 will be modified by linear extrapolation vs. the first two computing nodes. Similarly, the values at the last node will be calculated by linear extrapolation vs. the last two computing nodes.

This problem is solved by finite differences discretization vs. x' and the time marching will be made by the MacCormack method. Generally, having the values of $Z_{i,j}, V_{i,j}, \Phi_{i,j}$ calculated at the moment t_j ,

– we evaluate from the differential system $\frac{\partial Z}{\partial t'}$ and the other time derivatives at the moment t_j by replacing the spatial derivatives with the first order forward finite differences

$$\frac{g_{i+1,j} - g_{i,j}}{h}$$

where g is the generic notation of the right-hand side of the system;

– we evaluate $Z_{i,j+1}$ and the other quantities at the next time level by

$$Z_{i,j+1} = Z_{i,j} + \Delta t \left(\frac{\partial Z}{\partial t'} \right)_{i,j}, \dots ;$$

– we evaluate $\frac{\partial Z}{\partial t'}$ and the other quantities at the moment t_{j+1} using in the system the above calculated values and discretizing the spatial derivatives by backward finite differences

$$\frac{g_{i,j} - g_{i-1,j}}{h};$$

– we correct the values of the derivatives vs. t' by the average

$$\left(\frac{\partial Z}{\partial t'}\right)_{i,j+1} = \frac{1}{2} \left(\left(\frac{\partial Z}{\partial t'}\right)_{i,j} + \left(\frac{\partial Z}{\partial t'}\right)_{i,j+1} \right), \dots;$$

– we calculate Z, V, Φ at the next time level by

$$Z_{i,j+1} = Z_{i,j} + \Delta t \left(\frac{\partial Z}{\partial t'}\right)_{i,j+1}, \dots$$

and we resume the iterations.

The numerical results are presented in Figure 5.4 where the variation

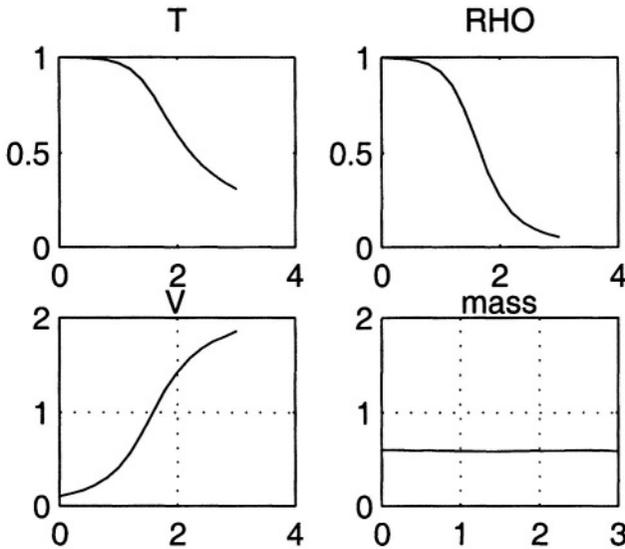


Figure 5.4. The nozzle fluid flow

of the temperature T' , density ρ' , velocity u' and mass transfer $\rho'u'A'$ in the steady state (after 250 iterations) are given vs. x' .

We remark that the spatial step size was chosen constant $\Delta x' = 0.2$ but the time step size was modified during the iterations such that

$$\Delta t' = \min_i \frac{\Delta x'}{u'_i + a'_i}$$

where a'_i is the sound speed at x'_i (corresponding to the temperature T'_i)

$$a' = \frac{\sqrt{\gamma RT}}{\sqrt{\gamma RT_0}} = \sqrt{T'}.$$

This adaptive time step size was imposed by the stability of calculations. The MATLAB code is

```
x=0:0.2:3;A=log(1+2.2*(x-1.5).^2);dx=0.2;er=1;
Z=log(-0.3*x+1);FI=log(-0.3*x+1);V=log(0.3*x+0.1);
for iter = 1:250 dt=min(dx./(exp(V)+sqrt(exp(FI)))));
DZ=-exp(V(2:15)).*(A(3:16)+V(3:16)+Z(3:16)-...
A(2:15)-V(2:15)-Z(2:15))/dx;
DV=-exp(FI(2:15))/1.4./exp(V(2:15)).*...
(FI(3:16)+Z(3:16)-FI(2:15)-Z(2:15))/dx-...
exp(V(2:15)).*(V(3:16)-V(2:15))/dx;
DFI=-exp(V(2:15)).*(0.4*V(3:16)+FI(3:16)+...
0.4*A(3:16)-0.4*V(2:15)-FI(2:15)-0.4*A(2:15))/dx;
ZN(2:15)=Z(2:15)+DZ*dt;VN(2:15)=V(2:15)+...
DV*dt;FIN(2:15)=FI(2:15)+DFI*dt;
ZN(1)=Z(1);VN(1)=max(-15,2*VN(2)-VN(3));FIN(1)=FI(1);
ZN(16)=2*ZN(15)-ZN(14);VN(16)=2*VN(15)-VN(14);
FIN(16)=2*FIN(15)-FIN(14);
DZN=-exp(VN(2:15)).*(A(2:15)+VN(2:15)+ZN(2:15)-...
A(1:14)-VN(1:14)-ZN(1:14))/dx;
DVN=-exp(FIN(2:15))/1.4./exp(VN(2:15)).*...
(FIN(2:15)+ZN(2:15)-FIN(1:14)-ZN(1:14))/dx-...
exp(VN(2:15)).*(VN(2:15)-VN(1:14))/dx;
DFIN=-exp(VN(2:15)).*(0.4*VN(2:15)+FIN(2:15)+...
0.4*A(2:15)-0.4*VN(1:14)-FIN(1:14)-0.4*A(1:14))/dx;
DZ=(DZ+DZN)/2;DV=(DV+DVN)/2;DFI=(DFI+DFIN)/2;
Z(2:15)=Z(2:15)+DZ*dt;V(2:15)=V(2:15)+DV*dt;
FI(2:15)=FI(2:15)+DFI*dt;
Z(1)=Z(1);V(1)=max(-15,2*V(2)-V(3));FI(1)=FI(1);
Z(16)=2*Z(15)-Z(14);V(16)=2*V(15)-V(14);
FI(16)=2*FI(15)-FI(14);
er=max([abs(DZ*dt) abs(DV*dt) abs(DFI*dt)]);
if rem(iter,10)==0
subplot(2,2,1);plot(x,exp(FI));title('T');
subplot(2,2,2);plot(x,exp(Z));title('RH0');
subplot(2,2,3);plot(x,exp(V));title('V');grid;
subplot(2,2,4);plot(x,exp(Z).*exp(V).*exp(A));
title('mass');axis([0 3 0 2]);grid;
disp([iter/100 dt er]); pause;
```

end; end;

which presents the time evolution from 10 by 10 iterations.

4. Diffusion Equation

Let us now consider an equation with second order derivatives with respect to the spatial variable, namely the one-dimensional diffusion equation

$$\frac{\partial u}{\partial t} - c^2 \frac{\partial^2 u}{\partial x^2} = 0, x \in (0, 1), t > 0.$$

This *parabolic* equation is also used as a test problem for different numerical methods. Let us add to it some boundary conditions, like Dirichlet conditions $u(0, t) = u_l, u(1, t) = u_r$.

We consider the same grid $x = x_j, j = 0, \dots, N + 1$, respectively, $t = t_k, k = 0, 1, \dots$ as in the case of the linear advection equation. The spatial derivative will be discretized by the central second order finite difference and the temporal derivative by one of the first order finite differences. From the boundary conditions we have $u_0^k = u_l$ and $u_{N+1}^k = u_r$, for all k .

4.1 Forward-Time Scheme

This is

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} = c^2 \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{\Delta x^2},$$

i.e.,

$$u_j^{k+1} = u_j^k + c^2 \frac{\Delta t}{\Delta x^2} (u_{j+1}^k - 2u_j^k + u_{j-1}^k). \quad (5.16)$$

It is an explicit scheme. Let us study the stability by the von Neumann method.

Consider a Fourier mode

$$u(x, t) = A_n(t)e^{inx},$$

with variable amplitude; by substituting into the diffusion equation we find

$$\frac{dA_n}{dt} = -c^2 n^2 A_n$$

and thus

$$A_n(t) = e^{-c^2 n^2 t}.$$

We remark the decay in time of the amplitudes (as in the attached physical phenomenon) and the decaying dependence on the wave number n .

Let us check the behaviour of the numerical solution, which must be the same in the case of stability. If the numerical solution does not decay it means that the scheme is unstable.

Choosing a test solution of the form

$$u_j^k = e^{-\lambda t_k + i n x_j}$$

and replacing into the equation (5.16) we find

$$u_j^k e^{-\lambda \Delta t} = u_j^k + \frac{c^2 \Delta t}{\Delta x^2} (e^{i n \Delta x} - 2 + e^{-i n \Delta x}) u_j^k$$

or

$$e^{-\lambda \Delta t} = 1 + 2 \frac{c^2 \Delta t}{\Delta x^2} (\cos(n \Delta x) - 1).$$

The numerical solution decreases in time if $e^{-\lambda \Delta t} < 1$. From the above equation we have

$$1 - 4 \frac{c^2 \Delta t}{\Delta x^2} \leq e^{-\lambda \Delta t} \leq 1.$$

There are three cases:

a) if $0 < 4 \frac{c^2 \Delta t}{\Delta x^2} < 1$, the numerical solution is monotonically decreasing,

b) if $1 < 4 \frac{c^2 \Delta t}{\Delta x^2} < 2$, the numerical solution is oscillatory decreasing (in this case λ may be complex and the amplification factor $e^{-\lambda \Delta t}$ is of magnitude less than 1),

c) if $4 \frac{c^2 \Delta t}{\Delta x^2} > 2$, the numerical scheme is unstable.

Concluding, it is necessary for stability that $\Delta t < \frac{\Delta x^2}{4c^2}$. So, this scheme is only *conditionally stable*. The stability requirement is very strong, we need very small time steps and thus this scheme is of less use. The truncation error is of order $O(\Delta t, \Delta x^2)$.

There are many explicit schemes, some of them unconditionally stable, like that of the *DuFort–Frankel* method

$$\frac{u_j^{k+1} - u_j^{k-1}}{2\Delta t} = c^2 \frac{u_{j-1}^k - u_j^{k-1} - u_j^{k+1} + u_{j+1}^k}{\Delta x^2}$$

which implies three time levels. The truncation error is better than in the previous, namely $O(\Delta t^2, \Delta x^2)$.

4.2 Centered-Time Scheme

Let us now consider the following approximation of the diffusion equation

$$\frac{u_j^{k+1} - u_j^{k-1}}{2\Delta t} = c^2 \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{\Delta x^2},$$

namely

$$u_j^{k+1} = u_j^{k-1} + 2c^2 \frac{\Delta t}{\Delta x^2} (u_{j+1}^k - 2u_j^k + u_{j-1}^k)$$

which is second order accurate in space and in time. The scheme is explicit but we need the numerical solution at two previous time levels t_k and t_{k-1} (in order to compute the next time level).

Resuming the stability computations, we find

$$e^{-\lambda \Delta t} = e^{\lambda \Delta t} + 4 \frac{c^2 \Delta t}{\Delta x^2} (\cos(n \Delta x) - 1).$$

If we denote $e^{-\lambda \Delta t} = \varepsilon$ and $-4 \frac{c^2 \Delta t}{\Delta x^2} (\cos(n \Delta x) - 1) = \tilde{C} > 0$, we have

$$\varepsilon^2 + \tilde{C} \varepsilon - 1 = 0$$

and thus

$$\varepsilon = \frac{1}{2} \left(-\tilde{C} \pm \sqrt{\tilde{C}^2 + 4} \right)$$

But the “-” sign always yields a solution with $\varepsilon < -1$ which represents (for complex λ) an oscillatory and growing in magnitude solution. This solution may cover the solution corresponding to the “+” sign. The scheme is therefore *unconditionally unstable*.

We must underline that a better accuracy does not yield a better stability.

4.3 Backward-Time Scheme

Finally, let us consider the scheme

$$\frac{u_j^k - u_j^{k-1}}{\Delta t} = c^2 \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{\Delta x^2}$$

which yields

$$u_j^k - \frac{c^2 \Delta t}{\Delta x^2} (u_{j+1}^k - 2u_j^k + u_{j-1}^k) = u_j^{k-1}.$$

This is an implicit scheme, the numerical solution at the k -time level is computed by solving a tridiagonal system formed with the known solution at the previous time level.

For stability, in this case, we find

$$1 - 2 \frac{c^2 \Delta t}{\Delta x^2} (\cos(k \Delta x) - 1) = e^{\lambda \Delta t}.$$

But the left-hand side is positive and of magnitude greater than 1, thus resulting in an amplification factor $e^{-\lambda \Delta t}$ which is always between 0 and 1, for all spatial or time steps. The scheme is therefore *unconditionally stable*.

4.4 Increasing the Scheme's Accuracy

In 1927 Richardson proposed the following technique to increase the accuracy of schemes with differences. Calculating the exact value u with a method of first order of accuracy, we obtain an approximation u_h of order h of u . Recomputing with a smaller step, like $\frac{h}{2}$, we obtain the approximation $u_{h/2}$.

If the exact solution is smooth, the scheme is stable and the round-off errors in the computer are negligible, then we may write

$$\begin{aligned}u_h &= u + Ah + O(h^2), \\u_{h/2} &= u + A\frac{h}{2} + O(h^2),\end{aligned}$$

where A is supposed constant. Eliminating A we obtain

$$2u_{h/2} - u_h = u + O(h^2)$$

so $2u_{h/2} - u_h$ approximates u by a second order accuracy.

Analogously, using second order schemes, we may obtain schemes of third order of accuracy

$$\begin{aligned}u_h &= u + Ah^2 + O(h^3), \\u_{h/2} &= u + A\frac{h^2}{4} + O(h^3),\end{aligned}$$

and thus

$$\frac{4u_{h/2} - u_h}{3} = u + O(h^3).$$

This procedure cannot be used indefinitely due to the accumulation of round-off errors.

4.5 Numerical Example

Let us use an implicit (backward-time) scheme for the problem of a starting flow in a channel, between two parallel infinite plates, caused by a suddenly imposed pressure gradient $\frac{dp}{dx}$ along the channel. The equation of the flow is

$$\frac{\partial u}{\partial t} = -\frac{1}{\rho} \frac{dp}{dx} + \nu \frac{\partial^2 u}{\partial t^2}.$$

If the distance between the plates is $2L$, the initial and boundary conditions will be $u|_{t=0} = 0$, $-L \leq y \leq L$ and $u|_{y=\pm L} = 0$, $t > 0$.

As the time increases, the solution $u(x, t)$ will approach the steady state distribution $u_s = -\frac{1}{2\mu} \frac{dp}{dx} (L^2 - y^2)$.

By introducing the dimensionless variables

$$T = \frac{t\nu}{L^2}, Y = \frac{y}{L}, U = \frac{u}{-\frac{L^2}{2\mu} \frac{dp}{dx}}$$

the deviation from the steady solution, which is

$$W = \frac{u_s - u}{\frac{L^2 dp}{2\mu dx}} = (1 - Y^2) - U,$$

satisfies the diffusion equation $\frac{\partial W}{\partial t} = \frac{\partial^2 W}{\partial Y^2}$ with the conditions $W|_{Y=\pm 1} = 0$ and $W|_{T=0} = 1 - Y^2$ for $-1 \leq Y \leq 1$.

By applying here this backward-time scheme, we obtain the MATLAB program

```
r=0.4;h=0.05;Y=(-1:h:1)'; wv=1-Y.^2;
e=ones(39,1);wn=zeros(41,1);
T=spdiags([r*e -(1+2*r)*e r*e],[-1 0 1],39,39);
p=plot(Y,1-Y.^2-wv,'EraseMode','none');
axis([-1 1 -0.1 1]);hold on;
for i=1:1000 wn(2:40)=T\(-wv(2:40));
set(p,'color','w');drawnow;
set(p,'Ydata',1-Y.^2-wn,'color','y');drawnow;
wv=wn; end;
plot(Y,1-Y.^2,'r');hold off;
```

which shows the time evolution of the velocity profile, see Figure 5.5.

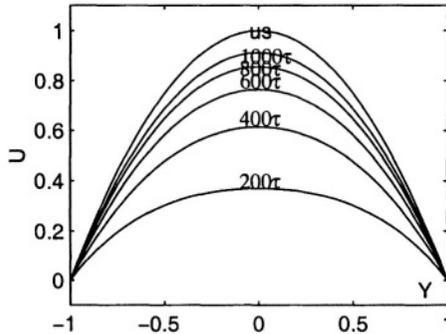


Figure 5.5. The fluid flow caused by a pressure gradient

The truncation error of the above scheme is of order $O(\Delta t, \Delta x^2)$. It could be improved by taking for the spatial derivative the average of the centered finite differences at the time levels $k - 1$ and k

$$\frac{u_j^k - u_j^{k-1}}{\Delta t} = \frac{c^2}{2} \left[\frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{\Delta x^2} + \frac{u_{j-1}^{k-1} - 2u_j^{k-1} + u_{j+1}^{k-1}}{\Delta x^2} \right]$$

which gives, after rearranging the terms,

$$Ru_{j-1}^k - 2(1 + R)u_j^k + Ru_{j+1}^k = -Ru_{j-1}^{k-1} - 2(1 - R)u_j^{k-1} - Ru_{j+1}^{k-1}.$$

This is the *Crank–Nicolson method* which is an implicit method and unconditionally stable too, with a truncation error of order $O(\Delta t^2, \Delta x^2)$.

We remark that the implicit methods have the advantage of stability for large values of the time step size (attention, *not every implicit method is unconditionally stable !*). This means fewer computing steps, resulting in a shorter computing time.

Unfortunately, the programming itself is more difficult, the computing time per each step being longer because a system of equations must be solved if R changes; also, larger truncation errors occur if the time step size Δt is chosen too large.

5. Burgers Equation Without Shock

We prefer the discretization of the conservative form of the equation,

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = 0 \tag{5.17}$$

which is closer to the physical conservation law modeled. In this section we will present some classical discretizations by finite differences schemes.

5.1 Lax Scheme

With the above notation, on a grid defined on $(x, t) \in \mathbb{R} \times \mathbb{R}^+$, we have

$$u_j^{k+1} = \frac{1}{2} \left(u_{j+1}^k + u_{j-1}^k \right) - \frac{\Delta t}{2\Delta x} \left[\left(\frac{u^2}{2} \right)_{j+1}^k - \left(\frac{u^2}{2} \right)_{j-1}^k \right]$$

obtained by discretization of the temporal derivative by a forward finite difference and of the spatial derivative by a centered finite difference together with the substitution of u_j^k by the averaged values at the spatial neighboring points. If we write the scheme in the form

$$u_j^{k+1} = \frac{1}{2} \left(u_{j+1}^k + u_{j-1}^k \right) - \frac{\Delta t}{2\Delta x} \frac{u_{j+1}^k + u_{j-1}^k}{2} \left(u_{j+1}^k - u_{j-1}^k \right),$$

we remark that it may be derived from the discretization of the nonconservative form of Burgers' equation where u_j^k was replaced by the above average. It is an explicit scheme, of first order of accuracy.

In order to study the stability, first a linearization is necessary, either of the original equation or of the nonlinear discretized form. Obviously,

in the nonlinear case, the method is only local, due to the high gradients of the solution in some domains.

Let us consider therefore the linearized scheme

$$u_j^{k+1} = \frac{1}{2} (u_{j+1}^k + u_{j-1}^k) - \frac{u\Delta t}{2\Delta x} (u_{j+1}^k - u_{j-1}^k) \quad (5.18)$$

where u is an averaged local value of the unknown.

Let us take an initial profile

$$u_0(x) = e^{inx}$$

which becomes at the point x_j ,

$$u_j^0 = e^{inj\Delta x}.$$

Substituting into (5.18) we get

$$u_j^1 = \alpha e^{inj\Delta x} = \frac{1}{2} (2 \cos nh) e^{inj\Delta x} - \frac{u\Delta t}{2\Delta x} (2i \sin nh) e^{inj\Delta x}$$

for which the amplification factor is

$$\varepsilon = \cos n\Delta x - \frac{u\Delta t}{\Delta x} i \sin n\Delta x.$$

But

$$|\varepsilon|^2 = 1 - \left(1 - \frac{u^2 \Delta t^2}{\Delta x^2}\right) \sin^2 n\Delta x$$

so that the stability condition is

$$\frac{\Delta t}{\Delta x} < \frac{1}{|u|}. \quad (5.19)$$

We remark now that the stability condition changes with the solution. The time and space step sizes must be automatically adapted while computing. For this, at each time level we should compute $\frac{1}{\sup_j |u_j|}$ and next an acceptable value of Δt is to be considered.

5.2 Leap-Frog Scheme

This is an explicit, of second order accuracy scheme, i.e.,

$$u_j^{k+1} = u_j^{k-1} - \frac{\Delta t}{\Delta x} \left[\left(\frac{u^2}{2}\right)_{j+1}^k - \left(\frac{u^2}{2}\right)_{j-1}^k \right]. \quad (5.20)$$

The stability analysis gives the same condition (5.19). To start we need a single step scheme in order to calculate, in addition, the second time level.

5.3 Lax–Wendroff Scheme

This is an explicit, of second order accuracy scheme, using intermediary grid points $x_{j+\frac{1}{2}} = x_j + \frac{\Delta x}{2}$ and $t_{k+\frac{1}{2}} = t_k + \frac{\Delta t}{2}$. We have two stages: a *predictor*, in which we compute, for all j ,

$$u_{j+\frac{1}{2}}^{k+\frac{1}{2}} = \frac{1}{2} (u_j^k + u_{j+1}^k) - \frac{\Delta t}{2\Delta x} \left[\left(\frac{u^2}{2} \right)_{j+1}^k - \left(\frac{u^2}{2} \right)_j^k \right] \quad (5.21)$$

and which is, in fact, a Lax scheme with steps $\frac{\Delta x}{2}, \frac{\Delta t}{2}$ and a *corrector*, in which we calculate, for all j ,

$$u_j^{k+1} = u_j^k - \frac{\Delta t}{\Delta x} \left[\left(\frac{u^2}{2} \right)_{j+\frac{1}{2}}^{k+\frac{1}{2}} - \left(\frac{u^2}{2} \right)_{j-\frac{1}{2}}^{k+\frac{1}{2}} \right], \quad (5.22)$$

which is a leap-frog scheme with halved steps.

For the stability study, we linearize the equations (5.21) and (5.22), i.e., we get

$$u_{j+\frac{1}{2}}^{k+\frac{1}{2}} = \frac{1}{2} (u_j^k + u_{j+1}^k) - \frac{u\Delta t}{2\Delta x} [u_{j+1}^k - u_j^k]$$

and

$$u_j^{k+1} = u_j^k - \frac{u\Delta t}{\Delta x} \left[u_{j+\frac{1}{2}}^{k+\frac{1}{2}} - u_{j-\frac{1}{2}}^{k+\frac{1}{2}} \right].$$

Eliminating the level $k + \frac{1}{2}$ we obtain

$$u_j^{k+1} = u_j^k - \frac{u\Delta t}{2\Delta x} [u_{j+1}^k - u_j^k] + \frac{u^2\Delta t^2}{2\Delta x^2} [u_{j+1}^k - 2u_j^k + u_{j-1}^k].$$

As above, the amplification factor ε is

$$\varepsilon = 1 - \frac{u\Delta t}{2\Delta x} 2i \sin n\Delta x + \frac{u^2\Delta t^2}{2\Delta x^2} (-2)(1 - \cos n\Delta x)$$

and

$$|\varepsilon|^2 = 1 + \frac{u^2\Delta t^2}{\Delta x^2} (1 - \cos n\Delta x)^2 \left(-1 + \frac{u^2\Delta t^2}{\Delta x^2} \right)$$

which is of magnitude less than 1 for

$$\frac{\Delta t}{\Delta x} < \frac{1}{|u|}$$

6. Hyperbolic Equations

6.1 Discretization of Hyperbolic Equations

Oscillatory flows in fluid dynamics are governed by partial differential hyperbolic equations. For example, the propagation of a one-dimensional sound wave of small amplitude is described by the equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

where t is the time and x is the spatial coordinate in the direction of propagation. The wave's velocity is c (considered constant in the linearized problem) and the flow velocity is u . Similar equations may be written for density, pressure, temperature.

Here u should be found for every time moment $t > 0$ in the spatial domain $x \in [0, 1]$. We also need the initial conditions

$$\begin{aligned} u(x, 0) &= f(x), \\ \frac{\partial u}{\partial t}(x, 0) &= g(x), \end{aligned} \tag{5.23}$$

and boundary conditions, at both ends of the interval. If, for example, one end is closed by a rigid wall, then there we have $u(1, t) = 0$ for all t . If the other end is open into the atmosphere, then the pressure should be constant at that end, i.e., $u_x(0, t) = 0$.

The discretization methods by finite differences for such problems are similar to the parabolic case, cf. [155] for example. We divide the spatial interval by a grid x_i of step size Δx , with the total number of points $N + 2$ and the time axis by a grid t_j of step size Δt , which now is not bounded, see Figure 5.6.

Using second order centered finite differences for partial derivatives, it follows that

$$u_j^{k+1} = 2u_j^k + C^2 \left(u_{j-1}^k - 2u_j^k + u_{j+1}^k \right) - u_j^{k-1}, \quad j = 1, \dots, N, \quad k = 2, 3, \dots \tag{5.24}$$

where C is the nondimensional parameter

$$C = \frac{c\Delta t}{\Delta x},$$

i.e., the *Courant number*. The above formulas calculate the approximate solution at the time level t_{j+1} from the known values at the *two* previous time levels..

Let us study the stability of the above scheme (5.24) using the *von Neumann* method. Suppose that the solution may be developed into a

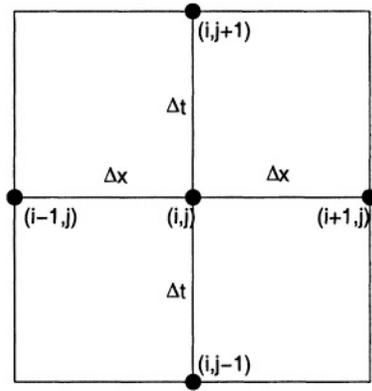


Figure 5.6. The grid for hyperbolic problems

Fourier series with respect to spatial variables. A typical term of this series is

$$u_j^k = A_k e^{ijn\Delta x}$$

where A_k is the amplitude at the moment t_k of the component with the wave number n . Analogously,

$$u_j^{k\pm 1} = A_{k\pm 1} e^{ijn\Delta x}, u_{j\pm 1}^k = A_k e^{i(j\pm 1)n\Delta x}.$$

Replacing into (5.24) we obtain

$$A_{k+1} = 2A_k + C^2 A_k (e^{-in\Delta x} + e^{in\Delta x} - 2) - A_{k-1}$$

or

$$A_{k+1} = \tilde{C} A_k - A_{k-1}$$

where $\tilde{C} = 2 [1 - C^2 (1 - \cos n\Delta x)]$. If we introduce the amplification factor ε for which

$$A_{k+1} = \varepsilon A_k = \varepsilon^2 A_{k-1},$$

the above equation becomes

$$\varepsilon^2 - \tilde{C}\varepsilon + 1 = 0$$

with the roots

$$\varepsilon_{1,2} = \frac{\tilde{C}}{2} \pm \sqrt{\left(\frac{\tilde{C}}{2}\right)^2 - 1}.$$

The stability is ensured only when $|\varepsilon_{1,2}| \leq 1$, that is $\tilde{C}^2 \leq 4$ or

$$C^2 \leq \frac{2}{1 - \cos n\Delta x}.$$

This inequality is always verified if $C^2 \leq 1$ or, finally,

$$\frac{c\Delta t}{\Delta x} \leq 1.$$

Concluding, for the stability we need a relationship between the time and space step sizes. In the particular case when $c\Delta t = \Delta x$ we obtain the scheme

$$u_j^{k+1} = u_{j-1}^k + u_{j+1}^k - u_j^{k-1}$$

which is, in fact, the “*leap-frog*” method.

It can be proved that this method yields the exact solution of the problem. Indeed, the exact solution verifying the initial conditions (5.23) is

$$u(x, t) = \frac{f(x + ct) + f(x - ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} g(s) ds$$

or, in short form,

$$u(x, t) = F(x - ct) + G(x + ct).$$

Here F and G represent waves that propagate without changing the profile, at constant velocity. The lines of slopes $\frac{dx}{dt} = \pm c$ in the plane $x-t$ are the *characteristics* of the wave equation and describe the advance in time of the waves.

So we have

$$u_j^{k+1} = F(x_j - ct_{k+1}) + G(x_j + ct_{k+1}).$$

But

$$x_j = x_1 + (j - 1)\Delta x, t_{k+1} = t_1 + k\Delta t$$

and therefore

$$u_j^{k+1} = F(x_1 + (j - 1)\Delta x - ct_1 - ck\Delta t) + G(x_1 + (j - 1)\Delta x + ct_1 + ck\Delta t).$$

On the other hand,

$$\begin{aligned} u_{j-1}^k + u_{j+1}^k - u_j^{k-1} &= F(x_1 + (j - 2)\Delta x - ct_1 - c(k - 1)\Delta t) \\ &\quad + G(x_1 + (j - 2)\Delta x + ct_1 + c(k - 1)\Delta t) \\ &\quad + F(x_1 + j\Delta x - ct_1 - c(k - 1)\Delta t) \\ &\quad + G(x_1 + j\Delta x + ct_1 + c(k - 1)\Delta t) \\ &\quad - F(x_1 + (j - 1)\Delta x - ct_1 - c(k - 2)\Delta t) \\ &\quad - G(x_1 + (j - 1)\Delta x + ct_1 + c(k - 2)\Delta t). \end{aligned}$$

Taking into account $\Delta x = c\Delta t$, the two equations coincide.

The above computations have a strong physical interpretation. The exact solution formula shows that the value of the solution at a point P of the plane $x - t$ depends on the values on the grid points between the diagonals PQ and PR of slopes $\frac{dx}{dt} = \pm \frac{\Delta x}{\Delta t}$ through P (see Figure 5.7).

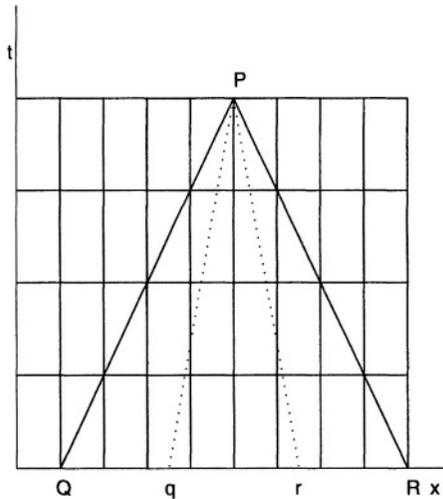


Figure 5.7. Physical interpretation of the stability criterion

The region PQR is the computational domain of the values of the solution in P . If Pq , respectively Pr , are the characteristics through P of slopes $\frac{dx}{dt} = \pm c$ for the exact solution, then Pqr is the physical domain of dependence for the exact solution in P . If $c < \frac{\Delta x}{\Delta t}$, as in the figure, then the computational domain contains the physical domain of dependence and the computations are stable. Computing errors appear by using values from the computational domain and not from the dependence domain.

But if $C = \frac{c\Delta t}{\Delta x} > 1$, then the computational domain PQR is included in the physical dependence domain Pqr and only a part of needed information for the value of the solution in P is available and this yields the instability. The limit case $C = 1$ or $c\Delta t = \Delta x$ yields the equality of the two domains and the algorithm computes the exact solution.

In the above formulas we need the numerical solution at two time levels in order to calculate it at a next time level. We have directly from the initial conditions

$$u_j^1 = f.$$

The equation at $k = 1$ is

$$u_j^2 = u_{j-1}^1 + u_{j+1}^1 - u_j^0.$$

But from the second initial condition, if we discretize the derivative by a centered finite difference, we get

$$\frac{u_j^2 - u_j^0}{2\Delta t} = g_j$$

from which, by replacing u_j^0 we have

$$u_j^2 = u_{j-1}^1 + u_{j+1}^1 - u_j^2 + 2\Delta t g_j$$

or

$$u_j^2 = \frac{1}{2}(f_{j-1} + f_{j+1}) + \Delta t g_j.$$

So, we have the starting formulas.

As an immediate application, let us study the sound waves in a tube [22]. Let us consider a tube of $1m$ in length, of uniform cross section, divided into two chambers by a diaphragm at the middle section and closed at the right end. Suppose that the air density in the tube is $\rho_0 + \rho'(x, t)$ where ρ_0 is the atmospheric density and

$$\rho'(x, 0) = f(x) = \begin{cases} 0, & x < \frac{1}{2} \\ 1, & x \geq \frac{1}{2} \end{cases}$$

with $\frac{d\rho'}{dt}(x, 0) = 0$. The boundary conditions will be $\rho'(0, t) = 0$ at the open end and $\rho'_x(1, t) = 0$ at the closed end.

Suppose that at $t = 0$ the diaphragm is suddenly removed. It may be proved, from Euler's equation, that

$$\rho \frac{DV}{Dt} = -\nabla p$$

and from the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho V) = 0,$$

by linearization, that the density fluctuation ρ' satisfies the wave equation

$$\frac{\partial^2 \rho'}{\partial t^2} = a^2 \frac{\partial^2 \rho'}{\partial x^2}$$

where a is the sound speed, considered to be $340m/s$.

By denoting, in the sequel, $\rho' \equiv u$ and using the above discretization, the boundary conditions become

$$\begin{aligned} u_1^k &= 0, \\ u_{N+1}^k &= u_{N-1}^k, \end{aligned}$$

which leads to the starting relations

$$\begin{aligned} u_j^1 &= f_j, \\ u_j^2 &= \frac{1}{2}(f_{j-1} + f_{j+1}), j = 2, \dots, N-1, \\ u_N^2 &= f_{N-1}, \end{aligned}$$

and to the iterations

$$\begin{aligned} u_j^{k+1} &= u_{j-1}^k + u_{j+1}^k - u_j^{k-1}, j = 2, \dots, N-1, \\ u_N^{k+1} &= 2u_{N-1}^k - u_N^{k-1} \end{aligned}$$

where the time step size is $\Delta t = \frac{\Delta x}{a}$. The following MATLAB program performs an animation of this phenomenon for 1000 time steps, describing the evolution of u as a function of x and time

```
a=340;m=201;h=1/(m-1);jmax=1000;
tau=h/a;u=zeros(m,1);u1=zeros(m,1);u2=zeros(m,1);
x=(0:h:1)';f=zeros(m,1);for i=(m-1)/2:m f(i)=1;end;
u1=f;
u2(1)=0;u2(2:m-1)=(f(1:m-2)+f(3:m))/2;
u2(m)=f(m-1);
p=plot(x,u1,'EraseMode','none');xlabel('x');ylabel('u');
axis([-0.1 1.1 -1.1 1.1]);pause(1);
set(p,'color','w');drawnow;
set(p,'Ydata',u2,'color','y');drawnow;
for j=3:jmax u(1)=0;u(2:m-1)=u2(1:m-2)+u2(3:m)-u1(2:m-1);
u(m)=2*u2(m-1)-u1(m);
set(p,'color','w');drawnow;
set(p,'Ydata',u,'color','y');drawnow;
u1=u2;u2=u;pause(0.01);
end;
```

The program shows the generation of two waves, a compression wave propagating toward the left and an expansion wave propagating towards the right and which specifically reflects at the ends of the tube.

6.2 Discretization in the Presence of a Shock

In a neighborhood of a shock the variation of u at the considered grid points does not tend to zero, and that induces numerical difficulties. These problems may be surpassed either by using some “shock-fitting” schemes, that treat the position of the shock as an unknown

and discretize the equation separately on each side of the shock, or by introducing an “artificial viscosity”.

The first methods are difficult to use. It is difficult to follow a shock that may appear or disappear, it is difficult to express numerically the entropy condition and it is difficult to keep the stability of the schemes in the case of using variable step sizes.

The methods based on an artificial viscosity are more often used. Although they may represent the shock more extended than it really is, its position and intensity are correctly represented. Moreover, we need not impose the entropy condition because at the limit, when $v \searrow 0$ we obtain just the entropy solution.

Practically, we add to the equation a term of the form $v \frac{\partial^2 u}{\partial x^2}$ with v positive and small, or we discretize the equation by a dissipative scheme, which automatically introduces a numerical diffusion.

Let us first take the linear advection equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0$$

where $c > 0$ is constant. The Lax scheme is

$$u_j^{k+1} = \frac{1}{2} (u_{j+1}^k + u_{j-1}^k) - \frac{c\Delta t}{2\Delta x} (u_{j+1}^k - u_{j-1}^k).$$

Let us suppose that u is sufficiently smooth and it can be developed in a neighborhood of the point (x_j, t_k) . Applying these developments in the above formula we obtain the equation

$$\begin{aligned} & u + \Delta t \frac{\partial u}{\partial t} + \frac{\Delta t^2}{2} \frac{\partial^2 u}{\partial t^2} + O(\Delta t^3) \\ &= u + \frac{\Delta x^2}{2} \frac{\partial^2 u}{\partial x^2} + O(\Delta x^4) - c\Delta t \left(\frac{\partial u}{\partial x} + O(\Delta x^2) \right). \end{aligned}$$

If we suppose $\Delta t = O(\Delta x)$, the equation may be written

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = \frac{1}{\Delta t} \left(\frac{\Delta x^2}{2} \frac{\partial^2 u}{\partial x^2} - \frac{\Delta t^2}{2} \frac{\partial^2 u}{\partial t^2} \right) + O(\Delta x^2).$$

We may say either that the Lax scheme discretizes the advection equation by a first order accuracy or that this scheme discretizes the equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = \frac{1}{\Delta t} \left(\frac{\Delta x^2}{2} \frac{\partial^2 u}{\partial x^2} - \frac{\Delta t^2}{2} \frac{\partial^2 u}{\partial t^2} \right) \quad (5.25)$$

by a second order accuracy.

From

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = O(\Delta x)$$

we obtain, if u is sufficiently smooth,

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= -c \frac{\partial}{\partial t} \left(\frac{\partial u}{\partial x} \right) + O(\Delta x) = -c \frac{\partial}{\partial x} \left(\frac{\partial u}{\partial t} \right) + O(\Delta x) \\ &= -c \frac{\partial}{\partial x} \left(-c \frac{\partial u}{\partial x} + O(\Delta x) \right) + O(\Delta x) = c^2 \frac{\partial^2 u}{\partial x^2} + O(\Delta x) \end{aligned}$$

and then the equation (5.25) becomes

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = \frac{\Delta x^2}{2\Delta t} \left(1 - \frac{c^2 \Delta t^2}{\Delta x^2} \right) \frac{\partial^2 u}{\partial x^2} + O(\Delta x^2).$$

If the stability condition is verified,

$$\frac{|c| \Delta t}{\Delta x} < 1,$$

then the numerical method introduces, in fact, in the right-hand side a dissipative term of intensity

$$\frac{\Delta x^2}{2\Delta t} \left(1 - \frac{c^2 \Delta t^2}{\Delta x^2} \right) = O(\Delta x) \searrow 0$$

when $\Delta x \rightarrow 0$.

The calculations are performed similarly for the Burgers equation. In fact, the second order approximating equation is

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = \frac{1}{\Delta t} \left(\frac{\Delta x^2}{2} \frac{\partial^2 u}{\partial x^2} - \frac{\Delta t^2}{2} \frac{\partial^2 u}{\partial t^2} \right) + O(\Delta x^2).$$

But we have

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= \frac{\partial}{\partial t} \left(\frac{\partial u}{\partial t} \right) = \frac{\partial}{\partial t} \left[-\frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) \right] = -\frac{\partial}{\partial x} \left[\frac{\partial}{\partial t} \left(\frac{u^2}{2} \right) \right] \\ &= -\frac{\partial}{\partial x} \left[u \frac{\partial u}{\partial t} \right] = \frac{\partial}{\partial x} \left[u \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) \right] = \frac{\partial}{\partial x} \left[u^2 \frac{\partial u}{\partial x} \right] = u^2 \frac{\partial^2 u}{\partial x^2} + 2u \left(\frac{\partial u}{\partial x} \right)^2 \end{aligned}$$

so our equation may be finally written

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = \frac{\Delta x^2}{2\Delta t} \left(1 - \frac{u^2 \Delta t^2}{\Delta x^2} \right) \frac{\partial^2 u}{\partial x^2} - u \Delta t \left(\frac{\partial u}{\partial x} \right)^2 + O(\Delta x^2).$$

The perturbation $-u\Delta t \left(\frac{\partial u}{\partial x}\right)^2$ is regular, so the qualitative behavior of the equation is the same as that of the Burgers equation with dissipative term. Of course, we should impose the condition

$$\frac{\Delta t}{\Delta x} < \frac{1}{|u|}$$

which ensures the correct sense of dissipation (with positive intensity).

This technique with artificial viscosity is often used for the stabilization of numerical schemes. Let us consider now a numerical example for the Burgers equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = 0$$

with the initial condition (see [42])

$$u(x, 0) \equiv u_0(x) = \begin{cases} 1, & x \leq 0, x \geq 2 \\ 0, & 0 < x < 1 \\ x - 1, & 1 \leq x < 2 \end{cases} \quad (5.26)$$

Starting from the parametric equations of the exact solution

$$\begin{aligned} x &= u_0(\xi)t + \xi, \\ u &= u_0(\xi), \end{aligned}$$

and from the shock condition

$$\frac{dx}{dt} = \frac{1}{2}(u_1 + u_2),$$

we can find the exact solution for the above problem, i.e.,

$$u(x, t) = \begin{cases} 1, & x < x_s(t) \\ 0, & x_s(t) < x \leq 1, 0 < t < 2 \\ \frac{x-1}{1+t}, & \max(1, x_s(t)) < x \leq t+2 \\ 1, & x > t+2 \end{cases}$$

where $x = x_s(t)$ is the shock equation,

$$x_s(t) = \begin{cases} \frac{t}{2}, & 0 < t \leq 2 \\ t + 2 - \sqrt{3}\sqrt{t+1}, & t > 2 \end{cases}$$

The equations of the characteristic lines, along which the initial values u should be transported, are, corresponding to the four lines from the

definition of u ,

$$\begin{aligned} t &= x - \xi, & \xi < 0, \\ x &= \xi, & 0 < \xi < 1, \\ t &= \frac{x - \xi}{\xi - 1}, & 1 < \xi < 2, \\ t &= x - \xi, & \xi > 2, \end{aligned}$$

see Figure 5.8.

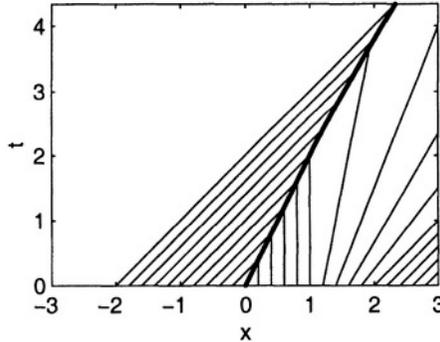


Figure 5.8. The characteristic lines of the equation

We will apply two discretization schemes and we will compare the numerical solutions with the exact one at $t = 3$, $x \in (0.5, 5.5)$, namely with

$$u(x, 3) = \begin{cases} 1, & x < 5 - 2\sqrt{3} \\ \frac{x - 1}{4}, & 5 - 2\sqrt{3} < x < 5 \\ 1, & x \geq 5 \end{cases}$$

First, we will use the Lax scheme,

$$u_j^{k+1} = \frac{1}{2} (u_{j+1}^k + u_{j-1}^k) - \frac{\Delta t}{2\Delta x} \left[\left(\frac{u^2}{2} \right)_{j+1}^k - \left(\frac{u^2}{2} \right)_{j-1}^k \right]$$

and then the predictor-corrector scheme S_β^α studied by Peyret and Lerat,

$$\begin{aligned} u_j &= (1 - \beta)u_j^k + \beta u_{j+1}^k - \alpha \frac{\Delta t}{\Delta x} \left[\left(\frac{u^2}{2} \right)_{j+1}^k - \left(\frac{u^2}{2} \right)_j^k \right], \\ u_j^{k+1} &= u_j^k - \frac{\Delta t}{2\alpha\Delta x} \left[(\alpha - \beta) \left(\frac{u^2}{2} \right)_{j+1}^k + (2\beta - 1) \left(\frac{u^2}{2} \right)_j^k \right. \\ &\quad \left. + (1 - \alpha - \beta) \left(\frac{u^2}{2} \right)_{j-1}^k + \left(\frac{u^2}{2} \right)_j^k - \left(\frac{u^2}{2} \right)_{j-1}^k \right], \end{aligned}$$

where the proposed optimal values $\alpha = 1 + \frac{\sqrt{5}}{2}, \beta = \frac{1}{2}$ were chosen. For both schemes we use the same stability condition

$$\frac{\Delta t}{\Delta x} < \frac{1}{\sup_j |u_j|}$$

which at each time level evaluates the new time step size Δt , the spatial step size Δx being fixed ($\Delta x = 0.01$) and therefore $\Delta t = \Delta x$.

Figures 5.9 and 5.10 show the numerical solutions together with the errors versus the exact solution.

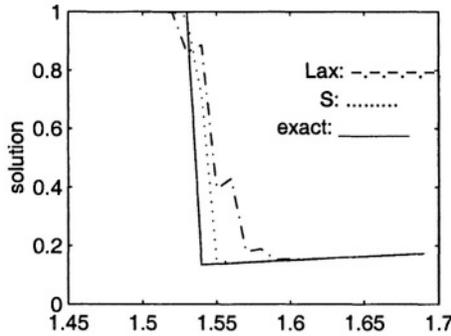


Figure 5.9. The numerical solutions

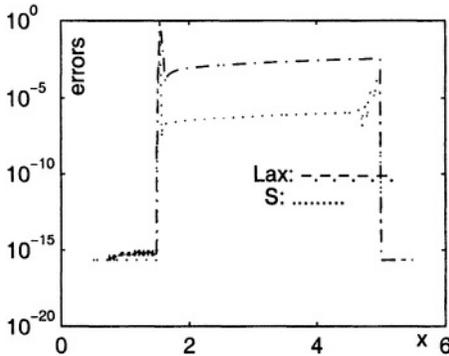


Figure 5.10. The errors

We could remark that the second accuracy scheme S_β^α gives better results than the Lax scheme, which is of the first order of accuracy. The MATLAB program, called `lax.m` is

```

clear;alpha=1+sqrt(5)/2;
for j=1:1101 x(j)=-2.5+(j-1)*0.01;
if x(j)<=0 u(j)=1; elseif x(j)<=1 u(j)=0;
elseif x(j)<=2 u(j)=x(j)-1; else u(j)=1;
end; end;v=u;
for k=1:300 u=(u(3:length(u))+u(1:length(u)-2))/2.*...
(1-(u(3:length(u))-u(1:length(u)-2))/2);
x=x(2:length(x)-1);
g=(v.*v)/2;
vi=(v(1:length(v)-1)+v(2:length(v)))/2-alpha*...
(g(2:length(v))-g(1:length(v)-1)); gi=(vi.*vi)/2;
v=v(2:length(v)-1)-1/2/alpha*((alpha-0.5)*...
(g(3:length(v))-g(1:length(v)-2))+...
gi(2:length(vi))-gi(1:length(vi)-1));
end;
for j=1:length(x)
if x(j)<5-2*sqrt(3) uex(j)=1;
elseif x(j)<5 uex(j)=(x(j)-1)/4;
else uex(j)=1; end; end;
plot(x,u,'-.',x,v,':',x,uex,'-');
xlabel('x');ylabel('solution');pause;
semilogy(x,eps+abs(u-uex),'-.',x,eps+abs(v-uex),':');
xlabel('x');ylabel('errors');
text(3,1.e-10,'Lax: -.-.-.-.-');
text(3.2,1.e-11,'S: .....');pause;
plot(x(100:120),u(100:120),'-.',x(100:120),...
v(100:120),':',x(100:120),uex(100:120),'-');
xlabel('x');ylabel('solution');
text(1.61,0.8,'Lax: -.-.-.-.-');
text(1.62,0.7,'S: .....');
text(1.603,0.6,'exact:-----');

```

6.3 Method of Characteristics

The numerical solution for the sound waves in a tube shows that a sound wave propagates at a constant velocity without changing its shape. This fact is a result of linearization of the governing equations, in the case of small perturbations about the equilibrium state.

These governing equations are, as we know, the Euler and the continuity equations

$$\begin{aligned}\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} &= 0, \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{a^2}{\rho} \frac{\partial \rho}{\partial x} &= 0,\end{aligned}\tag{5.27}$$

where the sound speed a is also a function of x and t . Assume that the fluid flow velocity u is less than the sonic speed. By eliminating ρ , after some transformations, we obtain the equations

$$\begin{aligned}\left[\frac{\partial}{\partial t} + (u + a) \frac{\partial}{\partial x} \right] \left(u + \frac{2a}{\gamma - 1} \right) &= 0, \\ \left[\frac{\partial}{\partial t} + (u - a) \frac{\partial}{\partial x} \right] \left(u - \frac{2a}{\gamma - 1} \right) &= 0,\end{aligned}$$

where

$$\frac{p}{p_0} = \left(\frac{\rho}{\rho_0} \right)^\gamma$$

and the subscript “0” indicates the undisturbed conditions.

The above equations show that

$$P = u + \frac{2a}{\gamma - 1}$$

is constant along a curve in the xt plane. From

$$dP = \frac{\partial P}{\partial t} dt + \frac{\partial P}{\partial x} dx = \left(\frac{\partial}{\partial t} + \frac{dx}{dt} \frac{\partial}{\partial x} \right) P = 0,$$

comparing with the above equations, we obtain

$$\frac{dx}{dt} = u + a$$

which is the expression for the slope of that curve.

Similarly,

$$Q = u - \frac{2a}{\gamma - 1}$$

is constant along a curve of slope

$$\frac{dx}{dt} = u - a$$

These curves are the so-called *characteristics* of the equations. As u and a depend on x and t , the characteristics are generally curves in the plane xt .

Since P , respectively Q , are constant along the characteristics, a so-called *method of characteristics* may be developed [22]. Suppose that the initial data are given at $t = 0$ and we must calculate those at a point C , at some $t > 0$. The two characteristics through C , of slopes $u + a$ respectively $u - a$, intersect the Ox -axis at A , respectively B . But $P_C = P_A$ and $Q_C = Q_B$ or

$$u_C + \frac{2a_C}{\gamma - 1} = u_A + \frac{2a_A}{\gamma - 1},$$

$$u_C - \frac{2a_C}{\gamma - 1} = u_B - \frac{2a_B}{\gamma - 1}.$$

Thus

$$\begin{aligned} u_C &= \frac{u_A + u_B}{2} + \frac{a_A - a_B}{\gamma - 1}, \\ a_C &= \left(\frac{\gamma - 1}{4}\right)(u_A - u_B) + \frac{a_A + a_B}{2}. \end{aligned} \tag{5.28}$$

If the distance between A and B is small, the characteristics can be approximated by two straight lines of slopes $u_A + a_A$, respectively $u_B - a_B$, and then, the values at C , which is the intersection of these lines, are approximately given by the above formulas.

Therefore, having the grid in the xt plane, we will draw the (linearized) characteristics through the new point C and the values at the previous time level at A and B will be calculated by interpolation from the known values on the grid at this time level, see Figure 5.11.

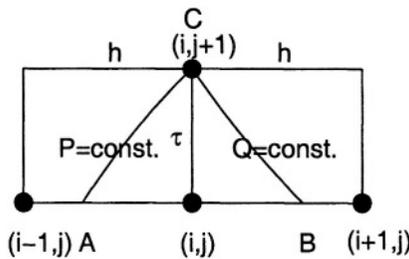


Figure 5.11. The characteristics method

Finally we have

$$\begin{aligned}u_A &= u_{i,j} + \frac{\tau}{h} (u_{i,j} + a_{i,j}) (u_{i-1,j} - u_{i,j}), \\a_A &= a_{i,j} + \frac{\tau}{h} (u_{i,j} + a_{i,j}) (a_{i-1,j} - a_{i,j}), \\u_B &= u_{i,j} - \frac{\tau}{h} (u_{i,j} - a_{i,j}) (u_{i+1,j} - u_{i,j}), \\a_B &= a_{i,j} - \frac{\tau}{h} (u_{i,j} - a_{i,j}) (a_{i+1,j} - a_{i,j}),\end{aligned}$$

and then the new values $u_C \equiv u_{i,j+1}$, respectively $a_C = a_{i,j+1}$, will be calculated from (5.28). Here h is the spatial step size Δx and τ is the time step size Δt .

The stability condition requires that the numerical domain of dependence at any grid point is not less than the physical domain of dependence determined by the characteristics, i.e., both

$$\frac{\tau}{h} |u + a| \leq 1, \quad \frac{\tau}{h} |u - a| \leq 1$$

must be satisfied in the whole computational domain.

For example, let us consider, as in the previous example, a tube with the left end closed and the right end open. Let $a_0 = 340\text{m/s}$ be the sound speed in the undisturbed state and at $t = 0$ we will consider a small perturbation of the shape described by a piecewise linear function determined by $u_{i,1}$ (as in the program); the initial condition for a is

$$a = a_0 \pm \frac{\gamma - 1}{2} u$$

where the sign is taken in correspondence with the sign of x . The boundary conditions become

$$\begin{aligned}u_{i,j+1} &= 0, \\a_{i,j+1} &= a_B - \frac{\gamma-1}{2} u_B\end{aligned}$$

for the left end, open in the atmosphere and where only the characteristic $Q = \text{const.}$ is used while

$$\begin{aligned}u_{m,j+1} &= u_A + \frac{2}{\gamma - 1} (a_A - a_0), \\a_{m,j+1} &= a_0\end{aligned}$$

for the closed right end where only the characteristic $P = \text{const.}$ is used.

The program uses $\gamma = 1.4$ for the air at sea level, $h = 0.02\text{m}$ and the starting time step size $\tau = 0.5h/a_0$ which can be modified by testing the numerical stability. The MATLAB program is

```
a=zeros(101,1);u=zeros(101,1);
a0=340;h=0.02;gamma=1.4;
m=101;jmax=2000;tau=0.5*h/a0;ratio=tau/h;
ampl=a0/2;coef=(gamma-1)/2;uv=zeros(101,1);
```

```

uv(1:13)=ampl*((1:13)'-1)/12;
uv(14:39)=ampl*(26-(14:39)')/13;
uv(40:51)=ampl*((40:51)'-51)/12;
av=a0+coef*uv;
p=plot(0:0.02:2,uv,'EraseMode','none');
xlabel('x');ylabel('y');pause(1);
for j=1:jmax u(1)=0;a(m)=a0;
upa=ratio*(uv(1)+av(1));uma=ratio*(uv(1)-av(1));
if (abs(upa)>1 abs(uma)>1) break;end;
ub=uv(1)-uma*(uv(2)-uv(1));
ab=av(1)-uma*(av(2)-av(1)); a(1)=ab-coef*ub;
for i=2:m-1
upa=ratio*(uv(i)+av(i));uma=ratio*(uv(i)-av(i));
if (abs(upa)>1 abs(uma)>1) break;end;
ua=uv(i)+upa*(uv(i-1)-uv(i));
aa=av(i)+upa*(av(i-1)-av(i));
ub=uv(i)-uma*(uv(i+1)-uv(i));
ab=av(i)-uma*(av(i+1)-av(i));
u(i)=0.5*((ua+ub)+(aa-ab)/coef);
a(i)=0.5*(coef*(ua-ub)+(aa+ab)); end;
upa=ratio*(uv(m)+av(m));uma=ratio*(uv(m)-av(m));
if (abs(upa)>1 abs(uma)>1) break;end;
ua=uv(m)+upa*(uv(m-1)-uv(m));
aa=av(m)+upa*(av(m-1)-av(m));
u(m)=ua+(aa-a0)/coef;
set(p,'color','w'); set(p,'Ydata',u,'color','y');drawnow;
uv=u;av=a;pause(0.01);end;

```

While running the program we remarked a distortion of the shape of the wave in the compression region, i.e., a shock wave is developed, see Figure 5.12 which represents u as a function of x at such a time instant.

Actually, the velocity gradient becomes so great that the viscosity of the fluid and heat transfer can no longer be neglected. In such regions the equations (5.27) break down and the computation should be stopped. The structure of a shock wave for a real gas was numerically studied in section 3.1, Chapter 4.

7. Elliptic Equations

Let us consider the Poisson equation

$$u_{xx} + u_{yy} = q(x, y) \quad (5.29)$$

to be solved in the rectangle $D = [a, b] \times [c, d]$. We know the values of u (for example $u = 0$) on the boundary of the domain. In order to

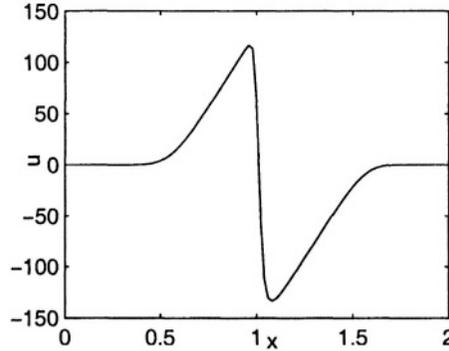


Figure 5.12. Shock waves

discretize the partial derivatives of u we introduce a grid on D , given by the lines $x = x_i = i\Delta x$, $i = 0, \dots, m + 1$, respectively $y = y_j = j\Delta y$, $j = 0, \dots, n + 1$. If we denote by u_{ij} the approximations of u at the point (x_i, y_j) and we use centered finite difference approximations for the derivatives, we have for each interior point of D ,

$$\frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\Delta x^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\Delta y^2} = q_{i,j}$$

with a truncation error of order $O(\Delta x^2 + \Delta y^2)$.

In the simple case when $\Delta x = \Delta y = h$, we find a system of $n \cdot m$ simultaneous equations and the same number of unknowns

$$u_{i,j} = \frac{1}{4} (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) - \frac{1}{4} h^2 q_{i,j}, \quad (5.30)$$

$$i = 1, \dots, m, j = 1, \dots, n.$$

7.1 Iterative Methods

We will analyze first an iterative method to solve the above system (which generally is a large system). We choose an initial approximation $u_{i,j}^{(0)}$, $i = 1, \dots, m$, $j = 1, \dots, n$ for the interior of the rectangle D . Supposing known $u_{i,j}^{(n)}$, we will calculate the next approximation from (5.30)

$$u_{i,j}^{(n+1)} = \frac{1}{4} \left(u_{i-1,j}^{(n)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n)} + u_{i,j+1}^{(n)} \right) - \frac{1}{4} h^2 q_{i,j}. \quad (5.31)$$

Let us prove that for $n \rightarrow \infty$, $u_{i,j}^{(n)} \rightarrow u_{i,j}$ (the solution of the finite difference system).

Indeed, by subtraction of the equations (5.30) and (5.31) we find for the errors at each point $e_{i,j}^{(n)} = u_{i,j}^{(n)} - u_{i,j}$,

$$e_{i,j}^{(n+1)} = \frac{1}{4} \left(e_{i-1,j}^{(n)} + e_{i+1,j}^{(n)} + e_{i,j-1}^{(n)} + e_{i,j+1}^{(n)} \right)$$

and thus

$$\left| e_{i,j}^{(n+1)} \right| \leq \frac{1}{4} \left(\left| e_{i-1,j}^{(n)} \right| + \left| e_{i+1,j}^{(n)} \right| + \left| e_{i,j-1}^{(n)} \right| + \left| e_{i,j+1}^{(n)} \right| \right).$$

Let us denote $E^{(n)}$ the greatest of these errors for the n -th iteration. Obviously, $E^{(n)} \leq E^{(n-1)}$. For the points having a neighbor on the boundary, (boundary points of the first layer) where the error vanishes, we have the estimation

$$\left| e_{i,j}^{(n+1)} \right| \leq \frac{3}{4} E^{(n)} = \left(1 - \frac{1}{4} \right) E^{(n)}.$$

For the points having as neighbor one of the above points (boundary points of the second layer), we have the estimation

$$\left| e_{i,j}^{(n+1)} \right| \leq \frac{1}{4} \left(3E^{(n)} + \left(1 - \frac{1}{4} \right) E^{(n-1)} \right) \leq \left(1 - \frac{1}{4^2} \right) E^{(n-1)}$$

So,

$$\left| e_{i,j}^{(n+1)} \right| \leq \left(1 - \frac{1}{4^M} \right) E^{(n-M+1)}$$

where M is the total number of layers in the grid. Consequently,

$$E^{(M)} \leq \left(1 - \frac{1}{4^M} \right) E^{(0)}, E^{(2M)} \leq \left(1 - \frac{1}{4^M} \right) E^{(M)}$$

and generally,

$$E^{(NM)} \leq \left(1 - \frac{1}{4^M} \right)^N E^{(0)} \rightarrow 0$$

as $N \rightarrow \infty$. Therefore, after sufficiently many iterations, the computed values will approximate as well as we wish (of course, within the limit of the computer's errors) the solution of the finite difference system (5.30).

7.1.1 Liebmann and SOR Methods

A faster iterative method is the Liebmann formula

$$u_{i,j}^{(n+1)} = \frac{1}{4} \left(u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)} \right) - \frac{1}{4} h^2 q_{ij} \quad (5.32)$$

where the new iteration is calculated from down to up and from left to right, the new computed values being immediately used.

In the particular case of a rectangle, with a uniform grid, a faster method is the *successive overrelaxation method*, shortly *S.O.R.*,

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)} + \frac{\omega}{4} \left(u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)} - 4u_{i,j}^{(n)} - h^2 q_{ij} \right)$$

where the optimal value of $\omega \in [1, 2)$ is

$$\omega = \frac{8 - 4\sqrt{4 - \alpha^2}}{\alpha^2}$$

and $\alpha = \cos \frac{\pi}{m+1} + \cos \frac{\pi}{n+1}$.

Let us consider, as an example, the fluid flow through a channel defined by $[-3, 3] \times [-2, 2]$ with an inside obstacle of boundary $f(x, y) = 0$, see [22]. The fluid enters in the channel by a hole $y \in (-0.25, 0.25)$, $x = -3$ and it freely exits through the outlet $x = 3$, as we can see in Figure 5.14.

The harmonic stream function Ψ will take the value 1 on the upper left and upper walls, the value -1 on the lower left and lower wall, it will be $4y$ on the hole and it will verify $\Psi_x = 0$ on the right wall (uniform stream).

In this case the presence of the obstacle (on the boundary of it we could take $\Psi = 0$) imposes a particular care for discretization.

Really, the (obstacle) boundary points do not usually coincide with the grid points, thus the grid points in the immediate neighborhood of the obstacle must be moved to its boundary, see Figure 5.13.

If a, b, c, d are the distances from the new nodes $(i-1, j)$, $(i+1, j)$, $(i, j-1)$, $(i, j+1)$ to the node (i, j) and we denote by $\Psi_a, \Psi_b, \Psi_c, \Psi_d$ the values of Ψ at these new nodes, we get the discretization formula of the Laplacian around the node (i, j) ,

$$\left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} \right)_{i,j} = \alpha_0 \Psi_{i,j} + \alpha_a \Psi_a + \alpha_b \Psi_b + \alpha_c \Psi_c + \alpha_d \Psi_d.$$

Let us expand Ψ_a and the others in Taylor's series and neglect the higher-order terms, obtaining

$$\Psi_a = \Psi_{i,j} - a \left(\frac{\partial \Psi}{\partial x} \right)_{i,j} + \frac{a^2}{2} \left(\frac{\partial^2 \Psi}{\partial x^2} \right)_{i,j} + \dots$$

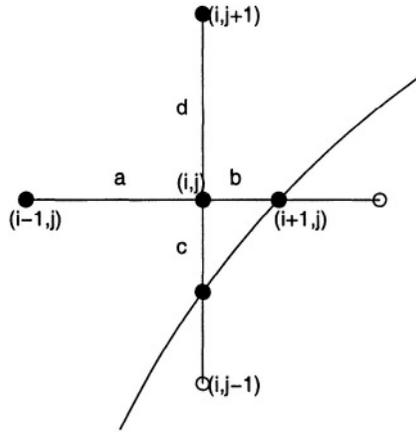


Figure 5.13. The grid near the boundary of the obstacle

and similar relations for the others. Substitution into the above relation gives

$$\begin{aligned} \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} \right)_{i,j} &= (\alpha_0 + \alpha_a + \alpha_b + \alpha_c + \alpha_d) \Psi_{i,j} \\ &+ (b\alpha_b - a\alpha_a) \left(\frac{\partial \Psi}{\partial x} \right)_{i,j} + (d\alpha_d - c\alpha_c) \left(\frac{\partial \Psi}{\partial y} \right)_{i,j} \\ &+ \frac{1}{2} (a^2\alpha_a + b^2\alpha_b) \left(\frac{\partial^2 \Psi}{\partial x^2} \right)_{i,j} + \frac{1}{2} (c^2\alpha_c + d^2\alpha_d) \left(\frac{\partial^2 \Psi}{\partial y^2} \right)_{i,j} + \dots \end{aligned}$$

and by equating corresponding coefficients and solving the obtained system we get

$$\begin{aligned} \alpha_0 &= -2 \left(\frac{1}{ab} + \frac{1}{cd} \right), \alpha_a = \frac{2}{a(a+b)}, \alpha_b = \frac{2}{b(a+b)}, \\ \alpha_c &= \frac{2}{c(c+d)}, \alpha_d = \frac{2}{d(c+d)}. \end{aligned}$$

Therefore, the iterative formula (Liebmann) (5.32) becomes

$$\Psi_{i,j} = \frac{\frac{\Psi_a}{a(a+b)} + \frac{\Psi_b}{b(a+b)} + \frac{\Psi_c}{c(c+d)} + \frac{\Psi_d}{d(c+d)}}{\frac{1}{ab} + \frac{1}{cd}}$$

for such a node.

For example, taking inside the channel an elliptical obstacle we obtain the streamlines from Figure 5.14.

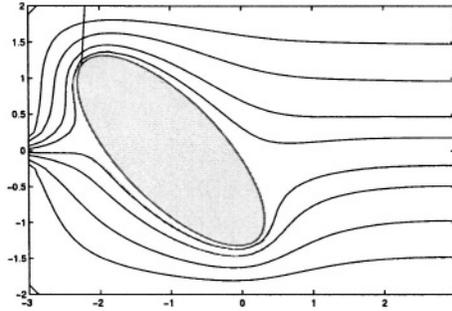


Figure 5.14. Channel flow past an elliptical obstacle

The MATLAB program is

```
clear; global x0 y0;
h=0.125; x=-3:h:3;y=-2:h:2;m=length(x);n=length(y);
for i=1:m for j=1:n
ID(i,j)=0;psi(i,j)=y(j)/2;a(i,j)=h;b(i,j)=h;c(i,j)=h;
d(i,j)=h;
if f(x(i),y(j))<=0 ID(i,j)=4;psi(i,j)=0;
elseif i==1 ID(i,j)=3;
    if y(j)<=-0.25 psi(i,j)=-1;
    elseif y(j)>=0.25 psi(i,j)=1;
    else psi(i,j)=4*y(j); end;
elseif j==1 ID(i,j)=3;psi(i,j)=-1;
elseif j==n ID(i,j)=3;psi(i,j)=1;
elseif i==m ID(i,j)=2;
    if y(j)<0 psi(i,j)=-1; elseif y(j)>0 psi(i,j)=1; end;
else
    if f(x(i-1),y(j))<0
    ID(i,j)=1;y0=y(j);a(i,j)=x(i)-fzero('fx',x(i));end;
    if f(x(i+1),y(j))<0
    ID(i,j)=1;y0=y(j);b(i,j)=fzero('fx',x(i))-x(i);end;
    if f(x(i),y(j-1))<0
    ID(i,j)=1;x0=x(i);c(i,j)=y(j)-fzero('fy',y(j));end;
    if f(x(i),y(j+1))<0
    ID(i,j)=1;x0=x(i);d(i,j)=fzero('fy',y(j))-y(j);end;
end;
end;end;
```

```

iter=1;er=1;psin=psi;
while er>1.e-2
  if rem(iter,10)==1 disp([iter er]);end;
  for i=1:m for j=1:n
  A=a(i,j);B=b(i,j);C=c(i,j);D=d(i,j);
  if ID(i,j)==0 psin(i,j)=(psin(i-1,j)+psi(i+1,j)+...
  psin(i,j-1)+psi(i,j+1))/4;
  elseif ID(i,j)==1
  psin(i,j)=1/(1/A/B+1/C/D)*(psin(i-1,j)/A/(A+B)+...
  psi(i+1,j)/B/(A+B)+psin(i,j-1)/C/(C+D)+...
  psi(i,j+1)/D/(C+D));
  elseif ID(i,j)==2
  psin(i,j)=(2*psin(i-1,j)+psin(i,j-1)+psi(i,j+1))/4;
  end;
  end;end;
iter=iter+1;er=norm(abs(psin-psi));psi=psin;
end;
contour(x,y,psi',[-1+eps -0.75 -0.5 -0.25 -0.1 0.1 ...
0.25 0.5 0.75 1]);
axis('equal');hold on;
[X,Y]=meshgrid(x,y);Z=f(X,Y);C=contour(X,Y,Z,[0 0]);
fill(C(1,:),C(2,:),'r');hold off;
which uses the function subprograms
function z=f(x,y)
global x0 y0;
z=(x+y+1).^2+(y-x-1).^2/6-1;

function z=fx(x)
global x0 y0;
y=y0;
z=f(x,y);

function z=fy(y)
global x0 y0;
x=x0;
z=f(x,y);

```

The program calculates the boundary of the obstacle from the equation $f(x, y) = 0$ and the new grid points on that boundary are calculated by solving the equations $f(x, y_j) \equiv fx(y) = 0$ respectively $f(x_i, y) \equiv fy(y) = 0$. An error (the difference between two successive iterations) less than 0.01 is obtained after about 60 iterations.

7.1.2 ADI Method

An interesting iterative algorithm, often used, could be obtained by introducing a fictitious diffusion problem

$$\begin{aligned}\frac{\partial u}{\partial t} &= c \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) - cf(x, y), \\ u(x, y, 0) &= U(x, y), \\ u|_{\partial\Omega} &= 0,\end{aligned}$$

with a suitable initial condition U . The solution of this unsteady problem for large t approximates the solution of the Poisson equation and thus, the time marching for the above equation gives an iterative method to solve the problem (5.29).

By discretizing the above equations with the finite differences method on the grid $x = x_i$, $i = 1, \dots, n$ respectively $y = y_j$, $j = 1, \dots, M$ with the step sizes $\Delta x, \Delta y$ and by denoting as usual u_{ij}^k the approximation of the exact solution $u(x_i, y_j)$ at the grid points and at the moment t_k (with the time step size Δt), the explicit Euler method with respect to the time leads to

$$\frac{u_{ij}^{k+1} - u_{ij}^k}{\Delta t} = c \left(\frac{u_{i+1,j}^k - 2u_{ij}^k + u_{i-1,j}^k}{\Delta x^2} + \frac{u_{i,j+1}^k - 2u_{ij}^k + u_{i,j-1}^k}{\Delta y^2} \right) - cf_{ij}.$$

Unfortunately, the stability requirements

$$c \left(\frac{\Delta t}{\Delta x^2} + \frac{\Delta t}{\Delta y^2} \right) < \frac{1}{2}$$

make the method useless. But Peaceman, Rachford and Douglas, in 1955, proposed the decomposition of each time step into two steps of length $\Delta t/2$ and a semi-implicit treatment of the spatial derivatives, obtaining the so-called *ADI (alternating direction implicit) method*, see,

for instance, [125], [120]:

$$\begin{aligned} & \frac{u_{ij}^{k+1/2} - u_{ij}^k}{\Delta t/2} \\ &= c \left(\frac{u_{i+1,j}^{k+1/2} - 2u_{ij}^{k+1/2} + u_{i-1,j}^{k+1/2}}{\Delta x^2} + \frac{u_{i,j+1}^k - 2u_{ij}^k + u_{i,j-1}^k}{\Delta y^2} \right) - cf_{ij}, \\ & \frac{u_{ij}^{k+1} - u_{ij}^{k+1/2}}{\Delta t/2} \\ &= c \left(\frac{u_{i+1,j}^{k+1/2} - 2u_{ij}^{k+1/2} + u_{i-1,j}^{k+1/2}}{\Delta x^2} + \frac{u_{i,j+1}^{k+1} - 2u_{ij}^{k+1} + u_{i,j-1}^{k+1}}{\Delta y^2} \right) - cf_{ij}. \end{aligned}$$

In fact, at the first half-step the x direction is implicitly treated and then, at the second half-step the y direction is implicitly treated too. In the sequel this order is reversed, to avoid the break of the solution into independent components. At every half-step we have to solve a tri-diagonal system of algebraic linear equations, which is no longer a difficult problem. The scheme has a second order of accuracy in space and in time and it is unconditionally stable, thus it is the usually used algorithm for practical problems. Of course, a sufficiently large number of time steps must be considered and the final accuracy depends on the spatial step sizes.

A project of MATLAB program is

```
n=64;h=1/n;x=linspace(0,1,n+1);y=x;
[X,Y]=meshgrid(x(2:n),y(2:n));
F=-5*pi^2*sin(pi*X').*sin(2*pi*Y');
U=zeros(n-1);dx=h;dt=0.1;al=dt/dx^2;e=ones(n-1,1);
Uex=sin(pi*X').*sin(2*pi*Y');
A=spdiags([al*e -2*(1+al)*e al*e],-1:1,n-1,n-1);
B=spdiags([-al*e -2*(1-al)*e -al*e],-1:1,n-1,n-1);
for kod=1:30
    Ui=A\(U*B'+dt*F);Uii=(B*Ui+dt*F)/A';
    Ui=(B*Uii+dt*F)/A';Un=A\(Ui*B'+dt*F);
    err=max(max(abs(U-Un)));disp([kod err]);U=Un;
end
surf(X',Y',abs(U-Uex));
```

for the problem

$$\begin{aligned} \Delta u &= -5\pi^2 \sin x \sin 2y, \quad \text{in } \Omega = [0, 1] \times [0, 1], \\ u|_{\partial\Omega} &= 0. \end{aligned}$$

With this data for the program an accuracy of about 8×10^{-4} is obtained. The algorithm can be easily extended to the three-dimensional cases.

7.2 **Direct Method**

In the case of a rectangle, we may also use an exact method for solving the system (5.30).

We remark that the computation of the second order derivative with respect to x is, in fact, the multiplication of the values of u on the grid, from left, with a differentiation matrix S_m ,

$$\begin{pmatrix} u_{xx1,1} & \cdots & u_{xx1,n} \\ \vdots & & \vdots \\ u_{xxm,1} & \cdots & u_{xxm,n} \end{pmatrix} = S_m \cdot \begin{pmatrix} u_{1,1} & \cdots & u_{1,n} \\ \vdots & & \vdots \\ u_{m,1} & \cdots & u_{m,n} \end{pmatrix},$$

$$S_m = \begin{pmatrix} -2 & 1 & & & & \\ \frac{1}{h^2} & \frac{1}{h^2} & & & & \\ 1 & -2 & \ddots & & & \\ \frac{1}{h^2} & \frac{1}{h^2} & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & \frac{1}{h^2} & \\ & & & & \frac{1}{h^2} & -2 \\ & & & & & \frac{1}{h^2} \end{pmatrix}$$

where we have taken into account the null values of u on the boundary (else they pass to the right-hand side of the discretized system). The second derivative with respect to y is similarly calculated, by multiplication of the values of u on the grid, from right, with the matrix S_n^T (the transposed differentiation matrix).

In the case of the problem

$$\begin{aligned} u_{xx} + u_{yy} &= q, \quad \text{in } \Omega = [0, a] \times [0, b], \\ u|_{\partial\Omega} &= 0, \end{aligned}$$

by discretization of the second derivatives at the points $(x_i, y_j) = (ih, jh)$, $i = 1, \dots, m$, $j = 1, \dots, n$ and $h = \frac{a}{m+1} = \frac{b}{n+1}$ we obtain a system with the unknowns $U_{ij} = u(x_i, y_j)$ of the form

$$S_m U + U S_n^T = F.$$

If we denote the right eigenvectors (columns) matrix of S_m by P_m , we have $P_m^{-1}S_mP_m = \Lambda_m$, where Λ_m is the diagonal matrix of eigenvalues. The computation is similar for S_n . So that, multiplying the above system by P_m^{-1} from the left and by P_n^{-1T} from the right, it becomes

$$P_m^{-1}S_mP_mP_m^{-1}UP_n^{-1T} + P_m^{-1}UP_n^{-1T}P_n^T S_n^T P_n^{-1T} = P_m^{-1}FP_n^{-1T}$$

or

$$\Lambda_m \tilde{U} + \tilde{U} \Lambda_n^T = \tilde{F}$$

from which,

$$\tilde{U}_{ij} = \frac{\tilde{F}_{ij}}{\lambda_i^{(m)} + \lambda_j^{(n)}}, i = 1, \dots, m, j = 1, \dots, n$$

and next, from \tilde{U} we calculate $U = P_m \tilde{U} P_n^T$.

The computing effort is the diagonalization of S_n and S_m , but this is performed only once. We remark that for tridiagonal and constant coefficients matrices, like ours, there are analytical formulas for eigenvalues and eigenvectors. So, for the matrix $n \times n$,

$$\begin{pmatrix} a & b & & \\ c & a & \ddots & \\ & \ddots & \ddots & b \\ & & c & a \end{pmatrix}$$

we have the eigenvalues [124]

$$\lambda_j = a + 2\sqrt{bc} \cos\left(\frac{j\pi}{n+1}\right), j = 1, \dots, n$$

and the right eigenvectors matrix is

$$P_{ij} = \sqrt{\frac{c}{b}} \sin\left(\frac{ij\pi}{n+1}\right), i, j = 1, \dots, n,$$

no other calculations being required.

For S_n we have, consequently,

$$\lambda_j^{(n)} = (n+1)^2 \left[-2 + 2 \cos\left(\frac{j\pi}{n+1}\right) \right], j = 1, \dots, n$$

and

$$P_{ij}^{(n)} = \sin\left(\frac{ij\pi}{n+1}\right), i, j = 1, \dots, n$$

(and similarly for S_m), which next should be normed, $P := P / \text{norm}(P)$. We have moreover $P^{-1} = P^T = P$.

We will present an example of a fluid flow leading to such a problem, i.e., a rectangular domain with vorticity.

Let us study the flow generated by a distribution of vorticity within a rectangular domain $[a, b] \times [c, d]$, following [22]. As we know, the vorticity is a vector in the Oz direction with the magnitude q , defined as the curl of the velocity vector $\nabla \times \mathbf{V} = \mathbf{q}$. Using the relation between the velocity components and the stream function Ψ , the above relationship can be written in the scalar form (passing to the magnitude q)

$$\Psi_{xx} + \Psi_{yy} = -q(x, y).$$

In a particular case of the domain $D = [-3, 3] \times [-2, 2]$, with the vorticity generated by the point vortices of strengths 100, respectively -50 , located at the points $(1, 1)$ respectively $(1, 0)$, one obtains the streamlines from Figure 5.15.

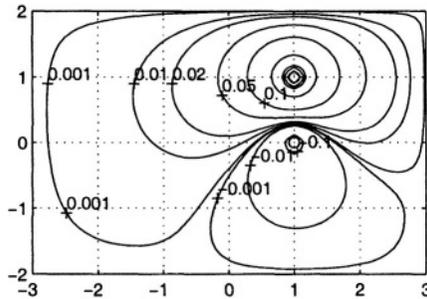


Figure 5.15. Streamlines generated by two line vortices

The MATLAB program is

```
x=-3:0.1:3;y=-2:0.1:2;
c=zeros(59,39);c(40,30)=-100;c(40,20)=-50;
e1=ones(59,1);e2=ones(39,1);
D23=spdiags([e1 -2*e1 e1],[-1 0 1],59,59)*100;
D15=spdiags([e2 -2*e2 e2],[-1 0 1],39,39)*100;
[U,L]=eig(full(D23));[V,P]=eig(full(D15'));
PSI=zeros(59,39);
C=inv(U)*c*V;
```

```

for i=1:59 for j=1:39
PSI(i,j)=C(i,j)/(L(i,i)+P(j,j));
end;end;
psi=U*PSI*inv(V);
S=zeros(61,41);S(2:60,2:40)=psi;
cs=contour(x,y,S',[0.01 0.02 0.05 0.1 0.2 0.3 0.4]);
clabel(cs,'manual');axis('equal');hold on;
plot(1,1,'o',1,0,'o');hold off;

```

where the discrete system was exactly solved. The stream function Ψ was chosen such that it takes the 0 value on the streamline which represents the boundary of the domain.

For further details concerning the solvability of large systems of equations we recommend [128].

7.3 Transonic Flows

Let us consider now the important problem of the calculation of a plane steady transonic flow. Precisely, we will present a computing procedure for a steady, inviscid, transonic fluid flow past an airfoil. In the case of a velocity close to the sound velocity, a zone with supersonic velocity appears near the airfoil, leading to the shock waves. Mathematically, the phenomenon is described by mixed partial differential equations: elliptic in the subsonic region and hyperbolic in the supersonic zone. The discretization procedure will take into account this aspect. Moreover, in the physical domain, the rapid changes of the flowfield around the airfoil arise so there we must refine the computing grid.

After the basic paper of E. M Murman [92], the simplest mode (but not the most accurate) to numerically calculate this flowfield is the use of the transonic small disturbance theory. What we are calculating is in fact the induced small disturbance on the uniform stream due to the presence of the airfoil.

We scale the y coordinate to $\tilde{y} = \delta^{1/3}y$ where δ is the airfoil thickness ratio and we consider the velocity potential Φ for which the velocity disturbances are

$$u = \Phi_x, v = \Phi_{\tilde{y}}.$$

The flow is governed by the unique equation

$$[K - (\gamma + 1)\Phi_x] \Phi_{xx} + \Phi_{\tilde{y}\tilde{y}} = 0$$

where K is a similarity parameter $K = \frac{(1-M_\infty^2)}{\delta^{2/3}}$ (when the unperturbed velocity increases to the sound speed, K decreases), M_∞ is the free stream Mach number and γ is the ratio of specific heats. For the concrete calculations we take $K = 1.3$ and $\gamma = 1.4$.

We can see that in the regions with higher velocity $\Phi_x > \frac{K}{\gamma+1}$ and the equation is locally hyperbolic, while in the regions with smaller velocity $\Phi_x < \frac{K}{\gamma+1}$ and the equation is elliptic.

As computational domain we choose a rectangle whose base $\tilde{y} = 0$ represents the (upper) airfoil surface on the interval $[-1, 1]$. The boundary conditions will be

$$\Phi_{\tilde{y}}(x, 0) = \begin{cases} F'(x), & |x| < 1 \\ 0, & |x| \geq 1 \end{cases}$$

where the airfoil equation is $y = \delta F(x)$. In the sequel, for sake of simplicity, we will consider $F(x) = 1 - x^2$. On the other sides of the computational rectangle we consider, as boundary conditions, the unperturbed values of Φ , the usual doublet for a closed body

$$\Phi(x, \tilde{y}) = \frac{\mathcal{D}}{2\pi\sqrt{K}} \frac{x}{x^2 + K\tilde{y}^2} + \dots$$

and we keep only the written term in the above series. Here \mathcal{D} is the doublet strength,

$$\mathcal{D} = 2 \int_{-1}^1 F(\xi) d\xi + \frac{\gamma + 1}{2} \int \int_{-\infty}^{\infty} u^2 d\xi d\eta$$

which, during the calculations, will be approximated (after every step) by reducing the double integral to an integral on the computational domain.

So, let us consider a mesh with meshlines $x = x_i, \tilde{y} = \tilde{y}_j$ and, as initial approximation, a uniform flow. We approximate at every node

$$(\Phi_x)_{i,j} \approx \frac{\Phi_{i+1,j} - \Phi_{i-1,j}}{2\Delta x}$$

and, depending on the result of the comparison with $\frac{K}{\gamma+1}$, the equation will be discretized as follows:

- in the elliptic case, $\Phi_x < \frac{K}{\gamma+1}$,

$$(\Phi_x)_{i,j} = \frac{\Phi_{i+1,j} - \Phi_{i-1,j}}{2\Delta x},$$

$$(\Phi_{xx})_{i,j} \approx \frac{\Phi_{i+1,j} - 2\Phi_{i,j} + \Phi_{i-1,j}}{\Delta x^2},$$

- in the hyperbolic case, $\Phi_x > \frac{K}{\gamma+1}$,

$$(\Phi_x)_{i,j} = \frac{\Phi_{i,j} - \Phi_{i-2,j}}{2\Delta x},$$

$$(\Phi_{xx})_{i,j} \approx \frac{\Phi_{i,j} - 2\Phi_{i-1,j} + \Phi_{i-2,j}}{\Delta x^2}.$$

In both cases,

$$(\Phi_{\tilde{y}\tilde{y}})_{i,j} \approx \frac{\Phi_{i,j+1} - 2\Phi_{i,j} + \Phi_{i,j-1}}{\Delta\tilde{y}^2}$$

excepting the first computational row $\tilde{y}_{i,1} = \Delta\tilde{y}/2$, where

$$(\Phi_{\tilde{y}\tilde{y}})_{i,1} \approx \frac{1}{\Delta\tilde{y}} \left[\frac{\Phi_{i,2} - \Phi_{i,1}}{\Delta\tilde{y}} - \Phi_{\tilde{y}}|_{\tilde{y}=0} \right]$$

Finally, we take at the horizontal axis

$$\Phi_{i,0} = \Phi_{i,1} - \Delta\tilde{y}\Phi_{\tilde{y}}|_{\tilde{y}=0}.$$

At every iteration we evaluate \mathcal{D} by

$$\mathcal{D} = \sum_i \sum_j (\Phi_x)_{i,j}^2 \Delta x \Delta\tilde{y}$$

and thus we can modify the values on the boundary of the computational domain. Finally, the values on the horizontal axis needed for the pressure coefficient are approximated by extrapolation from the internal nodes

$$\Phi_{i,0} = \frac{9}{8}\Phi_{i,1} - \frac{1}{8}\Phi_{i,2} - \frac{3}{8}\Delta\tilde{y} \cdot \Phi_{\tilde{y}}|_{\tilde{y}=0}.$$

The different discretizations of the equation in the different zones are imposed by the different dependence domains. In the elliptic case, this is the whole computational domain and the node for the new computed value is surrounded by the old ones. Conversely, in the hyperbolic case, the dependence domain is only the angle between the two characteristics through the node and the new value uses only those at the upwind nodes.

The discrete system is iteratively solved, considering the time evolution of the phenomenon. If we denote at each step the system to solve by $L\Phi = 0$, we attach to this problem the equation

$$\frac{\partial\Phi}{\partial t} = L\Phi$$

which may be discretized in time by, for instance,

$$\Phi^{(n+1)} = \Phi^{(n)} + \Delta t \cdot L\Phi^{(n)}.$$

Choosing the time step size sufficiently small in order to ensure the computational stability and performing a sufficiently large number of steps in order to approximate well the steady solution, we obtain the results from Figures 5.16, 5.17 and 5.18.

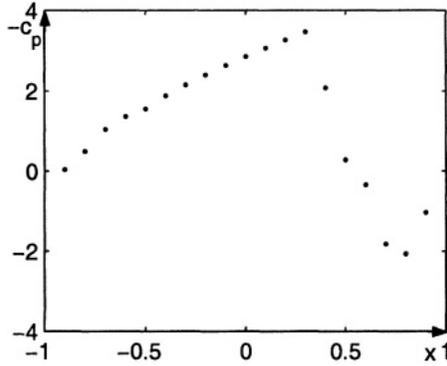


Figure 5.16. The pressure coefficient $-c_p$

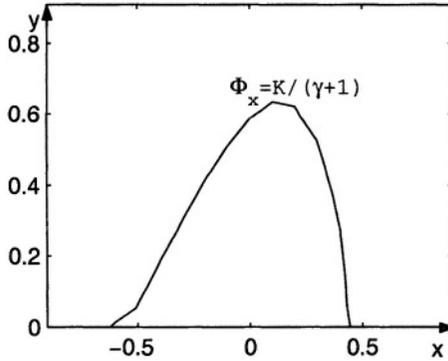


Figure 5.17. The sonic line

The first figure shows the pressure coefficient $c_p = -2u$ on the airfoil surface x (with changed sign). We can see the rapid change on the right, due to the presence of the shock wave. The second figure shows the sonic line shape $\Phi_x = \frac{K}{\gamma+1}$, which separates the subsonic zone (outside) and the supersonic zone (inside). The last figure shows the velocity field near the airfoil and the shock wave.

In order to increase the accuracy, we can use the transformed coordinates

$$\begin{aligned} \xi &= (x + c)^{-1} + d, \\ \eta &= (y + a)^{-1} + b, \end{aligned}$$

that refine the mesh near the airfoil. The computational domain is $[-5.6, 5.6] \times [0, 7.5/\sqrt{K}]$, using 111 nodes on Ox direction and 62 on Oy direction. We have performed 2000 time iterations. We remark that

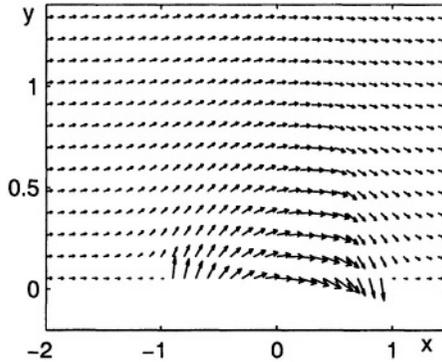


Figure 5.18. The velocity field

this procedure is not the fastest (or the most accurate) but it is easy to understand.

The MATLAB code is

```

K=1.3;g=1.4;m=111;n=62;D=8/3;
x=linspace(-5.6,5.6,m+2);dx=11.2/(m+1);dt=0.001;
dy=7.5/sqrt(K)/n;y=[0 linspace(dy/2,7.5/sqrt(K),n)];
[X,Y]=meshgrid(x,y);X=X';Y=Y';Fy=zeros(1,m+2);
for i=1:m+2 if (1-x(i)^2)>0 Fy(i)=-2*x(i); end;end;
F=ones(m+2,n+1)/20; F([1 2 m+2],:)=...
D/(2*pi*sqrt(K))*X([1 2 m+2],:)./...
(X([1 2 m+2],:).^2+K*Y([1 2 m+2],:).^2);
F(:,n+1)=D/(2*pi*sqrt(K))*X(:,n+1)./...
(X(:,n+1).^2+K*Y(:,n+1).^2);
F(:,1)=F(:,2)-dy*Fy'; mesh(F);
Fx=(F(4:m+2,2:n)-F(2:m,2:n))/2/dx;t=dt;iter=1;
while 1 a=zeros(m-1,n-1);
a(find(Fx<K/(g+1)))=ones(size(find(Fx<K/(g+1))));
Le=((K-(g+1)*(F(4:m+2,2:n)-F(2:m,2:n))/2/dx).*...
(F(4:m+2,2:n)+F(2:m,2:n)-2*F(3:m+1,2:n))/dx^2+...
(F(3:m+1,3:n+1)+F(3:m+1,1:n-1)-2*F(3:m+1,2:n))/dy^2);
Lh=((K-(g+1)*(F(3:m+1,2:n)-F(1:m-1,2:n))/2/dx).*...
(F(3:m+1,2:n)-2*F(2:m,2:n)+F(1:m-1,2:n))/dx^2+...
(F(3:m+1,3:n+1)+F(3:m+1,1:n-1)-2*F(3:m+1,2:n))/dy^2);
Fn=F(3:m+1,2:n)+dt*(a.*Le+(1-a).*Lh);
err=sum(sum(abs(F(3:m+1,2:n)-Fn)));
if rem(iter,10)==0 disp([iter/10 t err]);end;
F(3:m+1,2:n)=Fn;F(:,1)=F(:,2)-dy*Fy';

```

```

Fx=(F(4:m+2,2:n)-F(2:m,2:n))/2/dx;
Dn=8/3+(g+1)/2*sum(sum(Fx.^2))*dx*dy;
F([1 2 m+2],:)=Dn/D*F([1 2 m+2],:);
F(:,n+1)=Dn/D*F(:,n+1);
D=Dn;t=t+dt;iter=iter+1;
if err<0.05 break; end;
end;
F(:,1)=9/8*F(:,2)-F(:,3)/8-3/8*Fy'*dy;
Fx0=(F(4:m+2,1)-F(2:m,1))/2/dx;
cp=2*Fx0;I=find((1-x.^2)>0);J=find(y<1);
plot(x(I),cp(I),'.');pause;
u=(F(4:m+2,2:n)-F(2:m,2:n))/2/dx;
v=(F(3:m+1,3:n+1)-F(3:m+1,1:n-1))/2/dx;
v(:,1)=Fy(1,3:m+1)';
contour(x(I),y(J),u(I,J)',[K/(g+1) K/(g+1)]);pause;
quiver(X(3:m+1,2:n),Y(3:m+1,2:n),u+1,v);
axis([-2 1.5 -0.2 1.4]);

```

7.4 Stokes' Problem

We will firstly consider the steady state case

$$\begin{aligned}\nabla p &= \frac{1}{R} \nabla^2 \mathbf{V}, \\ \nabla \cdot \mathbf{V} &= 0, \\ \mathbf{V}|_{\partial\Omega} &= \mathbf{V}_{fr},\end{aligned}\tag{5.33}$$

where $\mathbf{V} = (u, v)$. Let us present an example leading to such a problem.

In the case of a viscous incompressible flow, the Reynolds number R measures the relative importance of the inertial forces vs. viscous forces in the flow. If the Reynolds number R is large, the viscous force terms in the Navier–Stokes equations become small in comparison with the others. In this case the viscous forces are important only in a relatively small region in the neighborhood of the surface of the fluid – the boundary layer. If R is much smaller than unity, the viscous forces are dominant on the fluid flow.

By eliminating the terms describing the inertial forces in the Navier–Stokes equations we obtain for the steady state the equation

$$\mu \nabla^2 \mathbf{V} = \nabla p.$$

Such flows for which $R \ll 1$ are called *Stokes flows* and the above equation is called *the Stokes equation* (see also sections 3.3 and 4.4.4). Taking the curl of the above equation we are led to

$$\nabla^2 \zeta = 0$$

where $\zeta = \nabla \times \mathbf{V}$ is the vorticity. Similarly, divergence of the equation yields, based on the incompressibility $\nabla \cdot \mathbf{V} = 0$,

$$\nabla^2 p = 0.$$

In the particular case of a two-dimensional Stokes flow in the xy plane, by introducing the stream function Ψ for which

$$u = \frac{\partial \Psi}{\partial y}, v = -\frac{\partial \Psi}{\partial x}$$

we find that the only nonvanishing vorticity component is that in the Oz direction and

$$\nabla^2 \Psi = -\zeta.$$

By using the (scalar) equation verified by the vorticity, we also get

$$\nabla^4 \Psi = 0$$

where

$$\nabla^4 = \nabla^2 \nabla^2 = \frac{\partial^4}{\partial x^4} + 2 \frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}$$

is the biharmonic operator and the above equation is the *biharmonic equation* for ψ .

Let us consider, for instance, following [22] a square cavity ABCD, see Figure 5.19.

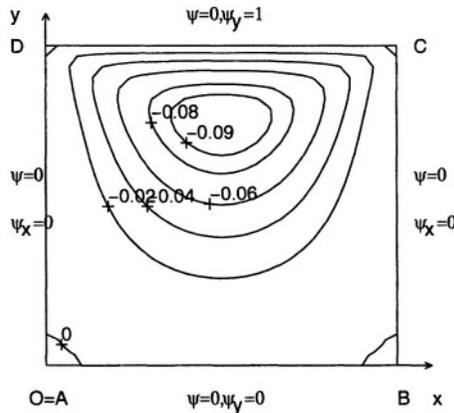


Figure 5.19. Driven cavity flow

Here the steady fluid flow is generated by sliding the lid (an infinitely long plate on the top of the cavity). We assume that the dimensions are

normalized, i.e., the cavity is the square $[0,1] \times [0,1]$, the horizontal lid velocity is 1 m/s and the Reynolds number is so small that we deal with a Stokes flow.

As there are no fluid changes between the cavity inside and outside, the fluid flow forms a closed path within the cavity. The surfaces DA , AB , BC , CD will determine the streamline $\Psi = 0$ and so, the normal velocities to these surfaces are all zero. We will require that the tangential velocity to these surfaces vanishes too, excepting on the lid CD , where it is equal to 1. So, the biharmonic equation for the stream function is joined with eight boundary conditions, precisely

$$\begin{aligned} \Psi|_{DA} = 0, \quad \frac{\partial \Psi}{\partial x} \Big|_{DA} &= 0, \\ \Psi|_{AB} = 0, \quad \frac{\partial \Psi}{\partial y} \Big|_{AB} &= 0, \\ \Psi|_{BC} = 0, \quad \frac{\partial \Psi}{\partial x} \Big|_{BC} &= 0, \\ \Psi|_{CD} = 0, \quad \frac{\partial \Psi}{\partial y} \Big|_{CD} &= 1. \end{aligned}$$

We would solve this problem by the finite differences method. Let us cover the cavity with a square mesh of step size h . The discretization of the biharmonic equation is [124]

$$\begin{aligned} \nabla^4 f_{i,j} \equiv \frac{1}{h^4} [20f_{i,j} - 8(f_{i+1,j} + f_{i-1,j} + f_{i,j+1} + f_{i,j-1}) \\ + 2(f_{i+1,j+1} + f_{i-1,j+1} + f_{i+1,j-1} + f_{i-1,j-1}) \\ + (f_{i,j+2} + f_{i,j-2} + f_{i+2,j} + f_{i-2,j})] = 0. \end{aligned}$$

At the boundary nodes we assume $f_{ij} = 0$. For the boundary conditions containing derivatives, we will use, for discretization, centered first order finite differences. So, we will consider a layer of fictitious nodes of step size h outside the domain. The nodes numbering will be:

- x_1 fictitious node at the left side (outside) of AD .
- x_2 boundary node at AD . Here $f = 0$.
- x_3 inside (computational) node in the Ox direction. Here f will be calculated.
-
- x_{m-1} inside (computational) node at the Ox direction. Here f will be calculated.
- x_m boundary node. Here $f = 0$.
- x_{m+1} fictitious node at the right (outside) of BC .

Analogously we make the numbering in the Oy direction : y_1, \dots, y_{m+1} . Thus, at AD we will have

$$\begin{aligned} f_{2,j} \equiv f(x_2, y_j) = 0, \quad j = 3, \dots, m - 1, \\ f_{1,j} = f_{3,j}, \quad j = 3, \dots, m - 1 \end{aligned}$$

and similarly at the sides AB and BC . At the side CD where $\Psi_y = 1$ the conditions will be

$$\begin{aligned} f_{i,m} &\equiv f(x_i, y_m) = 0, i = 3, \dots, m-1, \\ f_{i,m+1} &= f_{i,m-1} + 2h \end{aligned}$$

While we discretize the biharmonic equation at the inside boundary neighboring nodes, the values of f at the fictitious nodes appear. Here we will use the above equations.

So, the biharmonic equation will be discretized at the inside nodes (x_i, y_j) , where $i = 3, \dots, m-1$ and $j = 1, \dots, m-1$ obtaining $(m-3)^2$ linear equations with the same number of unknowns $f_{i,j}$, and a linear algebraic system of the form $R \equiv AS - b = 0$ is obtained. The matrix A of this system and the right-hand side terms are difficult to be manually written, but A is a sparse matrix and in our case, for $m = 32$, we are able to calculate the exact solution of the system.

The first part of the code, cf. [124], automatically determines the right-hand side b and then the matrix A . By systematically numbering the mesh nodes and arranging the unknowns f into a column vector S of size $(m-3)^2$, we observe that

$$b = -R|_{S=0}$$

and hence

$$A_{i,j} = \frac{\partial R_i}{\partial s_j} = R_i(S_3 = 0, \dots, S_j = 1, \dots, S_{m-1} = 0) + b$$

for $i, j = 3, \dots, m-1$. What we need is such a subprogram which calculates R from a given S , i.e., the subprogram `rez.m`

```
function R=rez(S)
m=length(S)-1;h=1/(m-2);
R=20*S(3:m-1,3:m-1)-8*(S(4:m,3:m-1)+...
S(2:m-2,3:m-1)+S(3:m-1,4:m)+S(3:m-1,2:m-2))+...
2*(S(4:m,4:m)+S(4:m,2:m-2)+S(2:m-2,4:m)+...
S(2:m-2,2:m-2))+S(3:m-1,5:m+1)+S(3:m-1,1:m-3)+...
S(5:m+1,3:m-1)+S(1:m-3,3:m-1));
R=reshape(R,(m-3)^2,1);
```

The main program must complete the boundary and fictitious layers of S (which is of size $(m+1) \times (m+1)$), with the above mentioned values and next it must compute b and A . Finally, it should solve the algebraic linear system and plot the solutions representing the stream function values on the mesh nodes. The MATLAB code is

```
m=32;S=zeros(m+1);h=1/(m-2);
for j=3:m-1 S(1,j)=S(3,j);S(m+1,j)=S(m-1,j);end;
```

```

for i=3:m-1 S(i,1)=S(i,3);S(i,m+1)=S(i,m-1)+2*h;end;
A=spalloc((m-3)^2,(m-3)^2,13*(m-3)^2);
B=-rez(S);p=0;
for jp=3:m-1 for ip=3:m-1 p=p+1;S(ip,jp)=1;
for j=3:m-1 S(1,j)=S(3,j);S(m+1,j)=S(m-1,j);end;
for i=3:m-1 S(i,1)=S(i,3);S(i,m+1)=S(i,m-1)+2*h;end;
A(:,p)=rez(S)+B;S(ip,jp)=0; end;end;
X=A\B;
X=reshape(X,m-3,m-3);S(3:m-1,3:m-1)=X;
x=linspace(0,1,m-1);h=1/(m-2);y=x;
cs=contour(x',y',S(2:m,2:m)',...
[0 -0.02 -0.04 -0.06 -0.08 -0.09]);
xlabel(cs,'manual');

```

The result, the streamlines, can be seen in Figure 5.19.

We will present other methods to bypass the difficulties generated by the presence of the equation $\nabla \cdot \mathbf{V} = 0$. Let us consider the evolution problem

$$\begin{aligned} \frac{\partial \mathbf{V}}{\partial t} + \nabla p &= \frac{1}{R} \nabla^2 \mathbf{V}, \\ \frac{\partial p}{\partial t} + c^2 \nabla \cdot \mathbf{V} &= 0 \end{aligned} \quad (5.34)$$

associated with the boundary conditions for \mathbf{V} and some suitable initial conditions for \mathbf{V} and p at $t = 0$. Here $c^2 > 0$ should be chosen so that the convergence of the solutions of the problem (5.34) toward the steady solution of the Stokes problem (5.33) when $t \rightarrow \infty$ is assured. Obviously, the second equation from (5.34) has no physical meaning before the steady state is reached. Consequently, the above method is only a tool to generate an iterative algorithm to approximate the steady solution of the Stokes problem.

The numerical solving of the problem (5.34) will be performed by the spatial discretization with finite differences on a mesh *MAC* (*marker and cell*), introduced by Harlow and Welsh, and with the simple forward Euler time discretization. We follow Peyret and Taylor [120], where the convergence and the stability of this scheme is analyzed.

The key element is the choice of the staggered mesh for the discretization of u, v respectively p (see Figure 5.20).

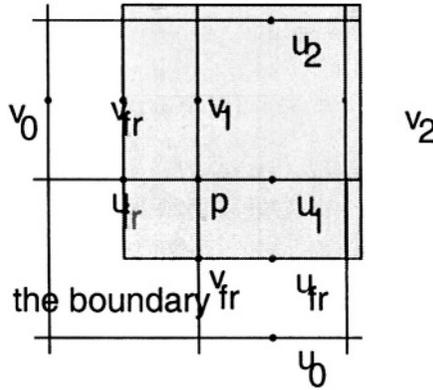


Figure 5.20. The MAC mesh

The discretized equations are:

$$\begin{aligned} & \frac{1}{\Delta t} \left(u_{i+\frac{1}{2},j}^{n+1} - u_{i+\frac{1}{2},j}^n \right) + \frac{1}{\Delta x} \left(p_{i+1,j}^n - p_{i,j}^n \right) \\ &= \frac{1}{R} \left(\frac{u_{i+\frac{3}{2},j}^n - 2u_{i+\frac{1}{2},j}^n + u_{i-\frac{1}{2},j}^n}{\Delta x^2} + \frac{u_{i+\frac{1}{2},j+1}^n - 2u_{i+\frac{1}{2},j}^n + u_{i+\frac{1}{2},j-1}^n}{\Delta y^2} \right), \\ & \frac{1}{\Delta t} \left(v_{i,j+\frac{1}{2}}^{n+1} - v_{i,j+\frac{1}{2}}^n \right) + \frac{1}{\Delta y} \left(p_{i,j+1}^n - p_{i,j}^n \right) \\ &= \frac{1}{R} \left(\frac{v_{i+1,j+\frac{1}{2}}^n - 2v_{i,j+\frac{1}{2}}^n + v_{i-1,j+\frac{1}{2}}^n}{\Delta x^2} + \frac{v_{i,j+\frac{3}{2}}^n - 2v_{i,j+\frac{1}{2}}^n + v_{i,j-\frac{1}{2}}^n}{\Delta y^2} \right), \\ & \frac{1}{\Delta t} \left(p_{i,j}^{n+1} - p_{i,j}^n \right) + \frac{c^2}{\Delta x} \left(u_{i+\frac{1}{2},j}^{n+1} - u_{i-\frac{1}{2},j}^{n+1} \right) + \frac{c^2}{\Delta y} \left(v_{i,j+\frac{1}{2}}^{n+1} - v_{i,j-\frac{1}{2}}^{n+1} \right) = 0 \end{aligned} \tag{5.35}$$

where $u_{i+\frac{1}{2},j}^n$ (and similarly for v, p) means the approximate value of u at the spatial node $(i + \frac{1}{2}, j)$ and at the time instant $t_n = n\Delta t$. The above approximations are of second order accuracy.

The necessary stability conditions of the above scheme are (Peyret, Taylor)

$$\frac{4\Delta t}{R\Delta x^2} \leq 1, \quad \frac{4\Delta t}{\Delta x^2} \left(\frac{1}{R} + \frac{c^2\Delta t}{2} \right) \leq 1.$$

As regards the behaviour of the discretization in the neighborhood of the boundary, we remark (Figure 5.21) that the pressure appears only at the inside nodes of the domain Ω , so we do not use pressure values on the boundary. We also remark that the above formulas involve the values of u only at vertical boundaries, respectively the values of v only at the horizontal boundaries. But we have values of v (in the discretization of Δv near the boundary) and of u (in the discretization of Δu near the boundary) at nodes outside the computing domain, values which should be calculated by extrapolation of the inside and boundary values.

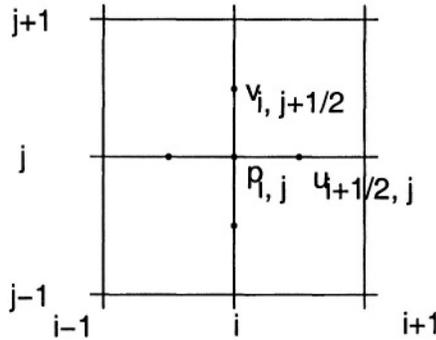


Figure 5.21. Boundary nodes

So, in order to calculate v_0 by left extrapolation of the inside values v_2, v_1 and of the boundary value v_{fr} with a quadratic polynomial, we find

$$v_0 = \frac{1}{3} (v_2 - 6v_1 + 8v_{fr}) \tag{5.36}$$

and, symmetrically, by right extrapolation,

$$v_{m+1} = \frac{1}{3} (v_{m-1} - 6v_m + 8v_{fr}).$$

Similar formulas may be also written for u .

Consequently, the algorithm consists in the following :

- step 1: from initial conditions we have u, v, p at the inside nodes
- step 2: from boundary conditions we have u, v at the boundary nodes
- step 3: we calculate u at the outside nodes (in the neighborhood of the horizontal boundaries)
- step 4: we calculate v at the outside nodes (in the neighborhood of the vertical boundaries), using (5.36)
- step 5: we calculate u, v, p at the inside nodes, using (5.35)

– step 6: we evaluate the differences between the old and the new values of u, v, p . If these differences are not sufficiently small, we will resume the algorithm from step 3; if the differences are sufficiently small, we extract the results, which represent the approximations of the solution of the steady Stokes problem.

As an example, let us solve the Stokes problem for the domain $\Omega = [0,1] \times [0,1]$, with boundary conditions $u, v|_{\partial\Omega} = 0$ excepting $u(x, 1) = 1, \forall x \in [0, 1]$. The MATLAB code is

```
n=32;m=32;dx=1/m;dy=1/n;dt=0.001;c2=10;re=100;
p(1:m,1:n)=zeros(m,n);
u(1:m+1,1:n+2)=zeros(m+1,n+2);
v(1:m+2,1:n+1)=zeros(m+2,n+1);
for k=1:10000
    t=k*dt;
    u(1,:)=zeros(1,n+2);u(m+1,:)=zeros(1,n+2);
    v(:,1)=zeros(m+2,1);v(:,n+1)=zeros(m+2,1);
    u(2:m,1)=(u(2:m,3)-6*u(2:m,2))/3;
    u(2:m,n+2)=(u(2:m,n)-6*u(2:m,n+1)+8)/3;
    v(1,2:n)=(v(3,2:n)-6*v(2,2:n))/3;
    v(m+2,2:n)=(v(m,2:n)-6*v(m+1,2:n))/3;
    un(2:m,2:n+1)=...
    u(2:m,2:n+1)-dt/dx*(p(2:m,1:n)-p(1:m-1,1:n))+...
    dt/dx^2/re*(u(3:m+1,2:n+1)+u(1:m-1,2:n+1)+...
    u(2:m,3:n+2)+u(2:m,1:n)-4*u(2:m,2:n+1));
    vn(2:m+1,2:n)=...
    v(2:m+1,2:n)-dt/dy*(p(1:m,2:n)-p(1:m,1:n-1))+...
    dt/dy^2/re*(v(2:m+1,3:n+1)+v(2:m+1,1:n-1)+...
    v(3:m+2,2:n)+v(1:m,2:n)-4*v(2:m+1,2:n));
    pn(1:m,1:n)=p(1:m,1:n)-dt*c2/dx*(u(2:m+1,2:n+1)-...
    u(1:m,2:n+1)+v(2:m+1,2:n+1)-v(2:m+1,1:n));
    erru(k)=sum(sum(abs(un(2:m,2:n+1)-u(2:m,2:n+1))));
    errv(k)=sum(sum(abs(vn(2:m+1,2:n)-v(2:m+1,2:n))));
    errp(k)=sum(sum(abs(pn(1:m,1:n)-p(1:m,1:n))));
    u(2:m,2:n+1)=un(2:m,2:n+1);v(2:m+1,2:n)=vn(2:m+1,2:n);
    p(1:m,1:n)=pn(1:m,1:n);
    display([t erru(k) errv(k) errp(k)]);
end;
x=linspace(0,1,m+1);y=linspace(0,1,n+1);
quiver(x(2:m+1)-dx/2,y(2:n+1)-dy/2,(u(1:m,2:n+1)+...
u(2:m+1,2:n+1))'/2,(v(2:m+1,1:n)+v(2:m+1,2:n+1))'/2);
pause;
k=1:10000;plot(k,erru,'-',k,errv,'--',k,errp,'.');
```

As result we obtain the velocities field (Figure 5.22) and the evolution of the errors of u, v, p for $t \in (0, 10)$ (Figure 5.23).

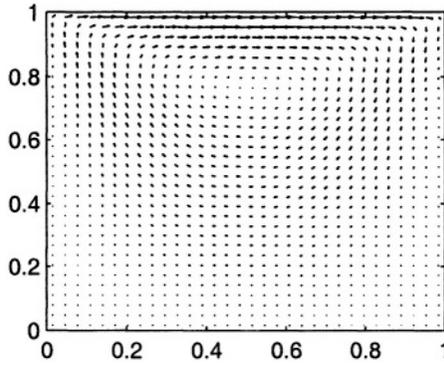


Figure 5.22. The steady solution of the Stokes problem

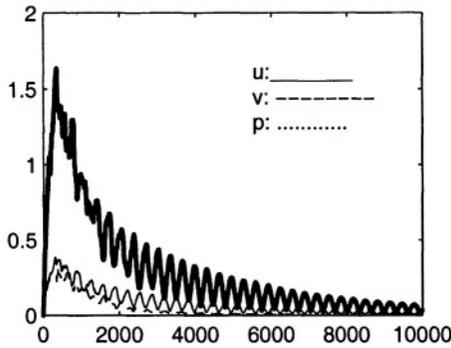


Figure 5.23. The time evolution of the errors of u, v, p

8. Compact Finite Differences

8.1 The Compact Finite Differences Method (CFDM)

For the usual finite differences methods, the accuracy could be increased by increasing the number of grid points, which complicates the obtained system and induces difficulties at the neighborhood of the boundary of the computational domain.

We could bypass these difficulties with the formulas using also the values of the derivatives at the nodes, together with formulas which link

these values, for instance,

$$u'_{j+1} + 4u'_j + u'_{j-1} - \frac{3}{h}(u_{j+1} - u_{j-1}) = 0$$

for the first derivative and

$$u''_{j+1} + 10u''_j + u''_{j-1} - \frac{12}{h^2}(u_{j+1} - 2u_j + u_{j-1}) = 0$$

for the second one. Both formulas have an accuracy of order $O(h^4)$ and associated to the equation $Lu_j = f_j$ and to the boundary conditions, lead to block tridiagonal systems with the unknowns (u_j, u'_j, u''_j) . This represents the exact solution with a higher accuracy using a smaller number of nodes.

If we know the values u_j of a single-variable function on a grid with the step size h , the values of the derivative on the grid u'_j may be approximated by combinations of the values u_j on the neighboring points. Our purpose is to obtain formulas of highest order of approximation and of the best spectral resolving power.

The above centered finite difference schemes of second order use the values u_{j-1} and u_{j+1} in order to approximate the derivative u'_j at the point x_j . High order schemes use more such values. In the spectral methods, the approximation of the first derivative is made using the values u_j on all points of the grid. The compact finite difference schemes simulate this behavior. We will present briefly the methods with compact finite differences, using the works of [81], [23], [29].

8.2 Approximation of the Derivatives

8.2.1 Approximation of the First Derivative

Let us seek an approximating formula of the form

$$\begin{aligned} & \beta u'_{j-2} + \alpha u'_{j-1} + u'_j + \alpha u'_{j+1} + \beta u'_{j+2} \\ &= c \frac{u_{j+3} - u_{j-3}}{6h} + b \frac{u_{j+2} - u_{j-2}}{4h} + a \frac{u_{j+1} - u_{j-1}}{2h}. \end{aligned} \quad (5.37)$$

The relationships between the coefficients α, β, a, b, c are obtained by matching the Taylor's series coefficients of different orders. So we have

$$\begin{aligned} u_{j+1} &= u_j + hu'_j + \frac{h^2}{2!}u''_j + \frac{h^3}{3!}u_j^{(3)} + \dots, \\ u_{j-1} &= u_j - hu'_j + \frac{h^2}{2!}u''_j - \frac{h^3}{3!}u_j^{(3)} + \dots, \\ u_{j+2} &= u_j + 2hu'_j + \frac{4h^2}{2!}u''_j + \frac{8h^3}{3!}u_j^{(3)} + \dots, \\ u_{j-2} &= u_j - 2hu'_j + \frac{4h^2}{2!}u''_j - \frac{8h^3}{3!}u_j^{(3)} + \dots, \\ u_{j+3} &= u_j + 3hu'_j + \frac{9h^2}{2!}u''_j + \frac{27h^3}{3!}u_j^{(3)} + \dots, \\ u_{j-3} &= u_j - 3hu'_j + \frac{9h^2}{2!}u''_j - \frac{27h^3}{3!}u_j^{(3)} + \dots. \end{aligned}$$

Analogously, we have

$$\begin{aligned} u'_{j+1} &= u'_j + hu''_j + \frac{h^2}{2!}u_j^{(3)} + \frac{h^3}{3!}u_j^{(4)} + \dots, \\ u'_{j-1} &= u'_j - hu''_j + \frac{h^2}{2!}u_j^{(3)} - \frac{h^3}{3!}u_j^{(4)} + \dots, \\ u'_{j+2} &= u'_j + 2hu''_j + \frac{4h^2}{2!}u_j^{(3)} + \frac{8h^3}{3!}u_j^{(4)} + \dots, \\ u'_{j-2} &= u'_j - 2hu''_j + \frac{4h^2}{2!}u_j^{(3)} - \frac{8h^3}{3!}u_j^{(4)} + \dots. \end{aligned}$$

Replacing into the above formulas we have

$$\begin{aligned} &\frac{a}{2h} \left(2hu'_j + \frac{2h^3}{3!}u_j^{(3)} + \dots \right) + \frac{b}{4h} \left(4hu'_j + \frac{16h^3}{3!}u_j^{(3)} + \dots \right) \\ &+ \frac{c}{6h} \left(6hu'_j + \frac{54h^3}{3!}u_j^{(3)} + \dots \right) = \beta \left(2u'_j + \frac{8h^2}{2!}u_j^{(3)} + \dots \right) \\ &+ \alpha \left(2u'_j + \frac{2h^2}{2!}u_j^{(3)} + \dots \right) + u'_j. \end{aligned}$$

By identification of the coefficients of u'_j we find

$$a + b + c = 2\beta + 2\alpha + 1. \quad (5.38)$$

The scheme (5.37) with the constraint (5.38) represents an approximating formula with four parameters of second order of accuracy.

By identification of the coefficients of $u_j^{(3)}$, we obtain, in addition,

$$a + 4b + 9c = 24\beta + 6\alpha. \quad (5.39)$$

Equation (5.37) with the constraints (5.38) and (5.39) represents an approximating formula with three parameters of fourth order of accuracy.

Analogously, if we add the relationship

$$a + 2^4b + 3^4c = 10(2^4\beta + \alpha)$$

we obtain a sixth order scheme with two parameters. Adding

$$a + 2^6b + 3^6c = 14(2^6\beta + \alpha)$$

we get an eighth order scheme with one parameter and, finally, by introducing

$$a + 2^8b + 3^8c = 18(2^8\beta + \alpha)$$

we are lead to a scheme of tenth order of accuracy.

If we write these formulas at all the points of the grid and if we add the special formulas for the boundary, we obtain a tri- or penta-diagonal system from which we can calculate the first order derivatives u'_j simultaneously on the whole grid.

Let us analyze in detail some particular compact schemes. From the approximating formula (5.37) with the relationships (5.38) and (5.39) we obtain a fourth order scheme, with three parameters. Choosing $\beta = 0$ we obtain tridiagonal systems to calculate the derivatives. Choosing $c = 0$ too, we get tridiagonal schemes with one parameter, of fourth order of accuracy.

From (5.38) and (5.39) we find

$$\beta = 0, a = \frac{2}{3}(\alpha + 2), b = \frac{1}{3}(4\alpha - 1), c = 0$$

so the approximating formula is

$$\alpha u'_{j-1} + u'_j + \alpha u'_{j+1} = \frac{4\alpha - 1}{12h} (u_{j+2} - u_{j-2}) + \frac{2(\alpha + 2)}{6h} (u_{j+1} - u_{j-1}) \quad (5.40)$$

with an error of the order

$$\frac{1}{30} (3\alpha - 1) h^4 u^{(5)}.$$

If $\alpha \rightarrow 0$ we obtain the well-known approximating formula of fourth order with centered differences. For $\alpha = \frac{1}{4}$ we obtain the classical scheme of Padé (which uses the values of u only at the neighboring points $x_{j\pm 1}$).

8.2.2 Approximation of the Second Order Derivative

As in the above section, we start from the approximating relationship

$$\begin{aligned} & \beta u_{j-2}'' + \alpha u_{j-1}'' + u_j'' + \alpha u_{j+1}'' + \beta u_{j+2}'' \\ &= c \frac{u_{j+3} - 2u_j + u_{j-3}}{9h^2} + b \frac{u_{j+2} - 2u_j + u_{j-2}}{4h^2} + a \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}. \end{aligned} \quad (5.44)$$

Next, based on the development in Taylor's series, we identify the coefficients and we obtain the relationships between α, β, a, b, c . So, for example, if

$$a + b + c = 2\beta + 2\alpha + 1$$

and

$$a + 4b + 9c = 12(4\beta + \alpha)$$

we get a scheme with three parameters, of fourth order.

Choosing $\beta = 0$ we obtain a tridiagonal scheme, while choosing $c = 0$ we get a five points scheme with one parameter,

$$\beta = 0, a = \frac{4}{3}(1 - \alpha), b = \frac{1}{3}(10\alpha - 1), c = 0.$$

The truncation error is of the order

$$-\frac{4}{6!}(11\alpha - 2)h^4 u^{(6)}.$$

For $\alpha \rightarrow 0$ we have the classical fourth order centered differences scheme. For $\alpha = \frac{1}{10}$ we obtain a three-points and fourth order scheme and for $\alpha = \frac{2}{11}$ the dominant term of the error vanishes and we get a sixth order scheme.

Obviously, in order to increase the order of the scheme to ten we may impose other conditions.

Similar relationships may be used for the high order approximation of the high derivatives.

In this case too, we need special formulas for computing the derivatives at the boundary points and their neighbors, in the cases of bounded domains. For example, at the left boundary point x_1 we impose an approximation formula

$$u_1'' + \alpha u_2'' = \frac{1}{h^2} (au_1 + bu_2 + cu_3 + du_4 + eu_5).$$

In order to obtain a third order of accuracy, the coefficients become

$$a = \frac{11\alpha + 35}{12}, b = -\frac{5\alpha + 26}{3}, c = \frac{\alpha + 19}{2}$$

$$d = \frac{\alpha - 14}{3}, e = \frac{11 - \alpha}{12}$$

with an error of the order

$$\frac{\alpha - 10}{12} h^3 u^{(5)}.$$

Choosing $\alpha = 10$, we obtain a fourth order scheme.

8.3 Fourier Analysis of the Errors

Let us now consider a periodic function u on the interval $[0, L]$, that is $u_{N+1} = u_1$ and $h = \frac{L}{N}$. Therefore

$$u(x) = \sum_{k=-N/2}^{k=N/2} \hat{u}_k e^{\frac{2\pi i k x}{L}}$$

where $\hat{u}_{\pm k}$ are complex conjugated and \hat{u}_0 is real. For facility, we introduce the scaled wave number $w_k = \frac{2\pi k h}{L} = \frac{2\pi k}{N}$ and the scaled coordinate $s = \frac{x}{h}$. So, the Fourier modes become $e^{i w_k s}$ and the scaled wave numbers $w_k \in [0, \pi]$.

The first derivative of u with respect to s generates a function with Fourier coefficients

$$\hat{u}'_k = i w_k \hat{u}_k$$

The differentiation error for the formulas in the above section may be evaluated by comparison between the derivative coefficients, from those formulas, and the exact coefficients.

For example, for the second order centered finite difference,

$$u'(s) = \sum_{k=-N/2}^{N/2} \hat{u}_k \frac{e^{i w_k (s+1)} - e^{i w_k (s-1)}}{2} = \sum_{k=-N/2}^{N/2} \hat{u}_k e^{i w_k s} i \sin w_k,$$

so the calculated Fourier coefficients are

$$i w'(w_k) \hat{u}_k$$

where $w'(w_k) = \sin w_k$ are the modified scaled wave numbers (by the numerical scheme).

To every numerical scheme one assigns a particular function w' . The exact derivative corresponds to $w'(w_k) = w_k$. The interval $[2\pi/N, W]$ on which $w'(w_k)$ corresponding to the numerical scheme, approximates well (within the limit of a given tolerance), the exact derivative $w'(w_k) = w_k$ defines the set of well solved waves. The shortest wave well solved

(corresponding to the largest wave number W) depends only on the numerical scheme and not on the number N of points of the grid.

A similar calculation as above, applied to the scheme (5.37) gives a modified wave number

$$w'(w_k) = \frac{a \sin w_k + \frac{b}{2} \sin 2w_k + \frac{c}{3} \sin 3w_k}{1 + 2\alpha \cos w_k + 2\beta \cos 2w_k}.$$

Indeed, considering the particular Fourier mode $\hat{u}_k e^{iw_k s}$, it is modified by (5.37) as follows:

$$\begin{aligned} & iw'(w_k) [\beta (e^{-2iw_k} + e^{2iw_k}) + \alpha (e^{-iw_k} + e^{iw_k}) + 1] \\ &= \frac{c}{6} (e^{3iw_k} - e^{-3iw_k}) + \frac{b}{4} (e^{2iw_k} - e^{-2iw_k}) + \frac{a}{2} (e^{iw_k} - e^{-iw_k}) \end{aligned}$$

or

$$\begin{aligned} & iw'(w_k) [2\beta \cos(2w_k) + 2\alpha \cos w_k + 1] \\ &= \frac{c}{6} 2i \sin(3w_k) + \frac{b}{4} 2i \sin(2w_k) + \frac{a}{2} 2i \sin w_k \end{aligned}$$

from which we obtain $w'(w_k)$.

For example, for the centered fourth order scheme ($\alpha = 0$) we have

$$w'(w_k) = \frac{4}{3} \sin w_k - \frac{1}{6} \sin 2w_k$$

and for the compact fourth order scheme ($\alpha = \frac{1}{4}$) we have

$$w'(w_k) = \frac{\frac{3}{2} \sin w_k}{1 + \frac{1}{2} \cos w_k}.$$

The graphs of the functions w' for each case are given in Figure 5.24.

It is obvious that the compact schemes have better spectral solving qualities than classical finite difference schemes. These qualities may be improved. For example, if we impose on the relationships (5.37) the conditions (5.38) and (5.39) which ensure the fourth order of accuracy, we still dispose of three parameters. They may be calculated from the conditions

$$w'(w_a) = w_a, w'(w_b) = w_b, w'(w_c) = w_c$$

where $w_a = 2.2$, $w_b = 2.3$, $w_c = 2.4$. This method yields a better pentadiagonal scheme with seven points, with a higher spectral resolving power, as it can be seen in Figure 5.24. In this case the parameters are

$$\begin{aligned} & \alpha = 0.5771439, \beta = 0.0896406, \\ & a = 1.3025166, b = 0.99355, c = 0.03750245. \end{aligned}$$

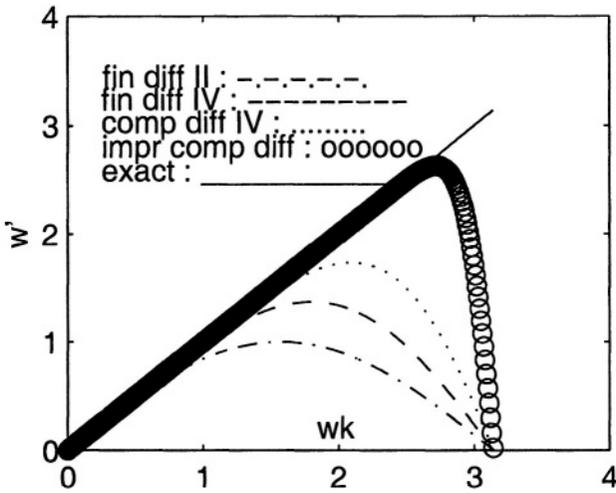


Figure 5.24. The modified wavenumbers

We remark that the spectral methods, which will be studied in a subsequent chapter, yield $w'(w_k) = w_k$ for all $w_k \neq \pi$. We have performed calculations and got similar results for higher order derivatives, too.

There is another manner to characterize the errors. Let us consider the linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

where every wave (with whatever wave number) propagates with the phasic velocity 1. By discretizing the spatial derivative we may prove that the phasic velocity for a wave with the wave number w_k is given by

$$c_f = \frac{w'(w_k)}{w}$$

and the more different this is from 1, the more inappropriate the numerical scheme represents that wave.

In multidimensional problems these phase errors also appear in an anisotropic form. Considering the equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial l} = 0$$

where l is a direction in the plane, while every wave has the phasic velocities 1 in each direction, the discretization schemes generate different

velocities depending on the wave numbers and *on the direction*. These velocities are given by

$$c_f(w, \theta) = \frac{\cos \theta w'(w_k \cos \theta) + \sin \theta w'(w_k \sin \theta)}{w}$$

where θ is the angle between the propagation direction and the Ox axis.

Figure 5.25 represents these velocities for classical centered-differen-

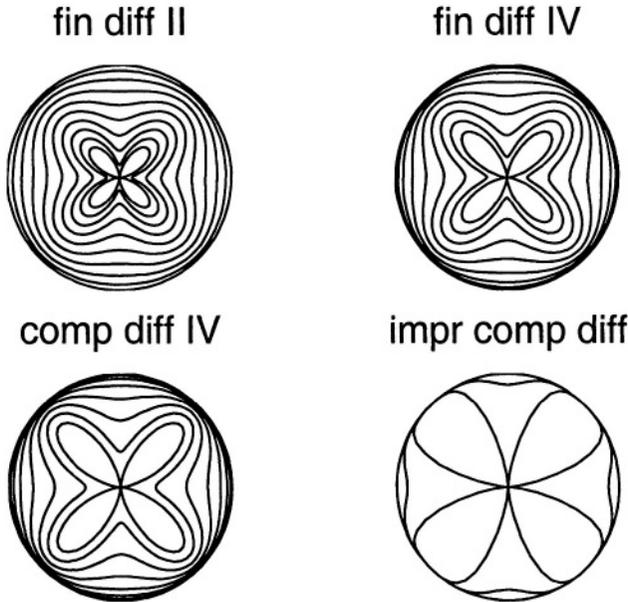


Figure 5.25. The anisotropy of the phase velocities

ces schemes of orders II and IV, for compact scheme of order IV and for compact spectrally improved scheme. Each curve corresponds to a wave number $w_k \in \{ \frac{5\pi}{50}, \frac{10\pi}{50}, \dots, \frac{50\pi}{50} \}$, the radius distance to the angle θ representing their phase velocity.

The outward curves correspond to small wave numbers and represent better solved waves (with phasic velocities closer to 1). The shorter waves, with larger wave numbers, have smaller phasic velocities, with anisotropic propagation. We remark the qualities of compact schemes over classical schemes.

8.4 Combined Compact Differences Schemes

The compact schemes were developed in many directions, in order to increase the accuracy, the resolving power and to make them easily

handled, especially in the treatment of boundary conditions. We will present such schemes, with three points, of order 6 of accuracy, with a similar accuracy (5) at the boundary points and their neighbor. Specific is the combination in the same relationship of the first and second order derivatives, which yields in applications twice- and triple-tridiagonal systems.

If the function to be approximated $u(x)$ is defined on $[0, L]$, we will use a uniform grid $0 = x_1 < x_2 < \dots < x_N < x_{N+1} = L$ with the step size $h = L/N$. If we denote u_i, u'_i, u''_i the exact values of the function and of the first and second order derivatives at the points $x_i, i = 1, \dots, N + 1$, we seek formulas of the type

$$\begin{aligned} \left(\frac{du}{dx}\right)_i + \alpha_1 \left[\left(\frac{du}{dx}\right)_{i+1} + \left(\frac{du}{dx}\right)_{i-1} \right] + \beta_1 h \left[\left(\frac{d^2u}{dx^2}\right)_{i+1} - \left(\frac{d^2u}{dx^2}\right)_{i-1} \right] + \dots \\ = \frac{\alpha_1}{h} (u_{i+1} - u_{i-1}), \\ \left(\frac{d^2u}{dx^2}\right)_i + \alpha_2 \left[\left(\frac{d^2u}{dx^2}\right)_{i+1} + \left(\frac{d^2u}{dx^2}\right)_{i-1} \right] + \frac{\beta_2}{2h} \left[\left(\frac{du}{dx}\right)_{i+1} - \left(\frac{du}{dx}\right)_{i-1} \right] + \dots \\ = \frac{\alpha_2}{h^2} (u_{i+1} - 2u_i + u_{i-1}), \end{aligned}$$

and so on. Here $\left(\frac{du}{dx}\right)_i, \dots$ represent the approximations of the corresponding derivatives and they will be calculated from the established formulas for $i = 2, 3, \dots, N$ by some systems of simultaneous equations.

In order to obtain a sixth order formula (as an example), we will build the Hermite polynomial $H_i(x)$, defined on $[x_{i-1}, x_{i+1}]$, and satisfying

$$\begin{aligned} H_i(x_{i-1}) = u_{i-1}, \quad H_i(x_i) = u_i, \quad H_i(x_{i+1}) = u_{i+1}, \\ H'_i(x_{i-1}) = u'_{i-1}, \quad H'_i(x_{i+1}) = u'_{i+1}, \\ H''_i(x_{i-1}) = u''_{i-1}, \quad H''_i(x_{i+1}) = u''_{i+1}. \end{aligned} \tag{5.45}$$

But

$$\begin{aligned} H_i(x) = H_i(x_i) + H'_i(x_i) (x - x_i) + \frac{H''_i(x_i)}{2!} (x - x_i)^2 + \frac{H_i^{(3)}(x_i)}{3!} (x - x_i)^3 \\ + \frac{H_i^{(4)}(x_i)}{4!} (x - x_i)^4 + \frac{H_i^{(5)}(x_i)}{5!} (x - x_i)^5 + \frac{H_i^{(6)}(x_i)}{6!} (x - x_i)^6. \end{aligned} \tag{5.46}$$

The seven coefficients from the above relationship may be calculated from the conditions (5.45)

$$H'_i(x_i) = \frac{15}{16h} (u_{i+1} - u_{i-1}) - \frac{7}{16} (u'_{i+1} + u'_{i-1}) + \frac{h}{16} (u''_{i+1} - u''_{i-1}),$$

$$H_i''(x_i) = \frac{3}{h^2} (u_{i+1} - 2u_i + u_{i-1}) - \frac{9}{8h} (u'_{i+1} - u'_{i-1}) + \frac{1}{8} (u''_{i+1} + u''_{i-1}), \tag{5.47}$$

and so on. Approximating the derivatives of u by the derivatives of H_i locally, in the neighborhood of x_i ,

$$u^{(k)}(x_i) \simeq H_i^{(k)}(x_i) \tag{5.48}$$

and substituting it into the relationships (5.47) we find

$$\frac{7}{16} (u'_{i+1} + u'_{i-1}) + u'_i - \frac{h}{16} (u''_{i+1} - u''_{i-1}) = \frac{15}{16h} (u_{i+1} - u_{i-1}) + R_1$$

and

$$\frac{9}{8h} (u'_{i+1} - u'_{i-1}) - \frac{1}{8} (u''_{i+1} + u''_{i-1}) + u''_i = \frac{3}{h^2} (u_{i+1} - 2u_i + u_{i-1}) + R_2. \tag{5.49}$$

If we neglect now the rests R_1 and R_2 (the truncation errors) we obtain the approximating formulas. Therefore

$$\alpha_1 = \frac{7}{16}, \beta_1 = -\frac{1}{16}, a_1 = \frac{15}{8},$$

$$\alpha_2 = -\frac{1}{8}, \beta_2 = \frac{9}{4}, a_2 = 3.$$

We remark that the rests are

$$R_1 \simeq \frac{1}{7!} h^6 u^{(7)}, R_2 \simeq \frac{1}{8!} h^6 u^{(8)}.$$

The Fourier analysis of the errors gives for the modified wave number

$$w'(w) = \frac{9(4 + \cos w) \sin w}{24 + 20 \cos w + \cos 2w}$$

which indicates a much increased resolving power vs. the non-combined schemes. Also, in the multidimensional case

$$c_f(w, \theta) = \frac{\cos \theta w'(w \cos \theta) + \sin \theta w'(w \sin \theta)}{w}$$

indicates a much decreased anisotropy over the non-combined schemes.

Let us take an example of this type of discretization. We consider the problem

$$a_2(x) \frac{d^2 u}{dx^2} + a_1(x) \frac{du}{dx} + a_0(x) u = s(x), x \in [0, L],$$

$$d_1(x) \frac{du}{dx} + d_0(x) u(x) = c(x), x = 0, x = L.$$

Using the above discretization we have

$$\begin{aligned}
 a_{2i} \left(\frac{d^2 u}{dx^2} \right)_i + a_{1i} \left(\frac{\partial u}{\partial x} \right)_i + a_{0i} u_i &= s_i, \quad i = 1, \dots, N + 1, \\
 d_{11} \left(\frac{du}{dx} \right)_1 + d_{01} u_1 &= c_1, \quad d_{1,N+1} \left(\frac{du}{dx} \right)_{N+1} + d_{0,N+1} u_{N+1} = c_{N+1},
 \end{aligned}
 \tag{5.50}$$

where, as above, the index i means the value of that function at x_i .

We remark that at every interior point $i = 2, \dots, N$ from $[0, L]$ we have three relationships: one of (5.50) and two of (5.49), relating the three unknowns for that point: u_i , $\left(\frac{\partial u}{\partial x}\right)_i$ and $\left(\frac{d^2 u}{dx^2}\right)_i$.

But, at each boundary point $i = 1$, respectively $i = N + 1$, we have only two relationships coming from (5.50). In order to solve the above system, we need one more relationship for each boundary point.

Let's now consider a fifth degree polynomial,

$$P(x) = P_0 + P_1 x + P_2 x^2 + P_3 x^3 + P_4 x^4 + P_5 x^5.$$

At the boundary point x_1 we impose that

$$\begin{aligned}
 P(x_1) &= u_1, \quad P(x_2) = u_2, \quad P(x_3) = u_3, \\
 P'(x_1) &= u'_1, \quad P'(x_2) = u'_2, \quad P''(x_2) = u''_2.
 \end{aligned}$$

If we now calculate the coefficients of P_k and we use the series developments of u in the neighborhood of x_2 , we find

$$14u'_1 + 16u'_2 + 2hu''_1 - 4hu''_2 + \frac{1}{h} (31u_1 - 32u_2 + u_3) = \frac{h^5}{90} u_2^{(6)} + O(h^6).
 \tag{5.51}$$

Neglecting the rest of the right-hand side, we obtain the needed formula.

Similarly, for the boundary point x_{N+1} we find

$$14u'_{N+1} + 16u'_N - 2hu''_{N+1} + 4hu''_N - \frac{1}{h} (31u_{N+1} - 32u_N + u_{N-1}) = 0.
 \tag{5.52}$$

If we add the equations (5.51) and (5.52) to the above system, we finally obtain a system of $3(N + 1)$ equations with $3(N + 1)$ unknowns: the values of u and of its derivatives of first and second order at all (interior and boundary) points. This system has a triple tridiagonal matrix and could be solved by special techniques for sparse systems.

The sixth-order accuracy in the interior and the fifth-order near the boundary make this method very efficient. For a similar error, the needed number of grid points over the centered second order finite difference method is much smaller: 18 nodes vs. 9400 in the case of an effective example (see [23]).

8.5 Supercompact Difference Schemes

For the classical finite difference schemes it is difficult to increase the order of accuracy; this can be performed only by increasing the number of nodes.

The compact schemes behave well, i.e., we may obtain a better accuracy with a small number of nodes. However, formulas with an “a priori” degree of accuracy are difficult to obtain.

The combined compact schemes give us a solution to this problem. By coupling the first and second derivatives we can increase the accuracy while maintaining the small number of grid points.

We can extend this idea and so the supercompact schemes are set up. With these schemes a needed accuracy (as high as we need) using only three points in the pattern may be obtained.

We will present (without calculations), after [29], this type of schemes. Let there be N -dimensional vectors

$$U = (u^{(1)}, u^{(3)}, \dots, u^{(2N-1)})^T,$$

$$e_1 = (1, 0, \dots, 0)^T$$

and the $N \times N$ matrices

$$A = \begin{pmatrix} \frac{1}{1!} & \frac{1}{3!} & \frac{1}{5!} & \dots & \frac{1}{(2N-1)!} \\ 0 & \frac{1}{2!} & \frac{1}{4!} & \dots & \frac{1}{(2(N-1))!} \\ 0 & 0 & \frac{1}{2!} & \dots & \frac{1}{(2(N-2))!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{2!} \end{pmatrix},$$

$$L = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

where $u^{(k)} = h^k \frac{d^k u}{dx^k}$. The simplest supercompact scheme to approximate the odd derivatives is

$$-\frac{1}{2}LU_{i-1} + (L + A)U_i - \frac{1}{2}LU_{i+1} = \delta_x^0 u_i e_1$$

where

$$\delta_x^0 = \frac{\delta_x^+ + \delta_x^-}{2}, \delta_x^\pm u_i = \mp (u_i - u_{i\pm 1}).$$

If we have u_i , we may calculate all $u_i^{(2k-1)}$ for $k = 1, 2, \dots, N$ and next, $u_i^{(2k-1)}/h^{2k-1}$ approximate the corresponding derivatives $\frac{d^{2k-1}u}{dx^{2k-1}}$ by an accuracy of order $2(N - k + 1)$.

Similarly, for even derivatives, we define the vector

$$S = \left(u^{(2)}, u^{(4)}, \dots, u^{(2N)} \right)^T$$

and the matrix

$$B = \begin{pmatrix} \frac{1}{2!} & \frac{1}{4!} & \frac{1}{6!} & \dots & \frac{1}{(2N)!} \\ 0 & \frac{1}{2!} & \frac{1}{4!} & \dots & \frac{1}{(2(N-1))!} \\ 0 & 0 & \frac{1}{2!} & \dots & \frac{1}{(2(N-2))!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{2!} \end{pmatrix}.$$

The relationship to approximate the even derivatives is

$$-\frac{1}{2}LS_{i-1} + (L + B)S_i - \frac{1}{2}LS_{i+1} = \frac{1}{2}\delta_x^2 u_i e_1$$

where $\delta_x^2 = \delta_x^+ \delta_x^-$. Next $u^{(2k)}/h^{2k}$ approximates $\frac{d^{2k}u}{dx^{2k}}$ by an accuracy of order $2(N - k + 1)$.

Here we can choose N as high as we need but, moreover, for the same order of accuracy, the supercompact schemes behave better. For example, for $N = 3$ which yields an approximation of the first derivative of sixth order, the classical centered difference scheme of the same order

$$U_i = \frac{1}{60} [45(u_{i+1} - u_{i-1}) - 9(u_{i+2} - u_{i-2}) + u_{i+3} - u_{i-3}]$$

where $U_i/h \simeq \frac{du}{dx}$, has a truncation error

$$2160 \frac{h^6}{7!} u^{(7)}.$$

The sixth order compact difference scheme

$$\frac{1}{5}U_{i+1} + \frac{3}{5}U_i + \frac{1}{5}U_{i-1} = \frac{14}{15}\delta_x^0 u_i + \frac{1}{30}\delta_x^0 (u_{i+1} - u_{i-1})$$

has the truncation error

$$\frac{12}{5} \frac{h^6}{7!} u^{(7)}$$

and the sixth order supercompact scheme has the truncation error

$$\frac{h^6}{7!} u^{(7)}.$$

Also, the resolving power and the anisotropy are better than those of the same order compact schemes.

9. Coordinate Transformation

In some particular cases, the physical domain of the fluid may be covered by a rectangular grid. Such a case is, as an example, the driven rectangular cavity, where the boundary may be depicted by some grid points lying exactly on it.

In other cases, as of the fluid flow past a cylinder, rectangular grids yield difficulties in the treatment of the boundary. The grid points are inside or outside the cylinder and only by exception do they lie on the boundary. Consequently, we must modify the grid points in the neighborhood of the boundary and this yields computing difficulties. In the case of the cylinder or other such bodies the problem may be solved using polar (or other kinds of) coordinates, in order to transform the computing domain also into a rectangle. Of course, the price is a change of the equation envisaged to be numerically solved. For example, the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

becomes, in polar coordinates

$$\frac{\partial^2 u}{\partial \rho^2} + \frac{1}{\rho^2} \frac{\partial^2 u}{\partial \varphi^2} + \frac{1}{\rho} \frac{\partial u}{\partial \rho} = 0.$$

In many cases, the advantage of working on a rectangular computational domain, with a uniform rectangular grid, is compensatory to the more complicated form of the equation. The problem is to find the coordinate transformation which maps the physical domain into the needed computational domain such that the uniform rectangular grid in the computational domain corresponds to a *non-uniform curvilinear grid* in the physical domain. The advantage is not only the discretization of the boundary of the physical domain. Due to specific conditions, the characteristics of the fluid flow may have large variations in some regions in the physical space. In these regions a refinement of the grid should be very useful, as it yields an increased accuracy, without a supplementary computing effort; see Figure 5.26 which presents the grid transformation in a neighborhood of a body in the boundary layer problem, for example.

Let us see now how we can transform the grids of the physical domain into some rectangular grids in the computational domain, after [4] and [155]. We will consider only the case of two-dimensional domains, but such formulas (more complex) exist also for the tridimensional cases.

Let us transform the variables (x, y, t) from the physical space into (ξ, η, τ) in the computational domain, by the relationships

$$\xi = \xi(x, y, t), \quad \eta = \eta(x, y, t), \quad \tau = \tau(t)$$

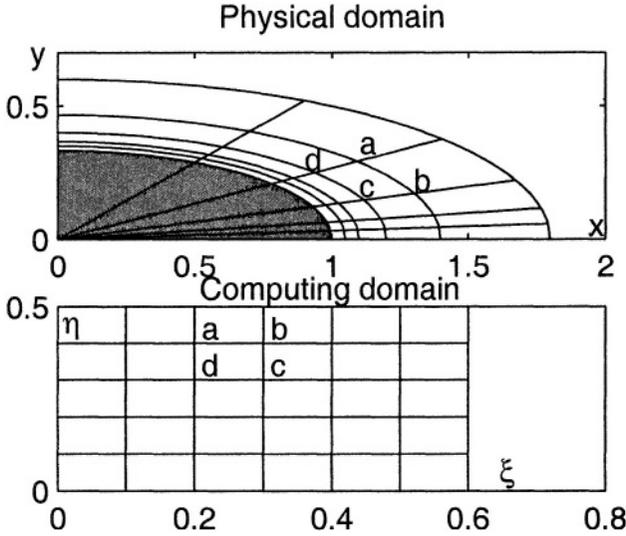


Figure 5.26. Curvilinear non-uniform grid

where often $\tau = t$. The derivatives in the partial differential equation are transformed by the formulas

$$\begin{aligned} \frac{\partial}{\partial x} &= \frac{\partial}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta} \frac{\partial \eta}{\partial x}, \\ \frac{\partial}{\partial y} &= \frac{\partial}{\partial \xi} \frac{\partial \xi}{\partial y} + \frac{\partial}{\partial \eta} \frac{\partial \eta}{\partial y}, \\ \frac{\partial}{\partial t} &= \frac{\partial}{\partial \xi} \frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta} \frac{\partial \eta}{\partial t} + \frac{\partial}{\partial \tau} \frac{d\tau}{dt}, \end{aligned} \tag{5.53}$$

and for the second order derivatives we have

$$\begin{aligned} \frac{\partial^2}{\partial x^2} &= \frac{\partial}{\partial \xi} \frac{\partial^2 \xi}{\partial x^2} + \frac{\partial}{\partial \eta} \frac{\partial^2 \eta}{\partial x^2} \\ &+ \frac{\partial^2}{\partial \xi^2} \left(\frac{\partial \xi}{\partial x} \right)^2 + \frac{\partial^2}{\partial \eta^2} \left(\frac{\partial \eta}{\partial x} \right)^2 + 2 \frac{\partial^2}{\partial \xi \partial \eta} \left(\frac{\partial \xi}{\partial x} \right) \left(\frac{\partial \eta}{\partial x} \right), \\ \frac{\partial^2}{\partial y^2} &= \frac{\partial}{\partial \xi} \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial}{\partial \eta} \frac{\partial^2 \eta}{\partial y^2} \\ &+ \frac{\partial^2}{\partial \xi^2} \left(\frac{\partial \xi}{\partial y} \right)^2 + \frac{\partial^2}{\partial \eta^2} \left(\frac{\partial \eta}{\partial y} \right)^2 + 2 \frac{\partial^2}{\partial \xi \partial \eta} \left(\frac{\partial \xi}{\partial y} \right) \left(\frac{\partial \eta}{\partial y} \right), \end{aligned}$$

$$\begin{aligned} \frac{\partial^2}{\partial x \partial y} &= \frac{\partial}{\partial \xi} \frac{\partial^2 \xi}{\partial x \partial y} + \frac{\partial}{\partial \eta} \frac{\partial^2 \eta}{\partial x \partial y} + \frac{\partial^2}{\partial \xi^2} \left(\frac{\partial \xi}{\partial x} \right) \left(\frac{\partial \xi}{\partial y} \right) \\ &+ \frac{\partial^2}{\partial \eta^2} \left(\frac{\partial \eta}{\partial x} \right) \left(\frac{\partial \eta}{\partial y} \right) + \frac{\partial^2}{\partial \xi \partial \eta} \left[\left(\frac{\partial \xi}{\partial x} \right) \left(\frac{\partial \eta}{\partial y} \right) + \left(\frac{\partial \xi}{\partial y} \right) \left(\frac{\partial \eta}{\partial x} \right) \right]. \end{aligned}$$

For example, in the case of the Laplace equation, verified by the velocities potential of an inviscid, steady, irrotational, incompressible fluid flow, i.e.,

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0,$$

we obtain, through the new coordinates (in the computational domain) the equation

$$\begin{aligned} \frac{\partial^2 \Phi}{\partial \xi^2} \left[\left(\frac{\partial \xi}{\partial x} \right)^2 + \left(\frac{\partial \xi}{\partial y} \right)^2 \right] &+ \frac{\partial^2 \Phi}{\partial \eta^2} \left[\left(\frac{\partial \eta}{\partial x} \right)^2 + \left(\frac{\partial \eta}{\partial y} \right)^2 \right] \\ &+ 2 \frac{\partial^2 \Phi}{\partial \xi \partial \eta} \left[\left(\frac{\partial \xi}{\partial x} \right) \left(\frac{\partial \eta}{\partial x} \right) + \left(\frac{\partial \xi}{\partial y} \right) \left(\frac{\partial \eta}{\partial y} \right) \right] \\ &+ \frac{\partial \Phi}{\partial \xi} \left(\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} \right) + \frac{\partial \Phi}{\partial \eta} \left(\frac{\partial^2 \eta}{\partial x^2} + \frac{\partial^2 \eta}{\partial y^2} \right) = 0. \end{aligned}$$

This equation will be discretized by a uniform grid in the computational domain. The corresponding algebraic system will be solved and next, using the inverse transformation, we will obtain the values of the potential Φ (and of the velocities too) in the physical domain. For this purpose, the coordinate transformation must be precisely given.

In the coordinate transformation formulas, the terms describing the geometry of the grid, like $\frac{\partial \xi}{\partial x}$ and others, are called *metrics*. If the coordinates transformation is given analytically, we may obtain formulas for these metrics. But in many applications, the coordinate transformations are given numerically and then the metrics are computed by finite differences.

We remark that it is more convenient to work with the inverse transformations

$$x = x(\xi, \eta, \tau), \quad y = y(\xi, \eta, \tau), \quad t = t(\tau)$$

because all computations are made in the computational domain, on uniform rectangular grids. For this, starting with a dependent variable, like the horizontal component of the velocity $u = u(x, y)$ in the steady

case, from the system

$$\frac{\partial u}{\partial \xi} = \frac{\partial u}{\partial x} \frac{\partial x}{\partial \xi} + \frac{\partial u}{\partial y} \frac{\partial y}{\partial \xi},$$

$$\frac{\partial u}{\partial \eta} = \frac{\partial u}{\partial x} \frac{\partial x}{\partial \eta} + \frac{\partial u}{\partial y} \frac{\partial y}{\partial \eta},$$

we obtain by direct solution,

$$\frac{\partial u}{\partial x} = \frac{1}{J} \left[\frac{\partial u}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial u}{\partial \eta} \frac{\partial y}{\partial \xi} \right],$$

$$\frac{\partial u}{\partial y} = \frac{1}{J} \left[\frac{\partial u}{\partial \eta} \frac{\partial x}{\partial \xi} - \frac{\partial u}{\partial \xi} \frac{\partial x}{\partial \eta} \right],$$

where

$$J \equiv \frac{\partial(x, y)}{\partial(\xi, \eta)} = \frac{\partial x}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial x}{\partial \eta} \frac{\partial y}{\partial \xi}.$$

Similar formulas may be also obtained for the second order derivatives.

In the following sections we will study three types of grid transformations.

9.1 Coordinate Stretching

In some cases, in the study of the boundary layer for example, the essential phenomenon happens in a little region, near the surface of the body. It is a good idea to refine the coordinate lines in this region, while maintaining a uniform rectangular grid in the computational domain.

For example, let us consider the viscous fluid flow over a flat surface $y = 0$ and we wish to refine the coordinate lines in a neighborhood of this surface. The simplest coordinate transformation for this is

$$\xi = x, \quad \eta = \ln(y + 1)$$

whose inverse transformation is

$$x = \xi, \quad y = e^\eta - 1.$$

So we obtain the inverse metrics

$$\frac{\partial x}{\partial \xi} = 1, \quad \frac{\partial x}{\partial \eta} = 0,$$

$$\frac{\partial y}{\partial \xi} = 0, \quad \frac{\partial y}{\partial \eta} = e^\eta.$$

For example, the continuity equation for the stationary bidimensional flow,

$$\frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} = 0$$

and using the above coordinates transformation formulas this equation becomes

$$\frac{1}{J} \left[\frac{\partial(\rho u)}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial(\rho u)}{\partial \eta} \frac{\partial y}{\partial \xi} \right] + \frac{1}{J} \left[\frac{\partial(\rho v)}{\partial \eta} \frac{\partial x}{\partial \xi} - \frac{\partial(\rho v)}{\partial \xi} \frac{\partial x}{\partial \eta} \right] = 0$$

or, by replacing the corresponding metrics,

$$e^\eta \frac{\partial(\rho u)}{\partial \xi} + \frac{\partial(\rho v)}{\partial \eta} = 0$$

which represents the continuity equation in the computational domain.

A more complex formula is

$$x = \frac{\xi_0}{A} [\sinh((\xi - x_0)\beta_x + A)]$$

where

$$A = \sinh(\beta_x x_0),$$

$$x_0 = \frac{1}{2\beta_x} \ln \left[\frac{1 + (e^{\beta_x} - 1)\xi_0}{1 + (e^{-\beta_x} - 1)\xi_0} \right].$$

Here ξ_0 is the point of the computational domain where the maximum clustering is to occur and β_x controls the degree of clustering (larger values of β_x provide a finer grid around ξ_0). By a similar formula, we may obtain simultaneous refinements in both directions x and y .

9.2 Boundary-Fitted Coordinate Systems

One of the great advantages of the coordinate transformations is the possibility to identify some coordinate lines with the boundaries of the physical domain. For example, if the physical domain is a rectangle, bounded (as an upper wall) by a curvilinear boundary of equation $y_s = f(x)$, then the transformation

$$\xi = x, \quad \eta = \frac{y}{f(x)}$$

will lead to a rectangular grid in the computational domain. The curvilinear boundary now coincides with the coordinate line $\eta = 1$.

Such transformations may be performed even in more complex cases. The domain around an airfoil, for example, may be transformed in a

rectangle in which the surface of the airfoil is one of the sides. It is necessary to find functions $x = x(\xi, \eta)$, $y = y(\xi, \eta)$ defined on a rectangle, in the computational domain, if we know those values on the rectangle's boundary. The transformation may be defined inside the rectangle by solving a Dirichlet problem for the Laplace equation (the simplest equation for which we have a maximum principle).

We must remark that this problem is not close to the physics of the fluid flow. It is used only to choose a suitable grid for our physical domain and, next, by the computed metrics, the equations of the physical model may be transformed and then discretized.

Let us present here an example of the automatic generation of the grid suitable for a given domain, using the work [143].

As above, the Laplace equation for x and y is transformed into

$$\begin{aligned} ax_{\xi\xi} - 2bx_{\xi\eta} + cx_{\eta\eta} &= 0, \\ ay_{\xi\xi} - 2by_{\xi\eta} + cy_{\eta\eta} &= 0, \end{aligned}$$

where

$$\begin{aligned} a &= x_{\eta}^2 + y_{\eta}^2, \\ b &= x_{\xi}x_{\eta} + y_{\xi}y_{\eta}, \\ c &= x_{\xi}^2 + y_{\xi}^2. \end{aligned}$$

The functions $x(\xi, \eta)$, $y(\xi, \eta)$ are effectively given on the boundary of the computational domain, corresponding to the boundary of the physical domain (or to some cuts in this domain). So, by discretization and solving these Dirichlet problems we find the discrete forms of the metrics, used next to transform the physical equations.

Therefore, we choose the grid points on the physical boundary, maybe closer in "difficult" regions, corresponding to uniformly spaced grid points on the boundary of the computational domain. By finite difference discretization of the derivatives, we have

$$\begin{aligned} a_{i,j} &= \left(\frac{x_{i,j+1} - x_{i,j-1}}{2h_1} \right)^2 + \left(\frac{y_{i,j+1} - y_{i,j-1}}{2h_2} \right)^2, \\ b_{i,j} &= \frac{x_{i+1,j} - x_{i-1,j}}{2h} \frac{x_{i,j+1} - x_{i,j-1}}{2h} + \frac{y_{i+1,j} - y_{i-1,j}}{2h} \frac{y_{i,j+1} - y_{i,j-1}}{2h}, \\ c_{i,j} &= \left(\frac{x_{i+1,j} - x_{i-1,j}}{2h_1} \right)^2 + \left(\frac{y_{i+1,j} - y_{i-1,j}}{2h_2} \right)^2. \end{aligned}$$

The equation will be discretized and then it will be solved by iterations. Starting from an initial approximation of the solutions $x_{i,j}^{(0)}$, $y_{i,j}^{(0)}$,

the new approximation will be calculated from

$$\begin{aligned}
 a_{i,j}^{(n)} \frac{x_{i+1,j}^{(n+1)} - 2x_{i,j}^{(n+1)} + x_{i-1,j}^{(n+1)}}{h_1^2} - 2b_{i,j}^{(n)} \frac{x_{i+1,j+1}^{(n+1)} + x_{i-1,j-1}^{(n+1)} - x_{i+1,j-1}^{(n+1)} - x_{i-1,j+1}^{(n+1)}}{4h_1h_2} \\
 + c_{i,j}^{(n)} \frac{x_{i,j+1}^{(n+1)} - 2x_{i,j}^{(n+1)} + x_{i,j-1}^{(n+1)}}{h_2^2} = 0, \\
 a_{i,j}^{(n)} \frac{y_{i+1,j}^{(n+1)} - 2y_{i,j}^{(n+1)} + y_{i-1,j}^{(n+1)}}{h_1^2} - 2b_{i,j}^{(n)} \frac{y_{i+1,j+1}^{(n+1)} + y_{i-1,j-1}^{(n+1)} - y_{i+1,j-1}^{(n+1)} - y_{i-1,j+1}^{(n+1)}}{4h_1h_2} \\
 + c_{i,j}^{(n)} \frac{y_{i,j+1}^{(n+1)} - 2y_{i,j}^{(n+1)} + y_{i,j-1}^{(n+1)}}{h_2^2} = 0.
 \end{aligned}$$

The boundary values for $i = 0, i = n + 1, j = 0, j = m + 1$ are introduced at the beginning and they do not change while iterate $n = 0, 1, \dots$. After the new iteration was calculated we recompute a, b, c at the grid points $i = 1, \dots, n, j = 1, \dots, m$ and then we pass to the next iteration. Finally, we obtain

$$x_{i,j} = x(\xi_i, \eta_j), y_{i,j} = y(\xi_i, \eta_j)$$

and the corresponding coordinate lines.

If we wish to study, for example, the inviscid, incompressible fluid flow, through a channel of variable section (see Figure 5.27), we choose the grid points on the boundary of the channel and we will transform this channel into the computing domain $(\xi, \eta) \in [0, a] \times [0, b]$, which may be covered by a uniform grid with step size h . The boundary values on the sides $\eta = 0, \eta = b$ are obtained from the grid points' coordinates chosen on the boundary of the channel and the boundary conditions on the sides $\xi = 0, \xi = a$, corresponding to the inlet or outlet, are obtained from the grid points chosen on these regions. So, we may control, to a certain extent, the density of the coordinate lines in different regions.

The above computation was not close to the physics of the phenomenon. In order to study the flow, we start from the streamlines equation

$$\Psi_{xx} + \Psi_{yy} = 0$$

which will be transformed into

$$a\Psi_{\xi\xi} - 2b\Psi_{\xi\eta} + c\Psi_{\eta\eta} = 0$$

where a, b, c are still calculated as above. The boundary conditions for Ψ , in the computational domain, will be $\Psi = 0$ on the lower horizontal wall, $\Psi = 1$ on the upper horizontal wall and $\Psi_\xi = 0$ on the vertical walls.

In our simple case these conditions are verified for the function $\Psi(\xi, \eta) = \eta$. So that, the coordinate lines $\eta = \text{const.}$ will be, in fact, streamlines. These results can be seen in Figures 5.27 and 5.28.

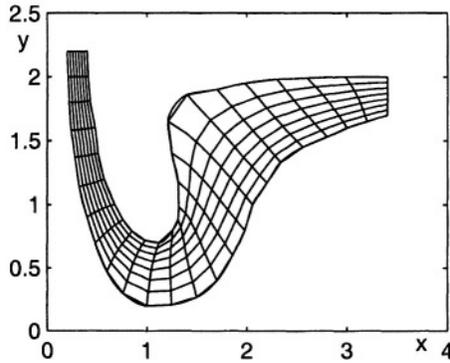


Figure 5.27. Curvilinear coordinates adapted to the physical domain

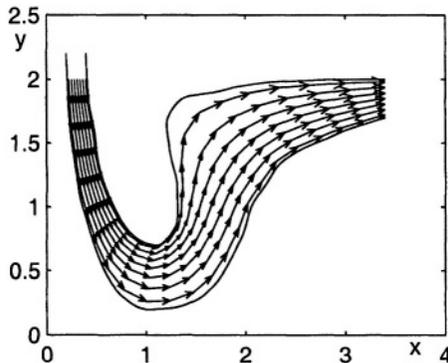


Figure 5.28. Velocity field through the channel

The MATLAB code is

```
x=zeros(24,15);y=zeros(24,15);
x(:,1)=[0.2;0.21;0.22;0.24;0.28;0.32;0.4;0.48;...
0.6;0.76;1;1.25;1.5;1.7;1.86;1.98;2.06;2.2;...
2.34;2.56;2.8;3;3.2;3.4];
x(:,15)=[0.4;0.41;0.43;0.48;0.51;0.57;0.68;0.76;...
0.85;0.98;1.12;1.26;1.31;1.32;1.31;1.25;1.21;...
1.4;1.7;1.98;2.3;2.6;2.98;3.4];
x(1,:)=linspace(0.2,0.4,15); x(24,:)=3.4*ones(1,15);
```

```

y(:,1)=[2.2;2;1.8;1.58;1.37;1.14;0.9;0.7;...
0.47;0.3;0.2;0.21;0.27;0.4;0.6;0.8;1;1.17;...
1.33;1.45;1.53;1.6;1.65;1.7];
y(:,15)=[2.2;2;1.81;1.6;1.4;1.2;1;0.88;0.79;...
0.72;0.7;0.8;0.88;0.98;1.18;1.4;1.65;1.86;1.9;...
1.95;1.98;1.99;2;2]; y(1,:)=2.2*ones(1,15);
y(24,:)=linspace(1.7,2,15);
A=ones(22,13);B=A;C=A;iter=1;
while 1
X=(A.*(x(3:24,2:14)+x(1:22,2:14))+C.*...
(x(2:23,3:15)+x(2:23,1:13))-B/2.*(x(3:24,3:15)+...
x(1:22,1:13)-x(3:24,1:13)-x(1:22,3:15)))/2./(A+C);
Y=(A.*(y(3:24,2:14)+y(1:22,2:14))+C.*...
(y(2:23,3:15)+y(2:23,1:13))-B/2.*(y(3:24,3:15)+...
y(1:22,1:13)-y(3:24,1:13)-y(1:22,3:15)))/2./(A+C);
err=max(max(abs(x(2:23,2:14)-X)))+...
max(max(abs(y(2:23,2:14)-Y)));
if rem(iter,30)==0 disp([iter err]);end;
x(2:23,2:14)=X;y(2:23,2:14)=Y;iter=iter+1;
A=((x(2:23,3:15)-x(2:23,1:13))/2/0.1).^2+...
((y(2:23,3:15)-y(2:23,1:13))/2/0.1).^2+eps;
B=(x(3:24,2:14)-x(1:22,2:14))/2/0.1.*...
(x(2:23,3:15)-x(2:23,1:13))/2/0.1+(y(3:24,2:14)-...
y(1:22,2:14))/2/0.1.*(y(2:23,3:15)-y(2:23,1:13))...
/2/0.1+eps;
C=((x(3:24,2:14)-x(1:22,2:14))/2/0.1).^2+...
((y(3:24,2:14)-y(1:22,2:14))/2/0.1).^2+eps;
if err<1.e-3 break;end; end;
ii=1:0.1:24;xi=interp1(1:24,x(:,1),ii,'cubic');
yi=interp1(1:24,y(:,1),ii,'cubic');
xii=interp1(1:24,x(:,15),ii,'cubic');
yii=interp1(1:24,y(:,15),ii,'cubic');
plot(xi,yi,'r',xii,yii,'r'); hold on;
for j=1:2:15 plot(x(:,j),y(:,j));end;
for i=1:24 plot(x(i,:),y(i,:));end;hold off; pause;
dx=x(3:24,2:14)-x(2:23,2:14);
dy=y(3:24,2:14)-y(2:23,2:14);
plot(xi,yi,xii,yii);hold on;
quiver(x(2:23,2:2:14),y(2:23,2:2:14),...
dx(:,1:2:13),dy(:,1:2:13));hold off;

```

9.3 Adaptive Grids

In many problems, either of evolution or steady state, solved by iterations, it is useful to dynamically adapt the grid according to the gradients of the calculated values. So, in the physical domain, the grid points evolve in conjunction with the solution. But in the computing domain they are fixed. In this case, the grid generation is linked to the computed solution, in contrast with what happened in the previous sections.

For example, the step sizes of the grid may be chosen by the formulas

$$\Delta x = \frac{B\Delta\xi}{1 + b\frac{\partial g}{\partial x}}, \quad \Delta y = \frac{C\Delta\eta}{1 + c\frac{\partial g}{\partial y}}$$

where g is one of the primitive variables of the fluid, like p , ρ or T . These formulas cluster the grid points in the regions with high gradients of that variable.

Now, in the transformation formulas, we should also take into account the time. In these cases, although the grid points in the computational domain are fixed, the coefficients of the form $\frac{\partial\xi}{\partial t}$ do not vanish, producing the movement of the grid points in the physical space. So, the changes of ξ , respectively η , for a fixed point (x, y) from the physical domain, are described.

If we exploit the formulas of the type

$$\left(\frac{\partial x}{\partial t}\right)_{x,y} = \left(\frac{\partial x}{\partial\xi}\right)\left(\frac{\partial\xi}{\partial t}\right)_{x,y} + \left(\frac{\partial x}{\partial\eta}\right)\left(\frac{\partial\eta}{\partial t}\right)_{x,y} + \left(\frac{\partial x}{\partial\tau}\right)\left(\frac{\partial\tau}{\partial t}\right)_{x,y}$$

where $\frac{\partial x}{\partial t} = 0$, $\frac{\partial\tau}{\partial t} = 1$ and similarly for y , those coefficients are obtained by solving the system

$$\frac{\partial\xi}{\partial t} = \frac{1}{J} \left[-\frac{\partial x}{\partial t} \frac{\partial y}{\partial\eta} + \frac{\partial y}{\partial t} \frac{\partial x}{\partial\eta} \right],$$

$$\frac{\partial\eta}{\partial t} = \frac{1}{J} \left[\frac{\partial x}{\partial t} \frac{\partial y}{\partial\xi} - \frac{\partial y}{\partial t} \frac{\partial x}{\partial\xi} \right].$$

In these formulas, at the points (ξ, η) , we approximate $\frac{\partial x}{\partial t} \approx \frac{\Delta x}{\Delta t}$, $\frac{\partial y}{\partial t} \approx \frac{\Delta y}{\Delta t}$ and $\Delta x, \Delta y$ are obtained from the formulas that govern the grid adaptation. We remark that by clustering the grid points in the regions of high gradients, such type of transformations are also flow visualization methods.

Chapter 6

FINITE ELEMENT AND BOUNDARY ELEMENT METHODS

1. Finite Element Method (FEM)

This is another method to transform a partial differential equation into a finite number of simple equations. Basically, the computational domain is divided into a finite number of subdomains — the *elements*. On each element we envisage a simple variation of the unknown functions and then the results are assembled to describe the numerical solution on the entire domain.

Let us suppose, in the one-dimensional case, that on the respective element the unknown function U has a linear variation. Then, the function could be expressed on the respective element using only its values at the ends of the elements (called *nodes*) and the distance from the computational point x to one end. For a quadratic variation we should use, in addition, the value of U at another point belonging to the element, for instance at its midpoint.

Using this representation, the derivative of U on that element is a constant while the second derivative is zero and carries no information about U . To eliminate this situation, the equations containing the second derivative are transformed into equations with the first derivatives only. The technique is called the *variational formulation* and consists of multiplication of the equation by a known function (the test function), followed by integration of the obtained equation on the respective domain and then the use of an integration by parts formula for terms containing higher order derivatives, in order to reduce the derivative order.

For example, consider the Laplace bidimensional equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0$$

where the unknown function U depends on the spatial coordinates x and y . By multiplication with the known function v followed by integration on the domain Ω we get

$$\int_{\Omega} v \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} \right) d\Omega = 0.$$

In order to reduce the derivative order, we integrate both terms by parts and so we obtain

$$\int_{\Gamma} v \left(\frac{\partial U}{\partial x} n_x + \frac{\partial U}{\partial y} n_y \right) d\Gamma - \int_{\Omega} \left(\frac{\partial U}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial U}{\partial y} \frac{\partial v}{\partial y} \right) d\Omega = 0$$

where Γ is the boundary of Ω and (n_x, n_y) is the unit outward normal vector drawn to the boundary of Ω . Therefore the derivative order of the unknown function is reduced but the values of its derivative on the boundary interfere.

Analogously, for a differential equation

$$U'' = 0$$

we obtain

$$\int_a^b U'' v dx = 0$$

and next

$$U'(b)v(b) - U'(a)v(a) - \int_a^b U' v' dx = 0.$$

From the variational form we can deduce the discrete form of the given equation. For example, in the one-dimensional case, on each element (x_1, x_2) with the nodes x_1 respectively x_2 we have the linear representation

$$U(x) = u_1 + \frac{x - x_1}{x_2 - x_1} (u_2 - u_1)$$

or

$$U(x) = u_1 \frac{x_2 - x}{x_2 - x_1} + u_2 \frac{x - x_1}{x_2 - x_1}.$$

Here the functions

$$\Phi_1 = \frac{x_2 - x}{x_2 - x_1}, \Phi_2 = \frac{x - x_1}{x_2 - x_1}$$

are the *shape functions* (they are linear, taking the value 0 at a node and the value 1 at the other one), generally chosen from a class of functions V_h , and u_1, u_2 are the *nodal values* (the values of U at the nodes).

Consequently, for each element we can write

$$U = \sum_{i=1}^n \Phi_i u_i$$

where n is the number of nodes belonging to that element. Then the derivatives are calculated through

$$\frac{dU}{dx} = \sum_{i=1}^n \frac{d\Phi_i}{dx} u_i.$$

By replacing into the variational form, decomposed now into a sum of integrals on each element, where

$$(a, b) = \bigcup (x_{i-1}, x_i)$$

and choosing a number of known test functions v from a test functions space W_h , a number equal to the number of nodes, we obtain a system which represents the discretized form of the given equation. For example, if we choose as v for each node the shape function corresponding to that node, we obtain the *Galerkin method*, but other choices are also possible. By solving this system we obtain the approximations of the values of the unknown solution at the nodes, which generate next the approximation of that solution on each element.

The finite element method is one of the most used methods for numerical solving of differential problems. It does not act directly on the differential equations; these are, firstly, set in a variational (integral) form. Next, the integrals are decomposed as sums of integrals on subdomains and the unknown functions are locally approximated by polynomials on those subdomains. This scheme leads to important advantages such as:

- a) the possibility to solve problems on domains with an arbitrary geometry and different type of boundary conditions,
- b) the possibility to use unstructured grids, the introduction or the elimination of some elements does not change the global structure of the data,
- c) the structural and flexible programming of the algorithms,
- d) a rigorous mathematical foundation.

Depending on the used variational principle, the finite element methods could be classified as the *Rayleigh–Ritz* method, the *Galerkin* method and the *least-squares* method.

The Rayleigh–Ritz method minimizes the “total potential energy”, that is the difference between the numerical and the exact solution of the problem is minimized in a certain energy norm. The algorithm leads to linear algebraic positively defined systems and it is practical especially for problems governed by self-adjoint elliptic operators.

The Galerkin method is based on the weighted residual form. If

$$\begin{aligned} Au &= f, \text{ in } \Omega, \\ Bu|_{\partial\Omega} &= 0 \end{aligned}$$

is the problem to solve, where A is a differential linear operator and B is a boundary operator, the unknown u is approximated by a linear combination of trial (basis) functions Φ_j , namely

$$u \simeq \sum_i u_i \Phi_i$$

whose coefficients u_j can be calculated from the system

$$\int_{\Omega} v_i^T (Au - f) d\Omega + \int_{\partial\Omega} \bar{v}_i^T B u d\sigma = 0.$$

Here v_i and \bar{v}_i are suitable test functions (for instance, $v_i = \bar{v}_i = \Phi_i$).

The method is applicable also for non-self-adjoint equations, in fluid dynamics for example. However, in many cases, especially for problems governed by first order equations, the method does not give the best approximation results.

The least-squares method is based on the minimization of the residuals in a least-squares sense, more precisely the method minimizes the functional

$$\int_{\Omega} (Au - f)^2 d\Omega$$

within the constraint of the boundary conditions. The approximate solution is calculated from the system

$$\int_{\Omega} (A\Phi_i)^T (Au - f) d\Omega = 0.$$

The most important advantages of this method are:

a) universality, i.e., in contrast with the classical methods where for every type of problem we should use a different type of schemes, the least-squares finite element method (LSFEM) has a unified formulation for all types of problems. For example, in the same mathematical and numerical frame, the method is able to simulate fluid dynamics problems for subsonic, transonic, supersonic or hypersonic flows.

b) efficiency, i.e., the method is suited for differential operators of the first order and leads to algebraic systems with symmetric positively defined matrices.

c) robustness, i.e., no special treatments such as artificial dissipation, staggered grids, operator-splitting, etc are necessary. The method contains the mechanism to automatically capture discontinuities or shocks.

d) optimality, i.e., the method leads to a solution with the best approximation (with an error of the same order as the interpolation error) and this error can be evaluated by an error indicator included in the form of residuals.

1.1 Flow in the Presence of a Permeable Wall

In the sequel we will consider a plane, potential, without circulation flow of an inviscid, incompressible fluid, generated by a general displacement of a profile, in the presence of an unlimited permeable wall. The fluid is assumed to be at rest at far distances [111]. The solution will be approximated by a finite element method, together with an analysis of the convergence and of the errors.

1.1.1 Variational Model Joined to the Mechanical Problem

We will suppose that the plane, unlimited, permeable wall Γ is located at a sufficiently large distance from the mobile profile C such that their working condition could be linearized. The determination of the complex potential of the considered fluid flow becomes a solution of the following boundary value problem for the uniform stream function Ψ

$$\left\{ \begin{array}{l} \Delta\Psi = 0 \quad \text{in } \Omega, \\ \Psi|_C = ny - mx - \frac{1}{2}\omega \left[(x - x_C)^2 + (y - y_C)^2 \right] + K(t), \\ \alpha_1 \frac{\partial\Psi}{\partial x} + \alpha_2 \frac{\partial\Psi}{\partial y} \Big|_{\Gamma} = l(x), \\ \lim_{|z| \rightarrow \infty} \frac{\partial\Psi}{\partial y} = \lim_{|z| \rightarrow \infty} \frac{\partial\Psi}{\partial x} = 0 \end{array} \right.$$

where Ω is the flow domain in the physical plane, the outside of the profile C bounded by the wall Γ , $n(t)$, $m(t)$, $\omega(t)$ are the components as functions of time in (x_C, y_C) , of the rototranslation of the profile C in the inviscid fluid mass, at rest at infinity (with respect to the fixed frame $Oxyz$ whose Ox axis coincides with the wall Γ) and $K(t)$ is an arbitrary function of time. The function $l(x)$ and the real constants α_1

and α_2 are related with the permeability P and the pressure p_e outside of Γ by $l(x) = \alpha_2(2 - p_e)$, $\alpha_1 P + \alpha_2 = 0$, $p > 0$; moreover, we assume $\lim_{|x| \rightarrow \infty} p_e(x) = p_\infty = 2$, this equality being in agreement with the condition on the wall Γ ($y = 0$) where $|x| \rightarrow \infty$.

But the rest condition at infinity under the acceptable hypothesis of the uniform convergence for $\lim_{|x| \rightarrow \infty} \frac{\partial \Psi}{\partial y} = 0$, $\lim_{|y| \rightarrow \infty} \frac{\partial \Psi}{\partial x} = 0$ leads also to the constancy of the limits $\lim_{|x| \rightarrow \infty} \Psi(x, y)$ and $\lim_{|y| \rightarrow \infty} \Psi(x, y)$ so that we have $\lim_{|z| \rightarrow \infty} \Psi(x, y) = \delta = \text{constant}$ (the constant δ being fixed to zero by a translation $\Psi \rightarrow \Psi + \delta$). Therefore, the above problem becomes

$$\left\{ \begin{array}{l} \Delta \Psi = 0 \quad \text{in } \Omega, \\ \Psi|_C = l_1(x, t), \\ \alpha_1 \frac{\partial \Psi}{\partial x} + \alpha_2 \frac{\partial \Psi}{\partial y} \Big|_\Gamma = l(x), \\ \lim_{|z| \rightarrow \infty} \Psi(x, y) = 0 \end{array} \right. \quad (6.1)$$

where $l_1(x, t) = ny - mx - \frac{1}{2}\omega \left[(x - x_C)^2 + (y - y_C)^2 \right] + K(t) - \delta$, while $y = K_1(x)$ at the upperside C_e of the profile C and $y = K_2(x)$ at the lowerside C_i of the same profile C ; $y = K_1(x)$ and $y = K_2(x)$, the equations of the upperside and lowerside belong to $C^2[a, b]$ and, more, $K_1^{(i)}(a) = K_2^{(i)}(a)$, $K_1^{(i)}(b) = K_2^{(i)}(b)$, $i = 1, 2$.

But the Dirichlet condition on C could be homogenized by ‘‘elevation’’, which means: being given an $\varepsilon > 0$, however small, and an $A > 0$, however large, one could introduce a function $w \in C^\infty(\Omega)$, with the support in the ‘‘half’’-disk $\{(x, y); x^2 + y^2 < A^2, y > \varepsilon\}$ and which verifies on the profile C the condition $w(x, y)|_C = l_1(x, t)$; denoting then $\Psi = u + w$, the function u satisfies the problem

$$\left\{ \begin{array}{l} -\Delta u = f \quad \text{in } \Omega \quad (f = \Delta w), \\ u|_C = 0, \\ \alpha_1 \frac{\partial u}{\partial x} + \alpha_2 \frac{\partial u}{\partial y} \Big|_\Gamma = l(x), \\ \lim_{|z| \rightarrow \infty} u(x, y) = 0. \end{array} \right. \quad (6.2)$$

In order to put the problem (6.2) into a suitable variational frame, we will introduce the space $V = \{v \in H^1(\Omega), v|_C = 0\}$ and we will define the following bilinear, skew-symmetric and continuous form on $V \times V$,

$$J(u, v) = \int_{\Omega} \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right) dx dy.$$

If the functions u and v are regular, v being zero on C , then

$$\begin{aligned} J(u, v) &= \int_{\Omega} \left[\frac{\partial}{\partial x} \left(-v \frac{\partial u}{\partial y} \right) + \frac{\partial}{\partial y} \left(v \frac{\partial u}{\partial x} \right) \right] dx dy \\ &- \int_{\Omega} \left(-v \frac{\partial^2 u}{\partial y \partial x} + v \frac{\partial^2 u}{\partial x \partial y} \right) dx dy = \int_C \left[\left(-v \frac{\partial u}{\partial y} \right) n_1 + v \frac{\partial u}{\partial x} n_2 \right] ds_1 \\ &+ \int_{\Gamma} v \frac{\partial u}{\partial x} ds_2 = \int_{-\infty}^{+\infty} v(x, 0) \frac{\partial u}{\partial x}(x, 0) dx = \left\langle \frac{\partial u}{\partial \tau}, v \right\rangle \end{aligned}$$

where $\frac{\partial}{\partial \tau}$ denotes the tangential derivative on Γ while $\mathbf{n}(n_1, n_2)$ is the unit outward normal at the contour C . The application $(u, v) \rightarrow \left\langle \frac{\partial u}{\partial \tau}, v \right\rangle$ is extended by density into the space $H^1(\Omega) \times V$.

Let us set now

$$a(u, v) = \alpha_2 \int_{\Omega} \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right) dx dy + \alpha_1 J(u, v)$$

and suppose $\alpha_2 > 0$. The problem (6.2) becomes:

Find a function $u \in V$ such that

$$a(u, v) = \alpha_2 (f, v) + \langle l, v \rangle, \quad \forall v \in V \tag{6.3}$$

where (\dots) denotes the inner product from $L^2(\Omega)$ while $\langle l, \cdot \rangle$ is the above considered linear and continuous functional on V which coincides with an inner product on $L^2(\Gamma)$ if $l \in L^2(\Gamma)$.

If we remark that $(f, v) = (\Delta w, v) = -\frac{1}{\alpha_2} a(w, v)$, we also have

$$\Psi = u + w, \quad u \in V, \tag{6.4}$$

$$a(\Psi, v) = \langle l, v \rangle, \quad \forall v \in V$$

and one proves

THEOREM 6.1. *If $l \in L^2(\Gamma)$ is given, there exists a unique function u , solution of the problem (6.3). The function $\Psi = u + w$ satisfies (6.4), it is unique and independent of the “elevation” w .*

1.1.2 Numerical Approximation of the Solution

In order to construct a numerical approximation of the unique solution of the problem (6.3) or (6.4) we must, first, replace the unbounded domain Ω by a bounded working domain. Therefore, let Ω_A be the bounded domain joined to the original domain Ω and defined by

$$\Omega_A = \Omega \cap \{(x, y); x^2 + y^2 < A^2\}$$

where the parameter $A > 0$ (which will tend to $+\infty$), should be chosen such that the contour C with its inside belongs to Ω_A .

Then we will have the approximated problem

$$\left\{ \begin{array}{l} \Delta \Psi_A = 0 \quad \text{in } \Omega_A, \\ \Psi_A|_C = l_1(x, t), \\ \Psi_A|_{\delta A} = 0, \\ \alpha_1 \frac{\partial \Psi}{\partial x} + \alpha_2 \frac{\partial \Psi}{\partial y} \Big|_{\Gamma_A} = l_0(x) \end{array} \right. \quad (6.5)$$

where Γ_A is the restriction of Γ onto the interval $(-A, A)$ and δA is the circumference from the positive half-plane $x^2 + y^2 = A^2$, defined by

$$\delta A = \{(x, y); x^2 + y^2 = A^2, y > 0\}.$$

We will show the existence and the uniqueness of the solution Ψ_A of the above problem (6.5) together with the fact that, under some hypotheses, the function Ψ_A converges (in a sense that will be made precise) towards the exact solution Ψ of the problem (6.1).

Following the same way as in the previous sub-section, we will introduce an “elevation” $w_A \equiv w \in C^\infty(\Omega_A)$ with the support belonging to the half-disk

$$\{(x, y); x^2 + y^2 < A^2, y > \varepsilon > 0\}$$

and verifying on the contour C the condition $w|_C = l_1(x, t)$.

Since $w|_{\delta A} = 0$ and $\Psi_A|_{\delta A} = 0$, if we set $\Psi_A = u_A + w$ we will be led to the problem

$$\begin{cases} -\Delta u_A = f_A & \text{in } \Omega_A \quad (f_A = \Delta w), \\ u_A|_{C \cup \delta A} = 0, \\ \alpha_1 \frac{\partial u_A}{\partial x} + \alpha_2 \frac{\partial u_A}{\partial y} \Big|_{\Gamma_A} = l_0(x). \end{cases} \tag{6.6}$$

Introducing now the space

$$V_A = \{v \in H^1(\Omega_A), v|_{\delta A \cup C} = 0\},$$

we obtain, for the problem (6.6), a variational formulation similar to that from the previous sub-section and we could state

THEOREM 6.2. *Under the hypotheses from the previous theorem, the problem (6.6) has a unique solution $u_A \in V_A$. The function $\Psi_A = u_A + w$ is unique, independent of the “elevation” w and it verifies the problem (6.5). Moreover, if \tilde{u}_A is the extension of u_A by zero onto Ω , the function $\tilde{\varphi}_A = \tilde{u}_A + w$ converges strongly towards the function Ψ in $H^1(\Omega)$.*

We will use now a finite element method for the effective approximation of the solution Ψ_A . Let $(\tau_K)_{K>0}$ be a regular sequence of triangulations of Ω_A , i.e., for which there exists θ_0 such that, θ_K being the smallest angle of all these triangles, we have the relation

$$\theta_K \geq \theta_0, \quad \forall K > 0.$$

Let us set $\Omega_K = \bigcup_{\mathcal{K} \in \tau_K} \mathcal{K}$, the largest side of all the triangles \mathcal{K} being less than a k . Consider now the finite dimensional space

$$V_K = \{V_K \in C^0(\overline{\Omega}_K); V_K|_{\mathcal{K}} \text{ is affine, } \forall \mathcal{K} \in \tau_K\}$$

and, correspondingly,

$$V_K^0 = \{v_K \in V_K; v_K|_{\Gamma_K} = 0\}$$

where $\Gamma_K = \delta A^K \cup C^K$, i.e., the union of all polygonal contours corresponding to δA and C , within the respective triangulation.

By denoting $w_K = \pi_K w$ ($w_K \in V_K$) the linear interpolation of $w = w_A$ on the nodes of \mathcal{K} and by putting

$$a_K(u, v) = \alpha_2 \int_{\Omega_K} \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right) dx dy \\ + \alpha_1 \int_{\Omega_K} \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right) dx dy$$

for all $u, v \in H^1(\Omega_K)$, the approximate problem comes to:

Find a function $u_K \in V_K^0$ such that $\Psi_K = u_K + w_K$ to be a solution of $a_K(\Psi_K, v_K) = \langle l, v_K \rangle$ for all $v_K \in V_K^0$.

The solving of this problem is given by

THEOREM 6.3. (1) The above approximate problem has a unique solution Ψ_K which converges towards Ψ_A when $\mathbf{k} \rightarrow 0$;

(2) If $\Psi_A \in H^2(\Omega_A)$ we have also the error estimation

$$\|\Psi_A - \Psi_K\|_{H^1(\Omega_A \cap \Omega_K)} \leq C_2 K \|\Psi_A\|_{H^2(\Omega_A)}$$

where the constant C_2 is independent of the parameter \mathbf{k} .

1.2 PDE-Toolbox of MATLAB

For the complicated shape domains and for more complicated equations one can call the Partial Differential Equations (PDE) toolbox from MATLAB. Shortly, it could be used in the following way:

- by the command `pdeTool` from the MATLAB work sheet the PDE Toolbox work sheet is open;

- from the menu *Options* activate *grid* and then, from the sub-menu *application* select *generic scalar* as equation type;

- from the menu *Draw* activate *draw mode* and then *polygon*;

- on the PDE Toolbox work sheet draw, with the mouse, the boundary of the considered (plane) computational domain (in our example a star-like domain);

- from the menu *Boundary* activate *boundary mode* and then *specify boundary conditions*; in the dialog window select the boundary condition type Dirichlet or Neumann and the corresponding parameters. The boundary conditions could be given separately on each boundary segment by a mouse click on that segment.

- from the menu *PDE* activate *PDE mode* and then *PDE specification*; in the dialog window choose the equation type - *elliptic*, which corresponds to the equation

$$-\text{div}(c * \text{grad}(u)) + a * u = f$$

and the coefficients: $c = 1$, $a = 0$, $f = \sin(x) \cdot \sin(y)$, for instance;

- from the menu *Mesh* activate *mesh mode* and then *initialize mesh*, which generates a starting triangular mesh that can be seen on the screen;

- from the menu *Solve* activate *Parameters* and in the dialog window activate *Adaptive mode*; This option allows the successive refinement of the mesh depending on the approximated solution. From the *Solve* menu too, activate *Solve PDE* and so the toolbox numerically solves the defined problem by the finite element method, performing also successive refinements of the mesh until a stopping criterion is verified;

- from the menu *Plot* activate *Parameters* and then, in the dialog window select *color* and *contour*, which determines the graphical visualization mode of the solution.

For any of the above actions there exist buttons, in the corresponding order, which facilitate the use of the toolbox. All the results on the PDE Toolbox work sheet may be exported on the MATLAB work sheet and used in complex programs.

For example, from the menu *Mesh*, activating *Export Mesh*, *OK* one could bring on the MATLAB work sheet the lists of the points, sides and triangles of the mesh, in the proposed variables p, e, t . Similarly, from the menu *Solve*, activating *Export Solution*, *OK* one could export on the MATLAB sheet the values of the numerical solution at the mesh points, in the proposed variable u .

Now, by the command `pdemesh(p,e,t)` on the MATLAB work sheet, the final computational triangular mesh could be graphically represented (see Figure 6.1). By the command `pdesurf(p,t,u)` the solution could be graphically represented (see Figure 6.2). By choosing a rectangular mesh, the calculated solution could be interpolated on that mesh using the commands `x=-1:0.01:1; y=x; uxy=tri2grid(p,t,u,x,y)`; The `uxy` variable will contain the numerical values of the solution at the grid points $(x-y)$ inside the computational domain and NaN at the other. The graphical representation can be performed now by `surf(x,y,uxy)`.

There are many other options, very well described in the *Help* pages of the toolbox and also in the demo examples.

For modeling and simulating many scientific and engineering problems based on partial differential equations, including 1D, 2D and 3D geometry, we recommend also the use of FEMLAB (www.femlab.com).

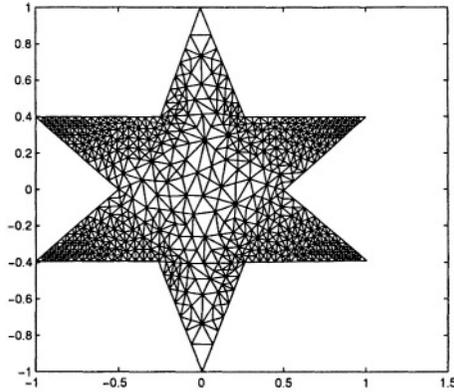


Figure 6.1. The triangular mesh on a star-like domain

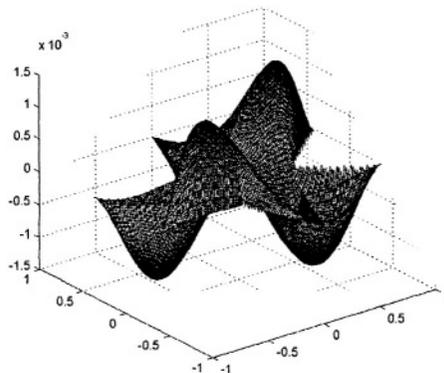


Figure 6.2. The numerical solution of the Dirichlet problem on the star-like domain

2. Least-Squares Finite Element Method (LSFEM)

We will briefly present, following [72], the least-squares finite element method (LSFEM).

2.1 First Order Model Problem

Let us consider the simplest first order differential problem

$$\begin{aligned} u'(x) &= f(x), x \in \Omega \equiv (0, 1), \\ u(0) &= 0. \end{aligned} \tag{6.7}$$

The classical solution is a function $u \in C^1(\Omega)$ which satisfies the above relations; it exists and it is unique for every $f \in C(\Omega)$. Moreover,

if $f \in C^m(\Omega)$ then $u \in C^{m+1}(\Omega)$ (we recall that $u \in C^m(\Omega)$ means that all $u(x), u'(x), \dots, u^{(m)}(x)$ are continuous on $\bar{\Omega}$).

Let us now convert the problem into a variational form, using the spaces $H^m(\Omega)$. For $m = 0$,

$$H^0(\Omega) \equiv L_2(\Omega) = \left\{ u : \Omega \rightarrow \mathbb{R}; \int_{\Omega} u^2 dx < \infty \right\}$$

with the norm and the inner product defined by

$$\|u\|_0 = \left\{ \int_{\Omega} u^2 dx \right\}^{\frac{1}{2}}, \quad (u, v) = \int_{\Omega} uv dx.$$

Typical examples of functions of H^0 are the continuous functions, the piecewise continuous functions, particularly piecewise constant functions, defined on $\bar{\Omega}$.

If $u \in C(\Omega), u' \in H^0(\Omega)$ and $\int_{x_0}^x u'(s) ds = u(x) - u(x_0)$, then $u \in H^1(\Omega)$. The corresponding norm is

$$\|u\|_1 = \left\{ \int_{\Omega} u^2 dx + \int_{\Omega} u'^2 dx \right\}^{\frac{1}{2}}.$$

Typical examples of functions of H^1 are the functions of $C^1(\Omega)$ or the piecewise differentiable continuous functions.

Generally, if $u \in C^{m-1}(\Omega)$ and $u^{(m)} \in H^0(\Omega)$, then $u \in H^m(\Omega)$ and we have the corresponding definitions of the norm. On H^1 we will also use the semi-norm

$$|u|_1 = \left\{ \int_{\Omega} u'^2 dx \right\}^{\frac{1}{2}}.$$

The following important inequalities hold:

a) *Friedrichs*, for $u \in H^1(0, l)$ and $u(0) = 0$,

$$\left\{ \int_0^l u^2 dx \right\}^{\frac{1}{2}} \leq l \left\{ \int_0^l u'^2 dx \right\}^{\frac{1}{2}};$$

b) *Sobolev*, for $u \in H^1(0, l)$,

$$|u(x)| \leq \sqrt{\frac{2}{l} + \frac{2l}{3}} \|u\|_1.$$

We will also use the basic lemma of the variational calculus

LEMMA. If $f \in C(\bar{\Omega})$ and

$$\int_{\Omega} f(x) \Phi(x) dx = 0, \quad \forall \Phi \in C(\bar{\Omega}),$$

then $f = 0$.

The finite element method to approximate the solution of the problem (6.7) consists in finding an approximate solution of the form

$$u_n(x) = a_1\Phi_1 + \dots + a_n\Phi_n$$

where Φ_i are known functions of a specific type, satisfying the condition $\Phi_i(0) = 0$. By denoting

$$R(x) = u'_n(x) - f(x),$$

the coefficients a_1, \dots, a_n can be calculated from the equations

$$\int_{\Omega} R(x)v(x)dx = 0$$

where v is an arbitrary continuous function. To calculate the n coefficients a_i we will choose n functions v .

A particular method is obtained from the above scheme by choosing the functions Φ_i and v from the subspace V_h of

$$V = \{u \in H^1(0, 1) | u(0) = 0\}$$

containing piecewise linear functions.

Let us consider the grid

$$0 = x_0 < x_1 < \dots < x_n = 1$$

which divides Ω into the *elements* $e_j = (x_{j-1}, x_j)$ of length h_j and let $h = \max h_j$ be. We will require that the elements u_h of V_h be continuous on $[0,1]$, linear on each element e_j and $u_h(0) = 0$.

The functions $u_h \in V_h$ could be described by their values u_j on the nodes. We have

$$u_h(x) = u_1\Phi_1(x) + \dots + u_n\Phi_n(x) \tag{6.8}$$

where

$$\Phi_j(x) = \begin{cases} 1, & x = x_j, \\ 0, & x = x_k \neq x_j, \\ \frac{x - x_{j-1}}{h_j}, & x \in (x_{j-1}, x_j), \\ \frac{x_j - x}{h_{j+1}}, & x \in (x_j, x_{j+1}), \\ 0, & x \in e_k, k \neq j, j + 1. \end{cases}$$

Then, the basis functions Φ_j have the value 1 at the corresponding nodes x_j , the value 0 at other nodes and are piecewise linear functions on each interval e_k . Obviously, $u_h(x_j) = u_j$ for $j = 1, \dots, n$.

It may be proved that the interpolation error of a function $u \in V$ by an interpolant function at the given grid $\Pi_h u$ is, if $u \in H^2(0, 1)$,

$$\|(u - \Pi_h u)'\|_0 \leq \frac{h}{\pi} |u|_2, \quad \|u - \Pi_h u\|_0 \leq \frac{h^2}{\pi^2} |u|_2.$$

The classical Galerkin finite difference method could be formulated as follows:

Find $u_h \in V_h$ such that

$$\int_0^1 (u'_h - f)v_h = 0, \quad \forall v_h \in V_h. \tag{6.9}$$

Since $u_h(x)$ is of the form (6.8), by choosing in (6.9) $v_h = \Phi_i(x)$ for $j = 1, \dots, n$ we obtain the system

$$\sum_{j=1}^n \int_0^1 \Phi'_j(x)\Phi_i(x)dx = \int_0^1 f(x)\Phi_i(x)dx, \quad i = 1, \dots, n$$

from which we can calculate the unknowns u_1, \dots, u_n . The above system can be rewritten

$$\sum_{j=1}^n (\Phi'_j, \Phi_i) u_j = (f, \Phi_i), \quad i = 1, \dots, n \tag{6.10}$$

or, in matrix form,

$$KU = F.$$

The elements K_{ij} of the matrix K of order $n \times n$ could be easily calculated (in the general case they are obtained by assembling the values on each element). We have, for $i = 1, \dots, n - 1$,

$$K_{ii} = (\Phi'_i, \Phi_i) = \int_{x_{i-1}}^{x_i} \frac{1}{h_i} \frac{x - x_{i-1}}{h_i} dx + \int_{x_i}^{x_{i+1}} \frac{-1}{h_{i+1}} \frac{x_{i+1} - x}{h_{i+1}} dx = 0,$$

and

$$K_{nn} = (\Phi'_n, \Phi_n) = \int_{x_{n-1}}^{x_n} \frac{1}{h_n} \frac{x - x_{n-1}}{h_n} dx = \frac{1}{2}.$$

Moreover, for $i = 2, \dots, n - 1$ we have

$$K_{i-1,i} = (\Phi'_{i-1}, \Phi_i) = \int_{x_{i-1}}^{x_i} \frac{-1}{h_i} \frac{x - x_{i-1}}{h_i} dx = -\frac{1}{2},$$

$$K_{i+1,i} = (\Phi'_{i+1}, \Phi_i) = \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}} \frac{x_{i+1} - x}{h_{i+1}} dx = \frac{1}{2}.$$

As regards the calculation of F , by some simple quadrature formulas (trapezoidal rule), we obtain, for $i = 1, \dots, n - 1$,

$$F_i = (f, \Phi_i) = \int_{x_{i-1}}^{x_i} f(x) \frac{x - x_{i-1}}{h_i} dx + \int_{x_i}^{x_{i+1}} f(x) \frac{x_{i+1} - x}{h_{i+1}} dx \simeq \frac{f_i h_i}{2} + \frac{f_i h_{i+1}}{2}$$

and

$$F_n = (f, \Phi_n) = \int_{x_{n-1}}^{x_n} f(x) \frac{x - x_{n-1}}{h_n} dx \simeq \frac{f_n h_n}{2}.$$

If we choose a uniform grid $h_i = h = \frac{1}{n}$, the system (6.10) becomes

$$\frac{1}{2} \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 0 & 1 \\ 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ \vdots \\ u_n \end{pmatrix} = h \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{n-1} \\ f_n/2 \end{pmatrix}.$$

We remark that the equations are of the form

$$\frac{u_2 - 0}{2h} = f_1, \frac{u_3 - u_1}{2h} = f_2, \frac{u_4 - u_3}{2h} = f_3, \dots,$$

incidentally identical to the equations obtained by the centered finite difference method

$$u'_i \simeq \frac{u_{i+1} - u_{i-1}}{2h}.$$

As we know, this structure of the matrix leads to a solution decoupling on odd-even nodes and then oscillatory numerical solutions appear. These oscillations persist even if the grid is refined. In practice, we choose instead of the centered difference an upwind difference

$$u'_i \simeq \frac{u_i - u_{i-1}}{h} = \frac{u_{i+1} - u_{i-1}}{2h} - \frac{h}{2} \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$$

which is equivalent to introducing a numerical dissipation. In fact, instead of the given equation, we numerically solve by centered finite differences the “perturbed” equation

$$u' - \frac{h}{2} u'' = f.$$

Let us present now the LSFEM. In this case we try to minimize the integral

$$I(u) = \int_0^1 (u' - f)^2 dx$$

onto the space V . The necessary minimum condition is that the first variation vanishes, i.e.,

$$\lim_{t \rightarrow 0} \frac{d}{dt} I(u + tv) \equiv 2 \int_0^1 (u' - f)v' dx = 0, \forall v \in V$$

or

$$(u', v') = (f, v'), \forall v \in V.$$

The discrete problem is now:

Find $u_h \in V$ such that

$$\int_0^1 (u'_h - f)v'_h dx = 0, \forall v_h \in V_h.$$

Since u_h is of the form (6.8), we now obtain the system

$$\sum_{j=1}^n (\Phi'_j, \Phi'_i) u_j = (f, \Phi'_i), i = 1, \dots, n \tag{6.11}$$

of the same form

$$KU = F.$$

By recalculating the matrix K and the right-hand side F , we find

$$K_{i,i} = (\Phi'_i, \Phi'_i) = \int_{x_{i-1}}^{x_i} \frac{1}{h_i^2} dx + \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}^2} dx = \frac{1}{h_i} + \frac{1}{h_{i+1}}, i = 1, \dots, n-1,$$

$$K_{n,n} = (\Phi'_n, \Phi'_n) = \int_{x_{n-1}}^{x_n} \frac{1}{h_n^2} dx = \frac{1}{h_n},$$

$$K_{i-1,i} = (\Phi'_{i-1}, \Phi'_i) = - \int_{x_{i-1}}^{x_i} \frac{1}{h_i^2} dx = -\frac{1}{h_i}, i = 2, \dots, n-1,$$

$$K_{i+1,i} = (\Phi'_{i+1}, \Phi'_i) = - \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}^2} dx = -\frac{1}{h_{i+1}}, i = 2, \dots, n-1,$$

$$F_i = (f, \Phi'_i) = \int_{x_{i-1}}^{x_i} f \frac{1}{h_i} dx + \int_{x_i}^{x_{i+1}} f \frac{-1}{h_{i+1}} dx$$

$$\simeq \frac{f_{i-1} + f_i}{2} - \frac{f_i + f_{i+1}}{2} = -\frac{f_{i+1} - f_{i-1}}{2}, i = 1, \dots, n-1,$$

$$F_n = (f, \Phi'_n) = \int_{x_{n-1}}^{x_n} f \frac{1}{h_n} dx \simeq \frac{f_{n-1} + f_n}{2}.$$

Now the matrix K is symmetric and positive-definite. In the particular case when $h_i = h = \frac{1}{n}$, the above system takes the form

$$-\frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ \vdots \\ u_n \end{pmatrix} = \frac{1}{2h} \begin{pmatrix} f_2 - f_0 \\ f_3 - f_1 \\ f_4 - f_2 \\ \vdots \\ f_n - f_{n-2} \\ -f_n - f_{n-1} \end{pmatrix}.$$

We remark now that the left-hand side can be interpreted as the centered finite differences discretization of u'' , while the right-hand side is the centered finite differences discretization of f' .

Here is the explanation of this fact. The variational problem was to find $u \in V$ for which

$$\int_0^1 (u' - f)v' = 0, \forall v \in V,$$

Assuming that u'' exists, upon integration by parts we find

$$(u' - f)v|_{x=1} - \int_0^1 (u'' - f')v dx = 0, \forall v \in V.$$

Consequently,

$$\begin{aligned} u'' - f' &= 0 \text{ in } (0, 1), \\ u' - f &= 0 \text{ in } x = 1, \\ u &= 0 \text{ in } x = 0. \end{aligned}$$

This means that the derivative of the original first order equation must be satisfied on the interval, the original equation must be satisfied at $x = 1$ as the natural boundary condition and the original boundary condition becomes an essential boundary condition. Therefore, we have the Galerkin formulation for a second order equation, which is very efficient. Moreover, the condition number of the matrix K is of order $O(h^{-2})$, which is similar to that from the classical Galerkin method for second order equations.

Concluding, the least squares method transforms the difficult (as regard the numerics) problem for a first order equation into an easily solvable second order equation.

If we study the error of the method, if $u \in H^2(0, 1)$ then we have

$$\|(u - u_h)'\|_0 \leq h |u|_2, \|u - u_h\|_0 \leq Ch^2 |u|_2,$$

i.e., an optimal error, of the same order as the interpolation by finite elements error.

We will present now a very simple example which illustrates the power of the LSEFEM. Let us consider the problem

$$\begin{aligned}
 u'(x) &= \frac{1 - e^{-\frac{1-x}{\epsilon}}}{1 - e^{-\frac{1}{\epsilon}}}, x \in [0, 1], \\
 u(0) &= 0
 \end{aligned}$$

with the exact solution

$$u_{ex}(x) = 1 - \frac{1 - e^{-\frac{1-x}{\epsilon}}}{1 - e^{-\frac{1}{\epsilon}}}.$$

For $\epsilon = 0.03$ we use the upwind scheme, the Galerkin finite element method and LSFEM. The comparison with the exact solution is given in Figure 6.3.

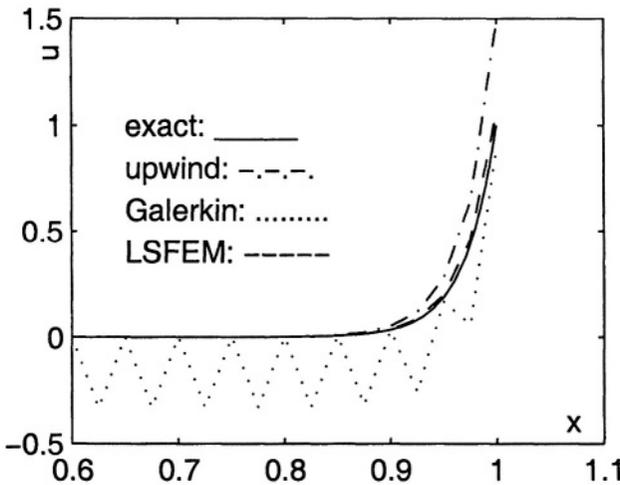


Figure 6.3. The approximate solutions for $\epsilon = 0.03$ and $h = 0.025$

2.2 The Mathematical Foundation of the Least-Squares Finite Element Method

Let $\Omega \subset \mathbb{R}^d$, with $d = 2$ or 3 , be an open bounded domain, with a piecewise smooth boundary Γ (i.e., it can be decomposed into a finite number of arcs (surfaces) and each of them can be locally represented by indefinitely differentiable functions; more, the angles between the arcs

(surfaces) are greater than zero). As examples, a sphere, cube, torus, triangle, polyhedron, etc. have piecewise smooth boundaries.

For $1 \leq p < \infty$ we have

$$L_p(\Omega) = \left\{ u : \Omega \rightarrow \mathbb{R} \mid \int_{\Omega} |u(x)|^p dx < \infty \right\}$$

which is a Banach space with the norm

$$\|u\|_{L_p} = \left(\int_{\Omega} |u(x)|^p dx \right)^{\frac{1}{p}}.$$

For $p = 2$, $L_2(\Omega)$ is a Hilbert space with the inner product, respectively the norm,

$$(u, v) = \int_{\Omega} u(x)v(x)dx, \|u\|_0 = (u, u)^{\frac{1}{2}}.$$

For every integer $k > 0$ we have the Sobolev space

$$H^k(\Omega) = \{u \in L_2(\Omega) \mid D^{\alpha}u \in L_2(\Omega), \forall |\alpha| \leq k\}$$

which is a Hilbert space, with the norm

$$\|u\|_k = \left(\|u_0\|^2 + \sum_{|\alpha| \leq k} \|D^{\alpha}u\|_0^2 \right)^{\frac{1}{2}}.$$

Obviously, $H^0(\Omega) = L_2(\Omega)$. Particularly, the space $H^1(\Omega)$ and its subspace

$$H_0^1(\Omega) = \{u \in H^1(\Omega) \mid u|_{\Gamma} = 0\}$$

with the norm

$$\|u\|_1 = \left(\|u_0\|^2 + \sum_{i=1}^d \left\| \frac{\partial u}{\partial x_i} \right\|_0^2 \right)^{\frac{1}{2}}$$

and, respectively, the semi-norm on $H^1(\Omega)$

$$|u|_1 = \left(\sum_{i=1}^d \left\| \frac{\partial u}{\partial x_i} \right\|_0^2 \right)^{\frac{1}{2}}$$

are of great interest.

For vector-valued functions \mathbf{u} with m components we consider the product spaces

$$\mathbf{L}_2(\Omega) = [L_2(\Omega)]^m, \quad \mathbf{H}^1(\Omega) = [H^1(\Omega)]^m \quad (6.12)$$

with the corresponding norms

$$\|\mathbf{u}\|_j^2 = \sum_{k=1}^m \|u_k\|_j^2, \quad j = 0, 1. \tag{6.13}$$

Let us consider now a linear equation

$$Au = f \tag{6.14}$$

where $f \in V$. Its solution is denoted by $u = A^{-1}f$. It is important that the practical (proposed) problems be well-posed. This means that the above equation has a solution for every $f \in V$, the solution is unique in a space U and if f changes “a little”, the solution u also changes “a little”. In operator language, this means that $A : U \rightarrow V$ is one-to-one and its inverse is continuous between the normed spaces U and V .

THEOREM 6.4. *The necessary and sufficient condition for a linear operator A to have a continuous inverse is that*

$$\exists \alpha > 0 : \alpha \|u\|_U \leq \|Au\|_V, \quad \forall u \in U. \tag{6.15}$$

In the case of an operator A satisfying the above condition, we have for the equation (6.14)

$$\|u\|_U \leq \frac{1}{\alpha} \|Au\|_V = \frac{1}{\alpha} \|f\|_V$$

which means that the solution u continuously depends on the data.

If u_h is an approximation (obtained by a certain method) of the exact solution, we have

$$\|u_h - u\|_U \leq \frac{1}{\alpha} \|Au_h - Au\|_V = \frac{1}{\alpha} \|Au_h - f\|_V.$$

Therefore, if the norm of the residual $R_h = Au_h - f$ tends towards zero for $h \rightarrow 0$, then $u_h \rightarrow u$ in U .

The proof of the property (6.15) uses

THEOREM 6.5. (The Friedrichs inequality) *If $u \in H^1(\Omega)$ and $u|_{\Gamma_1} = 0$, where $\Gamma_1 \subset \Gamma$, then there exists a real constant $C > 0$, which is independent of u , such that*

$$\|u\|_0 \leq C |u|_1.$$

We remark that one gets $\|u\|_1 \leq C |u|_1$ and consequently, on $H_0^1(\Omega)$ the semi-norm $|u|_1$ may be used instead of a norm.

THEOREM 6.6. (The Poincaré inequality) *If $u \in H^1(\Omega)$, then there exists a real constant $C > 0$, independent of u , such that*

$$\|u\|_0^2 \leq C \left\{ |u|_1^2 + \left(\int_{\Omega} u dx \right)^2 \right\}.$$

Let us define now the finite element spaces V_h . These spaces consist of piecewise polynomial functions. More precisely, the domain Ω is decomposed by a *triangulation* T_h with the elements K . In the case $d = 1$, the elements are intervals, for $d = 2$ the elements are triangles or quadrilaterals while for $d = 3$ they are tetrahedrons or hexahedrons.

We will denote by $P_r(K)$ the space of polynomials of order less than or equal to r , defined on K , and by $Q_r(K)$ the polynomials of order less than or equal to r in each of the variables.

We will define $V_h \subset H^1(\Omega)$ if we solve boundary value problems of first order and $V_h \subset H^2(\Omega)$ if we solve problems of second order. In the case of the *piecewise polynomial functions*, we have

$$V_h \subset H^1(\Omega) \Leftrightarrow V_h \subset C(\Omega),$$

$$V_h \subset H^2(\Omega) \Leftrightarrow V_h \subset C^1(\Omega)$$

where $\bar{\Omega} = \Omega \cup \Gamma$. Since the treatment of second order problems is difficult by the LSFEM, we will consider only first order systems and the high order problems can be reduced to this case.

Let us describe now the elements K . We will present only the case $d = 2$ and we will suppose that Ω is a polygonal plane domain. We will divide Ω into

$$\Omega = K_1 \cup K_2 \cup \dots \cup K_m,$$

generating the triangulation $T_h = \{K_1, \dots, K_m\}$. It is necessary that the triangles do not overlap and that no vertex of one triangle lies on the edge of another triangle (it can coincide only with another vertex).

We will define the parameter h of the triangulation as the maximum diameter of all circles circumscribing the triangles from T_h and ρ as the minimum diameter of all circles inscribed in the triangles from T_h . We suppose that there exists a constant $\beta > 0$ independent of h , such that

$$\frac{\rho}{h} \geq \beta. \quad (6.16)$$

This condition avoids the generation of arbitrarily thin triangles (or of interior angles arbitrarily small).

Let us consider now a triangle from such a triangulation. The nodes are the vertices A_1, A_2, A_3 of the triangle. We will construct the linear interpolant of u on this triangle

$$\Pi_h u(x, y) = \alpha_0 + \alpha_1 x + \alpha_2 y.$$

From the interpolation condition

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

we obtain the coefficients α_i and by substitution we find

$$\Pi_h u(x, y) = \psi_1(x, y)u_1 + \psi_2(x, y)u_2 + \psi_3(x, y)u_3$$

where

$$\begin{aligned} \psi_1(x, y) &= \frac{1}{D} \begin{vmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}, \psi_2(x, y) = \frac{1}{D} \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x & y \\ 1 & x_3 & y_3 \end{vmatrix}, \\ \psi_3(x, y) &= \frac{1}{D} \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x & y \end{vmatrix}, D = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}. \end{aligned} \quad (6.17)$$

We remark that the order of the nodes A_1, A_2, A_3 is important. The counterclockwise sense corresponds to $D > 0$, otherwise $D < 0$. We remark also

$$\sum_{i=1}^3 \psi_i(x, y) = 1.$$

About the interpolation errors, if we work with polynomials of order $r \geq 1$ and for functions u sufficiently smooth, we have

$$\|u - \Pi_h u\|_0 \leq Ch^{r+1} |u|_{r+1},$$

$$\|u - \Pi_h u\|_1 \leq Ch^r |u|_{r+1}.$$

Let us now give a general formulation of LSFEM. We will consider only steady state problems. In the evolution case, the time discretization leads at each step to a boundary value problem of this type.

The linear boundary value problem which we consider is

$$\begin{aligned} A\mathbf{u} &= \mathbf{f}, \text{ in } \Omega, \\ B\mathbf{u} &= \mathbf{g}, \text{ on } \Gamma, \end{aligned} \quad (6.18)$$

where A is a partial differential of the first order operator

$$A\mathbf{u} = \sum_{i=1}^d A_i \frac{\partial \mathbf{u}}{\partial x_i} + A_0 \mathbf{u}.$$

Here $\mathbf{u}^T = (u_1, \dots, u_m)$ is the vector of the unknown functions, $\mathbf{f}^T = (f_1, \dots, f_m)$ are given functions, A_i and A_0 are continuous matrices which depend on $x \in \Omega$ while B is a boundary operator.

We will suppose $\mathbf{f} \in \mathbf{L}_2(\Omega)$ and we choose a suitable subspace \mathbf{V} of $\mathbf{L}_2(\Omega)$ which involves the boundary conditions. Let $R = A\mathbf{v} - \mathbf{f}$ be the residual of \mathbf{v} and we have

$$\|R\|_0^2 = \int_{\Omega} (A\mathbf{v} - \mathbf{f})^2 dx \geq 0. \quad (6.19)$$

A solution \mathbf{u} of the problem (6.18) could be interpreted as an element of \mathbf{V} which minimizes the residual

$$0 = \|R(\mathbf{u})\|_0^2 \leq \|R(\mathbf{v})\|_0^2, \forall \mathbf{v} \in \mathbf{V}.$$

The least squares method minimizes (6.19) in \mathbf{V} , i.e.,

$$I(\mathbf{v}) = \|A\mathbf{v} - \mathbf{f}\|_0^2 = (A\mathbf{v} - \mathbf{f}, A\mathbf{v} - \mathbf{f}) \rightarrow \min.$$

A necessary minimum condition is that the first variation vanishes at \mathbf{u} ,

$$\lim_{t \rightarrow 0} \frac{d}{dt} I(\mathbf{u} + t\mathbf{v}) \equiv 2 \int_{\Omega} (A\mathbf{v})^T (A\mathbf{u} - \mathbf{f}) dx = 0, \forall \mathbf{v} \in \mathbf{V}.$$

Thus, the problem to solve is to find $\mathbf{u} \in \mathbf{V}$ such that

$$B(\mathbf{u}, \mathbf{v}) \equiv (A\mathbf{u}, A\mathbf{v}) = (\mathbf{f}, A\mathbf{v}) \equiv F(\mathbf{v}), \forall \mathbf{v} \in \mathbf{V}. \tag{6.20}$$

We remark that $B(\mathbf{u}, \mathbf{v})$ is symmetric and by discretization will lead to a symmetric positive-defined matrix.

In finite element discretization we choose a unique basis for all the unknown functions and we are looking for u_h in the form

$$\mathbf{u}_h(x) = \sum_{j=1}^N \psi_j(x) \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix}_j$$

where $u_{k,j}$ are the values of \mathbf{u} at the node j and N is the number of nodes of an element. Introducing this expression into the equation (6.20) we obtain the system

$$KU = F$$

where U is the global vector of the values at the nodes.

The global matrix K is assembled from the element matrices

$$K_e = \int_{\Omega_e} (A\psi_1, \dots, A\psi_N)^T (A\psi_1, \dots, A\psi_N) dx$$

where $\Omega_e \subset \Omega$ is the domain covered by the e^{th} element. F is obtained by assembling the element vectors

$$F_e = \int_{\Omega_e} (A\psi_1, \dots, A\psi_N)^T f dx.$$

We remark that the boundary condition could be also included into $B(u, v)$ and then no boundary conditions are imposed on the subspace V .

If we write now the Euler–Lagrange equation for the problem (6.20), upon application of the Green formula we find

$$(A^* \mathbf{A} \mathbf{u} - A^* \mathbf{f}, \mathbf{v}) + \langle \mathbf{A} \mathbf{u} - \mathbf{f}, \mathbf{v}_\Gamma \rangle = 0, \forall \mathbf{v} \in \mathbf{V}$$

where A^* is the adjoint operator of A . The Euler–Lagrange equation is therefore

$$A^* \mathbf{A} \mathbf{u} = A^* \mathbf{f} \text{ in } \Omega. \tag{6.21}$$

For this equation, the boundary condition $B \mathbf{u} = 0$ is an essential boundary condition and $\langle \mathbf{A} \mathbf{u} - \mathbf{f}, \mathbf{v}_\Gamma \rangle = 0$ is a natural boundary condition.

Concluding, the least squares method for first order systems is equivalent to the Galerkin method for the second order system (6.21). We remark also that A^*A is a self-adjoint operator, even if A itself is not self-adjoint.

We will estimate now the errors for this method. We need the following result:

THEOREM 6.7. *If the first order linear differential system $\mathbf{A} \mathbf{u} = \mathbf{f}$ has a unique solution \mathbf{u} which continuously depends on the data $\mathbf{f} \in \mathbf{L}^2(\Omega)$, then there exists a positive constant m such that*

$$m \|\mathbf{u}\|_0 \leq \|\mathbf{A} \mathbf{u}\|_0.$$

Moreover, if the solution $\mathbf{u} \in \mathbf{H}^1(\Omega)$, then there exists a positive constant M such that

$$\|\mathbf{A} \mathbf{u}\|_0 \leq M \|\mathbf{u}\|_1.$$

Let us suppose now that \mathbf{V}_h is a subspace of \mathbf{V} which consists of piecewise polynomials of order r and the problem $\mathbf{A} \mathbf{u} = \mathbf{f}$ is well-posed. Then for \mathbf{u} and \mathbf{u}_h we have

$$(\mathbf{A} \mathbf{u}, \mathbf{A} \mathbf{v}) = (\mathbf{f}, \mathbf{A} \mathbf{v}), \forall \mathbf{v} \in \mathbf{V},$$

$$(\mathbf{A} \mathbf{u}_h, \mathbf{A} \mathbf{v}_h) = (\mathbf{f}, \mathbf{A} \mathbf{v}_h), \forall \mathbf{v}_h \in \mathbf{V}_h.$$

Particularly, we have also

$$(\mathbf{A} \mathbf{u}, \mathbf{A} \mathbf{v}_h) = (\mathbf{f}, \mathbf{A} \mathbf{v}_h), \forall \mathbf{v}_h \in \mathbf{V}_h$$

from which, by subtracting, we obtain

$$(\mathbf{A}(\mathbf{u} - \mathbf{u}_h), \mathbf{A} \mathbf{v}_h) = 0, \forall \mathbf{v}_h \in \mathbf{V}_h.$$

Let now $\Pi_h \mathbf{u} \in \mathbf{V}_h$ be the interpolant of \mathbf{u} . Then, from the above relation, we have

$$\|\mathbf{A}(\mathbf{u} - \mathbf{u}_h)\|_0^2 = (\mathbf{A}(\mathbf{u} - \mathbf{u}_h), \mathbf{A}(\mathbf{u} - \mathbf{u}_h)) = (\mathbf{A}(\mathbf{u} - \mathbf{u}_h), \mathbf{A}(\mathbf{u} - \Pi_h \mathbf{u}))$$

$$\begin{aligned}
 + (A(\mathbf{u} - \mathbf{u}_h), A(\Pi_h \mathbf{u} - \mathbf{u}_h)) &= (A(\mathbf{u} - \mathbf{u}_h), A(\mathbf{u} - \Pi_h \mathbf{u})) \\
 &\leq \|A(\mathbf{u} - \mathbf{u}_h)\|_0 \|A(\mathbf{u} - \Pi_h \mathbf{u})\|_0 .
 \end{aligned}$$

By simplifying and from the above theorem, we get

$$m \|\mathbf{u} - \mathbf{u}_h\|_0 \leq \|A(\mathbf{u} - \mathbf{u}_h)\|_0 \leq \|A(\mathbf{u} - \Pi_h \mathbf{u})\|_0 \leq M \|\mathbf{u} - \Pi_h \mathbf{u}\|_1 .$$

Since $R(\mathbf{u}_h) = A\mathbf{u}_h - \mathbf{f} = -A(\mathbf{u} - \mathbf{u}_h)$, from the above relation and from the form of the interpolation error we obtain

THEOREM 6.8. *If the problem (6.18) is well-posed and its solution is sufficiently smooth, then we have the estimations of the error*

$$\|R(\mathbf{u}_h)\|_0 \leq C_1 h^r |\mathbf{u}|_{r+1} ,$$

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq C_2 h^r |\mathbf{u}|_{r+1} .$$

This theorem ensures the convergence of the method and it does not matter what type the first order system is, elliptic, hyperbolic, mixed, etc. We remark that in the elliptic case we are able to give an improved result. If A is elliptic and coercive, i.e.,

$$\alpha \|\mathbf{v}\|_1 \leq \|A\mathbf{v}\|_0, \forall \mathbf{v} \in \mathbf{V} = \{\mathbf{v} \in \mathbf{H}^1 \mid B\mathbf{v} = 0 \text{ on } \Gamma\} ,$$

then we have the optimal estimation:

THEOREM 6.9. *If \mathbf{u}_h is the solution generated by the LSFEM for the elliptic, coercive system (6.18), with piecewise polynomial of order r , then there exists constants C_1 and C_2 independent of \mathbf{u} and h such that*

$$\|\mathbf{u} - \mathbf{u}_h\|_1 \leq C_1 h^r |\mathbf{u}|_{r+1} ,$$

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq C_2 h^{r+1} |\mathbf{u}|_{r+1} .$$

2.3 Div-Curl (Rot) Systems

We will present on such types of systems the use of the LSFEM. Suppose that $\Omega \subset \mathbb{R}^3$ is a bounded domain, with the piecewise smooth boundary $\Gamma = \Gamma_1 \cup \Gamma_2$ (one or another of the components may be empty but not both; if both are not empty, they must have at least one common point). We will denote by \mathbf{n} the unit outward normal to the boundary, $\boldsymbol{\tau}$ a tangential vector to Γ at a boundary point.

We present, without proof, some technical properties.

THEOREM 6.10. *Let Ω be bounded and convex in \mathbb{R}^3 . Then for every function $\mathbf{u} \in \mathbf{H}^1(\Omega)$ satisfying $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ_1 and $\mathbf{n} \times \mathbf{u} = 0$ on Γ_2 we have*

$$|\mathbf{u}|_1^2 \leq \|\nabla \cdot \mathbf{u}\|_0^2 + \|\nabla \times \mathbf{u}\|_0^2 .$$

THEOREM 6.11. *If Ω is bounded simply connected in \mathbb{R}^3 and $\mathbf{u} \in \mathbf{H}^1(\Omega)$ satisfies*

$$\begin{aligned} \nabla \cdot \mathbf{u} &= 0, \nabla \times \mathbf{u} = 0 \text{ in } \Omega, \\ \mathbf{n} \cdot \mathbf{u} &= 0 \text{ on } \Gamma_1, \mathbf{n} \times \mathbf{u} = 0 \text{ on } \Gamma_2, \end{aligned}$$

then $\mathbf{u} = 0$ in Ω .

THEOREM 6.12. (The Friedrichs inequality). *Let Ω be a bounded and simply connected domain in \mathbb{R}^3 . Then for every function $\mathbf{u} \in \mathbf{H}^1(\Omega)$ satisfying $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ_1 and $\mathbf{n} \times \mathbf{u} = 0$ on Γ_2 we have*

$$\|\mathbf{u}\|_1^2 \leq C \left(\|\nabla \cdot \mathbf{u}\|_0^2 + \|\nabla \times \mathbf{u}\|_0^2 \right)$$

where C depends only on Ω .

The above result shows that on the space

$$\mathbf{H} = \left\{ [H^1(\Omega)]^3 \mid \mathbf{n} \cdot \mathbf{u} = 0 \text{ on } \Gamma_1, \mathbf{n} \times \mathbf{u} = 0 \text{ on } \Gamma_2 \right\}$$

the norms $\|\mathbf{u}\|_1$ and $\left(\|\nabla \cdot \mathbf{u}\|_0^2 + \|\nabla \times \mathbf{u}\|_0^2 \right)^{1/2}$ are equivalent (for Ω as above).

THEOREM 6.13. (The Gradient theorem). *If $g \in H^1(\Omega)$ satisfies*

$$\nabla g = 0 \text{ in } \Omega, \quad g = 0 \text{ on } \Gamma_1 \neq \emptyset,$$

then $g = 0$ in Ω .

THEOREM 6.14. *If $\mathbf{u} \in \mathbf{H}^1(\Omega)$ and $\mathbf{n} \times \mathbf{u} = 0$ on $\Gamma_2 \neq \emptyset$, then $\mathbf{n} \cdot \nabla \times \mathbf{u} = 0$ on Γ_2 .*

THEOREM 6.15. (The second Friedrichs inequality). *Let Ω be a bounded and simply connected domain in \mathbb{R}^3 , with the smooth boundary Γ . For every $\mathbf{u} \in \mathbf{H}^1(\Omega)$ we have*

$$\|\mathbf{u}\|_{1,\Omega}^2 \leq C \left(\|\mathbf{u}\|_{0,\Omega}^2 + \|\nabla \cdot \mathbf{u}\|_{0,\Omega}^2 + \|\nabla \times \mathbf{u}\|_{0,\Omega}^2 + \|\mathbf{n} \cdot \mathbf{u}\|_{1/2,\Gamma}^2 \right).$$

Let us consider now the 3D divergence-curl system

$$\begin{aligned} \nabla \times \mathbf{u} &= \boldsymbol{\omega}, \quad \nabla \cdot \mathbf{u} = \rho \text{ in } \Omega, \\ \mathbf{n} \cdot \mathbf{u} &= 0 \text{ on } \Gamma_1, \quad \mathbf{n} \times \mathbf{u} = 0 \text{ on } \Gamma_2. \end{aligned} \tag{6.22}$$

The given vector $\boldsymbol{\omega} \in \mathbf{L}^2(\Omega)$ must satisfy the compatibility conditions

$$\nabla \cdot \boldsymbol{\omega} = 0 \text{ in } \Omega, \quad \mathbf{n} \cdot \boldsymbol{\omega} = 0 \text{ on } \Gamma_2, \quad \int_{\Gamma} \mathbf{n} \cdot \boldsymbol{\omega} d\Gamma = 0, \tag{6.23}$$

and if $\Gamma_2 = \emptyset$, then $\rho \in L^2(\Omega)$ must satisfy

$$\int_{\Omega} \rho d\Omega = 0.$$

This system of four equations with three unknowns is not overdetermined. By introducing the dummy variable ϑ , the system can be rewritten

$$\begin{aligned}\nabla\vartheta + \nabla \times \mathbf{u} &= \boldsymbol{\omega}, \quad \nabla \cdot \mathbf{u} = \rho \quad \text{in } \Omega, \\ \mathbf{n} \cdot \mathbf{u} &= 0, \quad \vartheta = 0 \quad \text{on } \Gamma_1, \quad \mathbf{n} \times \mathbf{u} = 0 \quad \text{on } \Gamma_2.\end{aligned}\quad (6.24)$$

But, from Theorem 6.8, the first vector equation is equivalent with the system

$$\begin{aligned}\nabla \times (\nabla\vartheta + \nabla \times \mathbf{u} - \boldsymbol{\omega}) &= 0 \quad \text{in } \Omega, \\ \nabla \cdot (\nabla\vartheta + \nabla \times \mathbf{u} - \boldsymbol{\omega}) &= 0 \quad \text{in } \Omega, \\ \mathbf{n} \cdot (\nabla\vartheta + \nabla \times \mathbf{u} - \boldsymbol{\omega}) &= 0 \quad \text{on } \Gamma_1, \\ \mathbf{n} \times (\nabla\vartheta + \nabla \times \mathbf{u} - \boldsymbol{\omega}) &= 0 \quad \text{on } \Gamma_2.\end{aligned}$$

From the conditions (6.23), $\mathbf{n} \times \mathbf{u} = 0$ on Γ_2 and Theorem 6.11, the above relations yield

$$\begin{aligned}\Delta\vartheta &= 0 \quad \text{in } \Omega, \\ \vartheta &= 0 \quad \text{on } \Gamma_1, \quad \frac{\partial\vartheta}{\partial n} = 0 \quad \text{on } \Gamma_2,\end{aligned}$$

thus $\vartheta = 0$ in Ω and its introduction does not change the original system.

In Cartesian coordinates, for $\mathbf{u} = (u, v, w)^T$, the system is written

$$\begin{aligned}\vartheta_x + w_y - v_z &= \omega_x, \\ \vartheta_y + u_z - w_x &= \omega_y, \\ \vartheta_z + v_x - u_y &= \omega_z, \\ u_x + v_y + w_z &= \rho,\end{aligned}$$

while in matrix form

$$\mathbf{A}_1\mathbf{U}_x + \mathbf{A}_2\mathbf{U}_y + \mathbf{A}_3\mathbf{U}_z + \mathbf{A}_0\mathbf{U} = \mathbf{F},$$

where

$$\begin{aligned}\mathbf{A}_1 &= \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \\ \rho \end{pmatrix}, \\ \mathbf{A}_3 &= \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{A}_0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} u \\ v \\ w \\ \vartheta \end{pmatrix}.\end{aligned}$$

But the associated characteristic polynomial is

$$\det(\mathbf{A}_1\xi + \mathbf{A}_2\eta + \mathbf{A}_3\zeta) = \det \begin{pmatrix} 0 & -\zeta & \eta & \xi \\ \zeta & 0 & -\xi & \eta \\ -\eta & \xi & 0 & \zeta \\ \xi & \eta & \zeta & 0 \end{pmatrix} \\ = (\xi^2 + \eta^2 + \zeta^2)^2 > 0$$

for all $(\xi, \eta, \zeta) \neq \mathbf{0}$, thus the system is elliptic. We need two boundary conditions on each boundary. But $\vartheta = 0$ and $\mathbf{n} \cdot \mathbf{u} = 0$ are two conditions on Γ_1 while $\mathbf{n} \times \mathbf{u} = 0$ implies that two tangential components of \mathbf{u} on Γ_2 are zero.

Let us apply the least-squares method. We construct the functional

$$I : \mathbf{V} \rightarrow \mathbb{R}, I(\mathbf{u}) = \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u} - \rho\|_0^2$$

where

$$\mathbf{V} = \{ \mathbf{u} \in \mathbf{H}^1(\Omega) \mid \mathbf{n} \cdot \mathbf{u} = 0 \text{ on } \Gamma_1, \mathbf{n} \times \mathbf{u} = 0 \text{ on } \Gamma_2 \}.$$

If the variation of I vanishes, we obtain the variational formulation in the least squares sense: find $\mathbf{u} \in \mathbf{V}$ such that

$$(\mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}) = (\mathbf{f}, \mathbf{A}\mathbf{v}), \quad \forall \mathbf{v} \in \mathbf{V} \tag{6.25}$$

where

$$(\mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}) = (\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) + (\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v})$$

and

$$(\mathbf{f}, \mathbf{A}\mathbf{v}) = (\boldsymbol{\omega}, \nabla \times \mathbf{v}) + (\rho, \nabla \cdot \mathbf{v}).$$

From the Friedrichs inequality we obtain

$$\frac{1}{C} \|\mathbf{u}\|_1^2 \leq (\mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{u}) = (\mathbf{f}, \mathbf{A}\mathbf{u}) \leq \|\mathbf{u}\|_1 (\|\boldsymbol{\omega}\|_0 + \|\rho\|_0)$$

and then

THEOREM 6.16. *The solution of the problem (6.22) or (6.25) exists, it is unique and it satisfies*

$$\|\mathbf{u}\|_1 \leq C (\|\boldsymbol{\omega}\|_0 + \|\rho\|_0).$$

To apply the finite element method, we rewrite the equation (6.25) under the form

$$(\nabla \times \mathbf{u} - \boldsymbol{\omega}, \nabla \times \mathbf{v}) + (\nabla \cdot \mathbf{u} - \rho, \nabla \cdot \mathbf{v}) = 0, \quad \forall \mathbf{v} \in V.$$

With the hypothesis that all the functions are smooth enough, we use the Green formulas in the above relation and we get

$$(\nabla \times (\nabla \times \mathbf{u} - \boldsymbol{\omega}), \mathbf{v}) + \langle (\nabla \times \mathbf{u} - \boldsymbol{\omega}), \mathbf{n} \times \mathbf{v} \rangle_{\Gamma} - (\nabla (\nabla \cdot \mathbf{u} - \rho), \mathbf{v}) + \langle (\nabla \cdot \mathbf{u} - \rho), \mathbf{n} \cdot \mathbf{v} \rangle_{\Gamma} = 0, \quad \forall \mathbf{v} \in V.$$

Taking into account

$$\nabla \times \nabla \times \mathbf{u} = \nabla (\nabla \cdot \mathbf{u}) - \Delta \mathbf{u}$$

and the conditions satisfied by \mathbf{v} on Γ , the above relation becomes

$$(-\Delta \mathbf{u} - \nabla \times \boldsymbol{\omega} + \nabla \rho, \mathbf{v}) - \langle \mathbf{n} \times (\nabla \times \mathbf{u} - \boldsymbol{\omega}), \mathbf{v} \rangle_{\Gamma_1} + \langle (\nabla \cdot \mathbf{u} - \rho), \mathbf{n} \cdot \mathbf{v} \rangle_{\Gamma_2} = 0 \quad (6.26)$$

for all $\mathbf{v} \in V$. Then the Euler–Lagrange equation and the boundary conditions become

$$\Delta \mathbf{u} = -\nabla \times \boldsymbol{\omega} + \nabla \rho, \quad (6.27)$$

$$\mathbf{n} \cdot \mathbf{u} = 0, \quad \mathbf{n} \times (\nabla \times \mathbf{u}) = \mathbf{n} \times \boldsymbol{\omega} \text{ on } \Gamma_1, \quad (6.28)$$

$$\mathbf{n} \times \mathbf{u} = 0, \quad \nabla \cdot \mathbf{u} = \rho \text{ on } \Gamma_2. \quad (6.29)$$

We remark that now the divergence equation does not appear on the domain Ω . The solutions of the uncoupled Poisson system (6.27) with the mentioned boundary conditions automatically satisfy the divergence equation. In fact, if \mathbf{u} is smooth enough, the variational problem (6.26) is equivalent to the original problem (6.22).

We can now discretize the above problem by the finite element method. Let us construct the subspace $V_h \subset V$ of continuous, piecewise polynomial of order $r \geq 1$ functions and for the finite element solution $\mathbf{u}_h \in V_h$ we have

THEOREM 6.17. *The finite element method based on the equation (6.25) has an optimal convergence and an optimal satisfaction of the divergence equation, i.e.,*

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq C_1 h^{r+1} \|\mathbf{u}\|_{r+1},$$

$$\|\nabla \cdot \mathbf{u}_h - \rho\|_0 \leq C_2 h^r \|\mathbf{u}\|_{r+1}.$$

Concluding, the application of the LSFEM to the original problem (6.22) is reduced to the application of the Galerkin finite element method (6.26) to the system (6.27, 6.28, 6.29). This system contains three uncoupled Poisson equations; the essential boundary conditions come from the original boundary conditions while the natural boundary conditions come from the original equations considered on the boundary too.

2.4 Div-Curl (Rot)-Grad System

Let us consider now the second order elliptic boundary value problem

$$\begin{aligned} -\nabla \cdot \nabla \Phi &= f(x) \text{ in } \Omega, \\ \Phi &= 0 \text{ on } \Gamma, \end{aligned} \quad (6.30)$$

where $f \in L^2(\Omega)$. Here Φ could be a temperature, a potential, etc.

Let us try to transform this problem into a first order divergence-gradient problem

$$\begin{aligned} \mathbf{u} - \nabla \Phi &= 0, \quad \nabla \cdot \mathbf{u} = -f \text{ in } \Omega, \\ \Phi &= 0 \text{ on } \Gamma. \end{aligned} \quad (6.31)$$

The variational form is obtained by multiplying the first equation by $\mathbf{v} \in \mathbf{L}^2(\Omega)$ and integrating, multiplying then the second equation by $\psi \in H_0^1(\Omega)$ and integrating.

The variational problem is to find the pair

$$\{\Phi, \mathbf{u}\} \in H_0^1(\Omega) \times \mathbf{L}^2(\Omega)$$

for which

$$\begin{aligned} (\mathbf{u}, \mathbf{v}) - (\nabla \Phi, \mathbf{v}) &= 0, \quad \forall \mathbf{v} \in \mathbf{L}^2(\Omega), \\ -(\nabla \psi, \mathbf{u}) &= -(f, \psi), \quad \forall \psi \in H_0^1(\Omega). \end{aligned} \quad (6.32)$$

But, by finite element discretization, the associated matrix is not positively defined.

By applying the least-squares method in the classical form to the problem (6.32), which is *not* an elliptic system, we are led to a convergence which is not optimal.

The *optimal* least-squares method is based on the system

$$\begin{aligned} \nabla \cdot \mathbf{u} &= -f, \quad \nabla \times \mathbf{u} = 0, \quad \nabla \Phi - \mathbf{u} = 0 \text{ in } \Omega, \\ \Phi &= 0, \quad \mathbf{n} \times \mathbf{u} = 0 \text{ on } \Gamma. \end{aligned} \quad (6.33)$$

Although the second equation could be obtained from the third and the second boundary condition could be obtained from the first, the presence of these relations is very important.

In the two-dimensional case the system (6.33) consists of four equations with three unknown functions. As in the previous section, by introduction of a dummy variable ϑ , it will be shown that the system is well determined and elliptic. In the Cartesian coordinates, it is

$$\begin{aligned} u_x + v_y &= -f, \quad -u_y + v_x = 0 \text{ in } \Omega, \\ \Phi_x - \vartheta_y - u &= 0, \quad \Phi_y + \vartheta_x - v = 0 \text{ in } \Omega, \\ \Phi &= 0, \quad un_y - vn_x = 0 \text{ on } \Gamma. \end{aligned} \quad (6.34)$$

The equations containing ϑ are equivalent to

$$\begin{aligned} \nabla \cdot (-\mathit{curl} \vartheta + \nabla\Phi - \mathbf{u}) &= 0 \text{ in } \Omega, \\ \nabla \times (-\mathit{curl} \vartheta + \nabla\Phi - \mathbf{u}) &= 0 \text{ in } \Omega, \\ \mathbf{n} \times (-\mathit{curl} \vartheta + \nabla\Phi - \mathbf{u}) &= 0 \text{ on } \Gamma, \end{aligned}$$

where $\mathit{curl} \vartheta = \left(\frac{\partial\vartheta}{\partial y}, -\frac{\partial\vartheta}{\partial x} \right)$. But from the last two equations we get

$$\begin{aligned} \Delta\vartheta &= 0 \text{ in } \Omega, \\ \frac{\partial\vartheta}{\partial n} &= 0 \text{ on } \Gamma, \end{aligned}$$

i.e., ϑ is a constant and the introduction of it does not change the original system. In the matrix form we have

$$\mathbf{A}_1 \mathbf{U}_x + \mathbf{A}_2 \mathbf{U}_y + \mathbf{A}_0 \mathbf{U} = \mathbf{F}$$

where

$$\begin{aligned} \mathbf{A}_1 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \\ \mathbf{A}_0 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} -f \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} u \\ v \\ \Phi \\ \vartheta \end{pmatrix}. \end{aligned}$$

But

$$\det(\mathbf{A}_1\xi + \mathbf{A}_2\eta) = \det \begin{pmatrix} \xi & \eta & 0 & 0 \\ -\eta & \xi & 0 & 0 \\ 0 & 0 & \xi & -\eta \\ 0 & 0 & \eta & \xi \end{pmatrix} = (\xi^2 + \eta^2)^2 > 0$$

for every vector $(\xi, \eta) \neq \mathbf{0}$. Consequently, the extended system is elliptic with four equations and four unknowns, therefore we need two boundary conditions.

Let us study the errors. We will denote

$$H = H_0^1(\Omega), \quad S = \{\mathbf{v} \in \mathbf{H}^1(\Omega) \mid \mathbf{n} \times \mathbf{v} = 0 \text{ on } \Gamma\}.$$

The optimal least-squares method minimizes the functional

$$I : H \times S \rightarrow \mathbb{R}, \quad I(\Phi, \mathbf{u}) = \|\nabla \cdot \mathbf{u} + f\|_0^2 + \|\nabla \times \mathbf{u}\|_0^2 + \|\nabla\Phi - \mathbf{u}\|_0^2.$$

We remark again that the variable ϑ is indeed a dummy variable which has nothing to do with the numerical computation.

If the variation of I vanishes with respect to Φ and \mathbf{u} , we obtain the variational formulation: Find $\mathbf{U} = (\Phi, \mathbf{u}) \in H \times S$, such that

$$(A\mathbf{U}, A\mathbf{V}) = (f, A\mathbf{V}), \quad \forall \mathbf{V} = (\psi, \mathbf{v}) \in H \times S$$

where

$$(A\mathbf{U}, A\mathbf{V}) = (\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v}) + (\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) + (\nabla \Phi - \mathbf{u}, \nabla \psi - \mathbf{v}),$$

$$(f, A\mathbf{V}) = (-f, \nabla \cdot \mathbf{v}).$$

The discretized by the finite element method problem is to find

$$\mathbf{U}_h = (\Phi_h, \mathbf{u}_h) \in H_h \times S_h$$

such that

$$(A\mathbf{U}_h, A\mathbf{V}_h) = (f, A\mathbf{V}_h), \quad \forall \mathbf{V}_h = (\psi_h, \mathbf{v}_h) \in H_h \times S_h$$

where

$$(A\mathbf{U}_h, A\mathbf{V}_h) = (\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h) + (\nabla \times \mathbf{u}_h, \nabla \times \mathbf{v}_h) + (\nabla \Phi_h - \mathbf{u}_h, \nabla \psi_h - \mathbf{v}_h),$$

$$(f, A\mathbf{V}_h) = (-f, \nabla \cdot \mathbf{v}_h).$$

It can be proved that

$$\|A\mathbf{V}\|_0 \leq C \|\mathbf{V}\|_1, \quad \|A\mathbf{V}\|_0^2 \geq \alpha \|\mathbf{V}\|_1^2, \quad \forall \mathbf{V} \in H \times S$$

where

$$\|\mathbf{V}\|_1^2 = \|\Phi\|_1^2 + \|\mathbf{u}\|_1^2.$$

Consequently, A is continuous and coercive and therefore we have an optimal convergence.

2.5 Stokes' Problem

Let us consider now the Stokes problem

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f}, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0, & \text{in } \Omega, \\ \mathbf{u}|_{\partial\Omega} &= \mathbf{0}, \end{aligned}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with the sufficiently smooth boundary $\partial\Omega$. We define the bilinear forms $a(\mathbf{u}, \mathbf{v}) = (\nabla \mathbf{u}, \nabla \mathbf{v})$ and $b(p, \mathbf{v}) = -(p, \nabla \cdot \mathbf{v})$, so the weak formulation of the above Stokes problem is

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(p, \mathbf{v}) &= (\mathbf{f}, \mathbf{v}), & \forall \mathbf{v} \in \mathbf{H}_0^1, \\ b(q, \mathbf{u}) &= 0, & \forall q \in L_0^2, \end{aligned}$$

where the solution (\mathbf{u}, p) is looked for in a suitable space $\mathbf{u} \in \mathbf{H}_0^1, p \in L_0^2$.

To approximate this solution, we choose the finite dimensional subspaces H_h , included or not included in $H_0^1(\Omega)$, $L_h \subset L^2(\Omega)$ containing piecewise polynomial functions with respect to a simple decomposition T_h of Ω (triangles, quadrilaterals, etc.). Of course, we require a conservation of the elements shape condition during the refinement. In the discrete form, the Stokes problem becomes

$$\begin{aligned} a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(p_h, \mathbf{v}_h) &= (\mathbf{f}, \mathbf{v}_h), & \forall \mathbf{v}_h \in \mathbf{H}_h, \\ b_h(q_h, \mathbf{u}_h) &= 0, & \forall q_h \in L_h, \end{aligned}$$

where $(\mathbf{u}_h, \mathbf{v}_h)$ are looked for in $\mathbf{H}_h \times L_h$.

The necessary and sufficient conditions for the existence and the convergence of the approximations $(\mathbf{u}_h, \mathbf{v}_h)$ are

$$\begin{aligned} \inf_{\mathbf{v}_h \in \mathbf{H}_h} \|\mathbf{v} - \mathbf{v}_h\|_h &\leq ch^{m-1} \|\mathbf{v}\|_{\mathbf{H}^m}, & \forall \mathbf{v} \in \mathbf{H} \cap \mathbf{H}^m(\Omega), \\ \inf_{q_h \in L_h} \|q - q_h\|_0 &\leq ch^{m-1} \|q\|_{H^{m-1}}, & \forall q \in L \cap H^{m-1}(\Omega) \end{aligned}$$

for $m \geq 2$ while for the stability (the inf-sup or Babuška–Brezzi conditions) are

$$\min_{q_h \in L_h} \max_{\mathbf{v}_h \in \mathbf{H}_h} \frac{b_h(q_h, \mathbf{v}_h)}{\|\mathbf{v}_h\|_h \|q_h\|_0} \geq \beta$$

for a $\beta > 0$ independent of h .

Within these conditions we have the approximation result

$$\|\mathbf{u} - \mathbf{u}_h\|_0 + h \|\mathbf{u} - \mathbf{u}_h\|_h + h \|p - p_h\|_0 \leq ch^{m-1} \{ \|\mathbf{u}\|_{\mathbf{H}^m} + \|p\|_{H^{m-1}} \}.$$

In the literature many admissible pairs of finite elements spaces for velocities and for pressure are described.

An interesting procedure to study the Stokes problem is the use of the $u - p - \omega$ form together with the least-squares finite-element method. By introducing the vorticity $\omega = \nabla \times \mathbf{u}$, the Stokes system may be written

in a first order form, namely

$$\begin{aligned} \nabla p + \nabla \times \boldsymbol{\omega} &= \mathbf{f}, & \text{in } \Omega, \\ \nabla \cdot \boldsymbol{\omega} &= 0, & \text{in } \Omega, \\ \boldsymbol{\omega} - \nabla \times \mathbf{u} &= 0, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0, & \text{in } \Omega. \end{aligned}$$

Although in this form we have eight equations with seven unknowns (in the 3D case), the system is determined and elliptic. This fact may be remarked by introducing the auxiliary unknown Φ satisfying $\Phi|_{\partial\Omega} = 0$,

$$\begin{aligned} \nabla p + \nabla \times \boldsymbol{\omega} &= \mathbf{f}, & \text{in } \Omega, \\ \nabla \cdot \boldsymbol{\omega} &= 0, & \text{in } \Omega, \\ -\boldsymbol{\omega} + \nabla\Phi + \nabla \times \mathbf{u} &= 0, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0, & \text{in } \Omega. \end{aligned}$$

By introducing the third equation into the second, we find $\Delta\Phi = 0$ and thus $\Phi = 0$ on Ω , i.e., the use of Φ does not change the initial system, but now we have (in the 3D case) eight equations with eight unknowns.

If we write the above system on the components, ($\mathbf{u} = (u, v, w)$, $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)$), we have

$$\begin{aligned} \frac{\partial p}{\partial x} + \frac{\partial \omega_x}{\partial y} - \frac{\partial \omega_y}{\partial z} &= f_x, \quad \frac{\partial p}{\partial y} + \frac{\partial \omega_x}{\partial z} - \frac{\partial \omega_z}{\partial x} = f_y, \\ \frac{\partial p}{\partial z} + \frac{\partial \omega_y}{\partial x} - \frac{\partial \omega_x}{\partial y} &= f_z, \quad \frac{\partial \omega_x}{\partial x} + \frac{\partial \omega_y}{\partial y} + \frac{\partial \omega_z}{\partial z} = 0, \\ -\omega_x + \frac{\partial \Phi}{\partial x} + \frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} &= 0, \quad -\omega_y + \frac{\partial \Phi}{\partial y} + \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} = 0, \\ -\omega_z + \frac{\partial \Phi}{\partial z} + \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} &= 0, \quad \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \end{aligned}$$

or, in matrix form

$$A_1 \frac{\partial \mathbf{U}}{\partial x} + A_2 \frac{\partial \mathbf{U}}{\partial y} + A_3 \frac{\partial \mathbf{U}}{\partial z} + \mathbf{A}\mathbf{U} = \mathbf{F}, \tag{6.35}$$

where

$$\mathbf{F} = \begin{pmatrix} f_x \\ f_y \\ f_z \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} u \\ v \\ w \\ \omega_x \\ \omega_y \\ \omega_z \\ p \end{pmatrix}.$$

The associated characteristic polynomial satisfies

$$\det(A_1\xi + A_2\eta + A_3\zeta) = (\xi^2 + \eta^2 + \zeta^2)^4 > 0$$

for every $(\xi, \eta, \zeta) \neq 0$, which confirms the ellipticity of the system. Further, it is equivalent to the initial Stokes system.

Finally, to the system (6.35) one applies the least-squares finite-element method, see [72].

3. **Boundary Element Method (BEM)**

The boundary element method, developed especially after 1970, is a numerical method to solve boundary value problems. In fact, this method, using a solution of the homogeneous differential operator or a fundamental solution of the differential operator, associates to the given boundary value problem an integral representation which reduces solving of the differential problem to determining the solution of an integral equation on the boundary of the domain.

By the numerical (or analytical) integration of these integral equations on the boundary, which could require both a boundary discretization and the use of some quadrature formulas, we obtain numerical values which, through the associated integral representation, allow the evaluation of the solution at any interior point of the considered domain.

The integral equation on the boundary decreases by 1 the dimension of the problem to solve and, more, it incorporates the boundary conditions so that no other special relations are needed. Unbounded domains could be also treated without any special preparation, the conditions at infinity being incorporated into the respective integral equations.

Of course, the price that one must pay for the above facilities is the necessity to construct an *integral representation associated to the boundary value problem* and an *integral equation on the boundary*, the two tools required by this method. For the first one, generally, it is necessary to have explicit solutions of the homogeneous associated equation or fundamental solutions for the given equation and this fact will restrain

the applicability of the method foremost to differential operators with constant coefficients. Concerning determination of the integral equations on the boundary, different methods could be used, as for example passage to the limit in the associated integral representation when an arbitrary point “tends” towards a boundary point, or Green theorems or Somigliana identity or Betti theorems (for elasticity problems) etc. A unitary formulation, which is superior by its generality, could be obtained from the weighted residual formulation of the problem.

In particular cases, like the bidimensional boundary value problems for the Laplace operator, using the formalism of complex variables and functions, construction of integral equations on the boundary could be avoided, obviously using supplementary hypotheses (the so-called CVBEM variant).¹

The integral equations on the boundary, due to the diversity of the usually involved singularities, exceed the well studied Fredholm equations frame. This fact explains the lack of a unitary mathematical theory for BEM.

W.L. Wendland and his co-workers have obtained promising results in the construction of such a theory, situating BEM within the theory of pseudodifferential operators.

In what follows we will sketch some basic elements of BEM, using the book [100] and then the variant CVBEM will be developed, a variant which gives a convergent procedure and an extremely practical working instrument in plane hydrodynamics.

3.1 Abstract Formulation of the Boundary Element Method

Let \mathcal{L} be a differential operator defined on a certain space of functions with its values in another (maybe the same) space of functions. Considering then, on a domain Ω with the boundary Γ , the differential equation $\mathcal{L}(u) = b$, where u and b are functions from the domain of definition, respectively from the codomain (range) of \mathcal{L} (functions defined on the same Ω), to this equation one attaches, usually, a set of boundary conditions of the type $S(u) = s$ on Γ_1 , $G(u) = g$ on Γ_2 where $\Gamma = \Gamma_1 \cup \Gamma_2$, $\Gamma_1 \cap \Gamma_2 = \emptyset$.

The operators S and G , defined on the same space of functions as \mathcal{L} with values in some space of functions defined on Γ_1 , respectively Γ_2 , will correspond to the so-called *essential boundary conditions*, respectively *natural boundary conditions* (the essential ones having a determinant

¹ CVBEM (Complex Variables Boundary Element Method).

role for the uniqueness of the solution of the problem). In fact, we get these operators S and G by defining, first, on the working spaces (the domain and the codomain of \mathcal{L}) an inner (scalar) product, for instance $\langle \alpha, \beta \rangle \equiv \int_{\Omega} \alpha \beta d\Omega$. Then, if we integrate by parts the inner product $\langle \mathcal{L}(u), w \rangle \equiv \int_{\Omega} L(u) w d\Omega$ (w being a weight function defined on the same Ω , belonging to a space of functions which could coincide with that of u), until all the derivatives of the function u are eliminated, we will have for this inner product the “transposed” form

$$\int_{\Omega} \mathcal{L}(u) w d\Omega = \int_{\Omega} u \mathcal{L}^*(w) d\Omega + \int_{\Gamma} [S^*(w) G(u) - G^*(w) S(u)] d\Gamma,$$

S and G being the differential operators which appear after the integration by parts. By definition $S^*(w)$ contains w terms resulting from the first stage of integration while $S(u)$ contains the corresponding terms (of the same order of differentiability) in u . The operator \mathcal{L}^* is called the adjoint operator for \mathcal{L} . If $\mathcal{L} = \mathcal{L}^*$ we say that \mathcal{L} is self-adjoint and then we have also $G = G^*$ and $S = S^*$.

The above writing of the inner product does not give only the possibility to appreciate whether or not the operator is self-adjoint but also two different types of boundary conditions as the operators $S(u)$ or $G(u)$ are given at the points of Γ .

Of course the above form of the inner product anticipates that the working space (the domain and the codomain of \mathcal{L}) will be some subspaces, with some differentiability properties, of $L^2(\Omega)$ or, more generally, of $H^m(\Omega)$, m corresponding to the order of the differential operator \mathcal{L} . The “boundary values” of the functions u and of its derivatives at the points of Γ (and, implicitly, those of the operators S and G) will be understood everywhere in the sense of the values of the *trace operators* $T_S : H^m(\Omega) \rightarrow L^2(\Gamma)$, operators which exist by virtue of the Trace theorem, and which, for $u \in C^m(\bar{\Omega})$, $T_S u = D^i u|_{\Gamma}$ $i = 0, \dots, m - 1$.

Let now u_0 be the exact “punctual” solution of the boundary value problem and u an approximation of it belonging to the same space of functions. Obviously, corresponding to this approximation, we have either a “residue” (“error”) joined to the equation fulfilment, i.e., $R = \mathcal{L}(u) - b \neq 0$, or the residues linked with the boundary conditions satisfaction, i.e., $R_1 = S(u) - s \neq 0$ on Γ_1 and $R_2 = G(u) - g \neq 0$ on Γ_2 . The purpose of any approximation procedure is to make these residues (errors) as small as possible. Depending on the manner of performing this task, we have different types of approximation. So, if we require R to be zero at certain points or subdomains of Ω , we obtain *the*

points (on subdomains) collocation method which generalizes the finite differences method. If we consider a weight function w of a suitable space of functions, we could ask that the error R satisfy the requirement $\langle R, w \rangle = \int_{\Omega} R w d\Omega = 0$. This implies the “mean” satisfaction of the given equation and we are led to a weighted residuals method (the Galerkin method, respectively the moment method for w belonging to the same class as u).

If both $R_1 \neq 0$ and $R_2 \neq 0$ it is natural to have also (with respect to the weight function w),

$$\langle R_1, G(w) \rangle = \int_{\Gamma_1} [S(u) - s] G(w) d\Gamma = 0,$$

$$\langle R_2, w \rangle = \int_{\Gamma_2} [G(u) - g] w d\Gamma = 0.$$

It is expected to use, instead of the given initial problem, as starting point in the construction of the approximation u , the unique “weighted” equation

$$\langle R, w \rangle = -\langle R_1, G(w) \rangle + \langle R_2, w \rangle.$$

In fact, this equation could be obtained from the equation $\langle R, w \rangle = 0$, performing the integration by parts and, once the operators S and G have appeared, imposing on the approximation u the fulfilment of the conditions $S(u) = s$ and $G(u) = g$.

In what follows, in order to extend the domain of the possible approximations, we will try, first, to weaken the regularity conditions on the approximating function u (with the price of the corresponding “strengthening” of the requirements on the weight functions w) and then, to get the exact satisfaction of the equation (or of its adjoint), with the price of losing the only approximative fulfilment of the boundary conditions. This way will lead, finally, to the boundary element method (BEM).

Thus we obtain the *weak formulation* (a first reduction of the regularity requirements on u) and the *inverse formulation* (the complete elimination of the derivatives of the function u to the obvious detriment of the function w which takes over the respective derivatives). Of course, any solution of the initial “weighted” equations will be also a solution of the weak or inverse formulation equation but the reciprocal statement, generally, is not true.

We retain the requirement that the boundary element method will be correlated with the inverse formulation of the weighted equation. If in

the previous numerical methods (finite element method, finite differences method, etc.) we constructed the functions by approximating the desired solution on the domain Ω while satisfying the boundary conditions on Γ , in BEM we act contrarily: we choose the exact or fundamental solutions for the differential operator (or for its adjoint) and then we try to satisfy approximately the boundary conditions.

Basically, the working instruments of the boundary element method are, as we previously stated, both an integral representation of the solution, associated to the boundary value problem on the considered domain, and an integral equation on the boundary whose solving allows then, by the associated integral representation, the construction of the solution at any interior point of the considered domain.

Let us suppose now, as an example, that the operator $\mathcal{L} \equiv \nabla^2$, obviously a self-adjoint operator, and the boundary conditions joined to the equation $\nabla^2 u_0 = b$ in Ω are $u_0 = \bar{u}$ on Γ_1 (essential), respectively $q_0 \left(\equiv \frac{\partial u_0}{\partial n} \right) = \bar{q}_0$ on Γ_2 (natural). If the exact solution u_0 is approximated by u and obviously q_0 by $q \equiv \frac{\partial u}{\partial n}$, we will have also, together with the residue (error) $R_2 = \nabla^2 u - b$ in Ω , the boundary residues $R_1 = u - \bar{u}$ on Γ_1 and $R_2 = q - \bar{q}$ on Γ_2 .

Considering then the weighted equation $\langle R, w \rangle = -\langle R_1, G(w) \rangle + \langle R_2, w \rangle$ which is synonymous with imposing on the approximation u the conditions $u = \bar{u}$ on Γ_1 and $q = \bar{q}_0$ on Γ_2 , we are led to:

(i) *the original formulation*

$$\int_{\Omega} (\nabla^2 u - b) w d\Omega = \int_{\Gamma_2} (q - \bar{q}) w d\Gamma - \int_{\Gamma_1} (u - \bar{u}) \frac{\partial w}{\partial n} d\Gamma$$

(which represents the starting point in the genuine Galerkin method, when u and w belong to the same class, and in the weighted residuals method and implicitly in the finite differences method, when u and w belong to different classes);

(ii) *the weak formulation*

$$\int_{\Omega} \frac{\partial u}{\partial x_k} \frac{\partial w}{\partial x_k} d\Omega - \int_{\Omega} b w d\Omega = \int_{\Gamma_2} \bar{q} w d\Gamma + \int_{\Gamma_1} q w d\Gamma + \int_{\Gamma_1} (u - \bar{u}) \frac{\partial w}{\partial n} d\Gamma$$

(the starting point for the finite element method, for u and w belonging to the same class, and for the weak weighted residuals formulations);

(iii) *inverse formulation*

$$\int_{\Omega} (\nabla^2 w) u d\Omega - \int_{\Omega} b w d\Omega = - \int_{\Gamma_2} \bar{q} w d\Gamma - \int_{\Gamma_1} q w d\Gamma + \int_{\Gamma_2} u \frac{\partial w}{\partial n} d\Gamma + \int_{\Gamma_1} \bar{u} \frac{\partial w}{\partial n} d\Gamma$$

(the origin of the Trefftz method, for u and w belonging to the same space).

Confining ourselves to the inverse formulation but in the case when u and w belong to different classes and $\nabla^2 w = 0$ or $\nabla^2 w = \delta_{x_i-x}$, the inverse formulation will lead to a set of boundary relations which allow us to calculate the approximation u by satisfying the boundary requirements and implicitly, the construction “a posteriori” of the approximative solution at any point of the domain. Accepting, for instance, $\nabla^2 w = \delta_{x_i-x}$ because $\int_{\Omega} u \delta_{x_i-x} d\Omega = u(x)$ (for any function u with a compact support and continuous in a vicinity of x), the inverse formulation gives the so-called integral representation attached to the problem. Moreover, if in this integral representation we make $x \rightarrow \xi \in \Gamma$, by denoting the fundamental solution u^* while $q^* = \frac{\partial u^*}{\partial n}$, we will obtain an integral equation on the boundary of the type

$$c(\xi)u(\xi) + \int_{\Gamma} u(x_i)q^*(\xi, x_i)d\Gamma(x_i) = \int_{\Gamma} q(x_i)u^*(\xi, x_i)d\Gamma(x_i),$$

which represents, in fact, the compatibility condition of the boundary data and which could be the integral equation attached to the problem, an essential tool for BEM.

Concerning the coefficient $c(\xi)$, if we confine ourselves only to the bidimensional case, it will be equal to π if ξ belongs to a smooth portion of Γ , and it will be equal to $\pi + \alpha_1 - \alpha_2$ if ξ is a cuspidal point, framed by the smooth portions Γ_1 and Γ_2 of the boundary Γ whose outward normals form the angles α_1 and α_2 respectively with the Ox_1 axis.

We remark that both the integral representation and the integral equation on the boundary are not uniquely determined. They could be obtained in different ways, but the principles of the BEM are the same, not depending on the used technique.

3.2 Variant of the Complex Variables Boundary Element Method [112]

In the sequel we will give a variant of BEM, the so-called CVBEM, which provides total satisfaction in the problems where the unknown is a holomorphic function, as many of the plane hydrodynamics problems are. In this case, the simple use of the Cauchy formula already gives an

integral representation attached to the considered problem. Moreover, the use of an appropriate system of interpolating functions allows us to avoid the boundary integral equation; the data on the boundary could be calculated by solving an algebraic system, without any approximation of the boundary and without any numerical quadrature. It is also remarkable that CVBEM is a convergent procedure, within quite large conditions.

Let then $f(z)$ be a holomorphic function in the simply connected domain D , the outside of a Jordan rectifiable curve C . Suppose that $f(z)$ is continuous on $D \cup C$ and its real or imaginary part or even a combination of them being known on the boundary C . The Cauchy formula, that is

$$f(z) = \frac{1}{2\pi i} \int_{\tilde{C}} \frac{f(\zeta)}{\zeta - z} d\zeta + a_0, \quad z \in D,$$

($a_0 = \lim_{|z| \rightarrow \infty} f(z)$) will be the integral representation of the envisaged problem. Now we want to determine $f(\zeta)$ i.e., $f(z)|_C$. For that let us consider a system of nodes z_0, z_1, \dots, z_n ($z_0 \equiv z_n$) on C , placed counterclockwise, separating the contour C into the arcs C_j ($j = \overline{1, n}$), C_j being the arc joining the node z_{j-1} with z_j . Considering then the approximation $\tilde{f}(\zeta)$ of the unknown function $f(\zeta)$, defined by $\tilde{f}(\zeta) = \sum_{j=1}^n f_j L_j(\zeta)$ where $f_j = f(z_j)$ while $L_j(\zeta)$ are the Lagrange interpolating functions, constructed for every respective arc, i.e.,

$$L_j(\zeta) = \begin{cases} \frac{\zeta - z_{j-1}}{z_j - z_{j-1}}, & \zeta \in C_j, \\ \frac{\zeta - z_{j+1}}{z_j - z_{j+1}}, & \zeta \in C_{j+1}, \\ 0, & \text{otherwise,} \end{cases}$$

the Cauchy integral becomes (up to a constant a_0)

$$f^*(\zeta) = \sum_{j=1}^n f_j \tilde{L}_j(z)$$

where

$$\begin{aligned} \tilde{L}_j(z) &= \frac{1}{2\pi i} \int_{\tilde{C}} \frac{L_j(\zeta)}{\zeta - z} d\zeta \\ &= \frac{1}{2\pi i} \left(\frac{z - z_{j-1}}{z_j - z_{j-1}} \log \frac{z - z_j}{z - z_{j-1}} + \frac{z - z_{j+1}}{z_j - z_{j+1}} \log \frac{z - z_{j+1}}{z - z_j} \right) \end{aligned}$$

(we accept that the principal determination for the complex logarithm has been considered).

Supposing now that $f^*(z)$ is evaluated at the nodes z_k ($k = \overline{1, n}$), that is

$$f^*(z_k) (\equiv f_k \equiv u_k + iv_k) = \sum_{j=1}^n f_j \tilde{L}_j(z_k) \left(\equiv \sum_{j=1}^n f_j L_{jk} \right),$$

we are led to the algebraic system of n equations with n unknowns

$$\begin{cases} u_k = \sum_{j=1}^n M_{kj} u_j - \sum_{j=1}^n N_{kj} v_j, \\ v_k = \sum_{j=1}^n M_{kj} v_j + \sum_{j=1}^n N_{kj} u_j, \end{cases}$$

where $M_{jk} + iN_{jk} = L_{jk} (\equiv \tilde{L}_j(z_k))$.

By its solving we will obtain the approximation $\tilde{f}(\zeta)$ of the function $f(\zeta)$ and implicitly, via the Cauchy formula, the solution at any point of the domain D .

If the Jordan curve C has at the node z_p a cuspidal point, the angle of the semi-tangents at this point being $\pi - \mu\pi$ with $-1 \leq \mu < 0$, then the Cauchy formula is applicable again, the behaviour of $f(z)$ in a vicinity $V(z_p)$ being given by $f(z) - f(z_0) = (z - z_0)^{\frac{1}{1-\mu}} h(z)$ with $h(z_0) \neq 0$, i.e., $\frac{df}{dz} = o \left[(z - z_0)^{\frac{\mu}{1-\mu}} \right]$. In this case, the piecewise interpolation must take into account this behaviour in $V(z_p)$, which is performed by a similar approximation on $C_j, j \neq p, p + 1$ while on C_p and C_{p+1} we will take

$$\tilde{f}(\zeta) = \begin{cases} f_p + (f_{p-1} - f_p) \frac{(\zeta - z_p)^{\frac{1}{1-\mu}}}{z_{p-1} - z_p}, & \zeta \in C_p, \\ f_p + (f_{p+1} - f_p) \frac{(\zeta - z_p)^{\frac{1}{1-\mu}}}{z_{p+1} - z_p}, & \zeta \in C_{p+1}. \end{cases}$$

This choice does not change the structure of $L_j(z)$ for $j \neq p - 1, p, p + 1$ while for $j = p - 1, p, p + 1$ we have respectively

$$\tilde{L}_{p-1}(z) = \frac{1}{2\pi i} \left\{ \frac{z - z_{p-2}}{z_{p-1} - z_{p-2}} \log \frac{z - z_{p-1}}{z - z_{p-2}} - F_{\frac{1}{1-\mu}} \left(\frac{z - z_p}{z_{p-1} - z_p} \right) \right\},$$

$$\tilde{L}_p(z) = \frac{1}{2\pi i} \left\{ \log \frac{z - z_{p+1}}{z - z_{p-1}} + F_{\frac{1}{1-\mu}} \left(\frac{z - z_p}{z_{p-1} - z_p} \right) - F_{\frac{1}{1-\mu}} \left(\frac{z - z_p}{z_{p+1} - z_p} \right) \right\},$$

$$\tilde{L}_{p+1}(z) = \frac{1}{2\pi i} \left\{ \frac{z - z_{p+2}}{z_{p+1} - z_{p+2}} \log \frac{z - z_{p+2}}{z_p - z_{p+1}} - 1 + F_{\frac{1}{1-\mu}} \left(\frac{z - z_p}{z_{p+1} - z_p} \right) \right\}$$

where $F_\alpha(z) = \int_0^1 \frac{t^\alpha}{t-z} dt$

Concerning the coefficients $L_{kj} = \tilde{L}_k(z_j)$ for $k \neq j$, they could be directly calculated from the expression of L_j using, in the case $k = j - 1$ and $k = j + 1$, the equality

$$\lim_{z \rightarrow z_p} (z - z_p) \log(z - z_p) = 0.$$

For the case $k = j$ we have

$$L_{jj} = \frac{1}{2\pi i} \log \frac{z_j - z_{j+1}}{z_j - z_{j-1}}$$

where we take again the principal determination of the logarithmic function.

We will consider now the convergence problem of this (CVBEM) procedure. Precisely, we will determine under which conditions $\tilde{f}(\zeta) \rightarrow f(\zeta)$ holds and, more, when $f^*(z) \rightarrow f(z)$ is valid.

Definition. A grid $d := z_0, z_1, \dots, z_n$ ($z_0 \equiv z_n$) of the closed contour C is called “acceptable” if, for any $\zeta \in C_j, j = \overline{1, n}$, the condition

$$\max \{ |\zeta - z_j|, |\zeta - z_{j-1}| \} < |z_j - z_{j-1}|.$$

is fulfilled.

Let now d be an acceptable grid on the boundary C and let $\delta = \max |z_j - z_{j-1}|$ be the *norm* of this grid. Then we have the following theorems:

THEOREM 6.18. *If $\tilde{f}(\zeta) = \sum_{j=1}^n f_j L_j(\zeta)$ is a “piecewise linear” Lagrange approximation (i.e., constructed on each C_j of the contour C , as above) of the function $f(\zeta)$, with respect to an acceptable grid d , of norm δ , then*

$$\lim_{\delta \rightarrow 0} \tilde{f}(\zeta) = f(\zeta), \forall \zeta \in C$$

and

THEOREM 6.19. *If $f^*(z) = \sum_{j=1}^n f_j \tilde{L}_j(z)$ is a “piecewise linear” Lagrange approximation of the function $f(z)$, with respect to an acceptable grid d , of norm δ , then*

$$\lim_{\delta \rightarrow 0} f^*(z) = f(z), \forall z \in D.$$

The proofs of these theorems are immediate, based on the uniform continuity of $f(\zeta)$ [112].

Now we remark that this convergence of the Lagrange approximation should not surprise because it is an approximation on segments which is, generally, a spline function of first order. The result is still valid (based on the same remark) in the case when the “piecewise linear” Lagrange approximations are replaced, on every arc C_j , by arbitrary powers of them [112]. This generalization will be important in the cases of contours with cuspidal points such as the case of profiles with sharp trailing edge. An application of this procedure will be given in the next section.

Obviously, the solving of the finally obtained homogeneous algebraic systems needs supplementary conditions (like an “a priori” given circulation). More details can be found in [112].

3.3 The Motion of a Dirigible Balloon

As a particular application, let us consider the fluid flow produced by the motion of a self-propelled dirigible balloon in a uniform stream of wind whose velocity is “a priori” given.

We assume that both the dirigible motion and the velocity of the wind stream depend explicitly on time and the motion is plane and potential. Neither external forces nor the influence of the ground are considered (the dirigible being all the time at a sufficiently great distance from the ground).

Suppose that the contour of the dirigible is expressed by an explicit equation of the type

$$x+iy = \frac{8}{3 \left[1 - \left(1 - \frac{2}{s+1 \pm i\sqrt{2-s-s^2}} \right)^k \right]} - k, \quad s \in [-2, 1], \quad (6.36)$$

This equation implies, besides the symmetry of the balloon vs. the real axis, the existence of a sharp trailing edge, located at the point of

the abscissa $x = k$, and where the semitangents with the real axis are respectively $\pm k\pi$.

In fact, the above profile is of Karmann–Trefftz type [69], the connection between the parameter μ of the above theory and the just introduced parameter k being given by $\frac{\mu}{1-\mu} = \frac{1-k}{k}$. In the sequel we shall use the value $k = \frac{4}{3}$. As regards the stream of wind (the basic flow in terms of the previous theory), it will be defined by, for instance, the complex velocity

$$w_B(z, t) = u_B - iv_B = (2t + 1) - i3t^2,$$

while the displacement of the dirigible would be defined by $l(t) = -3t^2$, $m(t) = t$, where $t = 0, 1, \dots$ (the successive time instants), see Figure 6.4.

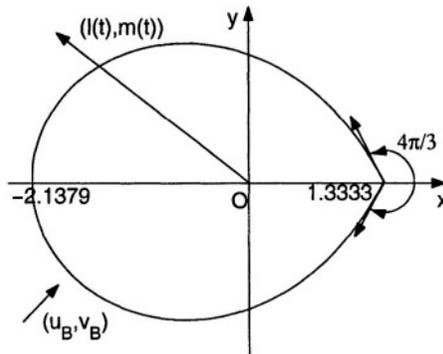


Figure 6.4. The balloon’s profile

The value of the circulation will be established by considering, instead of the flow produced by the dirigible motion, the “dual problem”, i.e., that of an opposite fluid stream of velocity $(-l(t), -m(t))$ past our profile, cumulated with the velocity of the wind (u_B, v_B) . The Jukovski hypothesis leads to the following value of the circulation, $\Gamma = -6\pi(v_B - m)$.

Above we have taken into consideration the fact that the image of our profile through the mapping $\frac{z - 4/3}{z + 4/3} = \left(\frac{Z - 1}{Z + 1}\right)^{4/3}$ is a circumference centered at $(-\frac{1}{2}, 0)$ of radius $\frac{3}{2}$ and whose point $Z = 1$ corresponds to the sharp trailing edge.

The slip condition at the points of the dirigible contour will be written as

$$\frac{v + v_B - m}{u + u_B - l} = \frac{dy}{dx} \Big|_C,$$

where $w = u - iv$ is the looked for complex relative velocity of the fluid flow vs. the system of axes xOy rigidly linked to the profile.

As regards to the nodes z_1, \dots, z_{30} ($z_{31} \equiv z_1$) chosen counterclockwise on the contour of the dirigible, they are obtained by allowing the real parameter s of (6.36) to take the values $-2, -1.9, -1.7, -1.4, -1.2, -1, -0.8, -0.6, -0.4, -0.2, 0, 0.5, 0.7, 0.9, 1$. The leading edge is the node $z_1 = -2.1379$ while the trailing edge is the node $z_F = z_{16} = 1.3333$.

By imposing also the additional conditions, which state the equality of the flow velocity at the sharp trailing edge z_F and at its neighboring nodes, i.e., $u_{F-1} = u_F = u_{F+1}$ and $v_{F-1} = v_F = v_{F+1}$ with $F = 16$ (in order to avoid some logarithmic singularities in the calculation of $\tilde{L}_F(z_{F-1})$ and $\tilde{L}_F(z_{F+1})$), we are led to the solving of a linear algebraic system of $60+2$ (circulation condition) real equations with 56 unknowns.

Since the slip conditions are written at all the 27 remaining nodes ($j \neq F - 1, F, F + 1$) and at z_{F-1} or z_{F+1} , that means

$$\frac{v_j + v_B - m}{u_j + u_B - l} = \frac{dy}{dx} \Big|_{z=z_j}, \quad j = 1, \dots, 30, \quad j \neq 16, 17$$

and by the elimination of v_j in the favor of u_j , we are led again to an overdetermined nonhomogeneous system but this time of 62 equations with 29 unknowns.

By solving this system we find its unique solution i.e., we find $u_j - iv_j = \tilde{w}(z_j)$ at the node $z_j \in C$, for $j = 1, \dots, 30$. We can proceed now to the determination of the unknown function

$$w(z) \approx w^*(z) = \sum_{j=1}^{30} w_j \tilde{L}_j(z).$$

This will be done at the mesh points of a squared neighborhood, of size $[-5,5] \times [-5,5]$ of the profile, both the x - and y - steps of the respective mesh being equal to $\frac{1}{3}$, which means 961 points. Finally, the (absolute) velocity of the resultant fluid flow vs. a fixed system of axes will be determined by calculating the vector $\mathbf{V}(u + l, v + m)$ at the same mesh points and at different time moments. For details and figures, see [107].

3.4 Coupling of the Boundary Element Method and the Finite Element Method

The complexity of the practical problems, the simultaneous presence of structures and systems which are much different by their properties, require a proper treatment, a special mobility to manage the computational techniques, all in order to obtain solutions that are as accurate as

possible. Obviously, the envisaged numerical method should also take into account the computational effort, the economical efficiency, short computational time being an essential component for practical mathematical modeling. In this context, if the FEM distinguishes itself as a very good method, easily applicable, for example, in anisotropic problems, it could not be used, without important losses of accuracy, for problems with geometrical singularities (cuts or concave breaks) as they appear, for instance, in the mechanics of “breaks”, being also uncomfortable for unbounded domains or domains with stresses or fluxes concentrations. Conversely, the BEM is recommended for the solving of these last types of problems, the boundary integral formulations containing both the data at infinity (for unbounded domains) and the possibility to model quite exactly and by a minimal effort (adopting a suitable elements system on the boundary with convenable nodes) the possible geometrical singularities.

In what follows, we will sketch some problems of the “coupling” of these two methods, FEM and BEM, for the same practical problem, but involving regions with different properties which require the use of one or the other of the two methods². Of course, the possibility to use boundary elements of higher order allows the “coupling” of the neighboring regions, distinctly treated by the two methods, without loss of continuity. Once the problems are approached in the distinct manner of the two methods, the resulting algebraic systems should be “fitted” in order to obtain a unique system (with the same unknowns). This could be performed either by transforming the FEM region into a boundary element, a real possibility in the case of the use of mixed finite element formulation or, conversely, by transforming the boundary element into an equivalent finite element.

We will develop this idea, confining ourselves to a potential problem in a domain $\Omega = \Omega^1 \cap \Omega^2$. We will study the problem by FEM in Ω^1 and by BEM in Ω^2 , considering, on the common interface Γ_I , both the continuity conditions (the potentials evaluated at the same point of Γ_I by the two methods, in the domains Ω^1 and Ω^2 must be the same) and the equilibrium conditions (the fluxes, the derivatives of the potentials in the direction of the outward normal, evaluated at the same point of Γ_I , by the two methods, in the domains Ω^1 respectively Ω^2 , must be opposite).

Thus let us consider a potential problem in the domain Ω with the boundary $\Gamma = \Gamma_1 \cup \Gamma_2$, governed by the Poisson equation $\Delta u_0 = b$, with

²This section follows the exposure from the book [100], p. 267.

the joined essential boundary conditions ($u_0 = \bar{u}$ on Γ_I) and natural boundary conditions ($q_0 = \bar{q}$ on Γ_2). Constructing then the expression of the weighted residual of this problem, with the supplementary requirement on the approximation u (of the solution u_0) to satisfy identically the essential boundary conditions ($u = \bar{u}$ on Γ_I), we will have, for the region Ω^1 (with FEM),

$$\int_{\Omega} \frac{\partial u}{\partial x_i} \frac{\partial w}{\partial x_i} d\Omega = \int_{\Gamma_1} \bar{q} w d\Gamma - \int_{\Omega} b w d\Omega$$

(the weak formulation with $w = 0$ on Γ_I) while for the region Ω^2 (with BEM),

$$\int_{\Omega} (\nabla^2 u^*) u d\Omega = - \int_{\Gamma_1} q u^* d\Gamma - \int_{\Gamma_2} \bar{q} u^* d\Gamma + \int_{\Gamma_1} \bar{u} q^* d\Gamma + \int_{\Gamma_2} u q^* d\Gamma + \int_{\Omega} b u^* d\Omega$$

(the inverse formulation with $w = u^*$).

If we remain for the moment, in the subdomain Ω^1 , then if the weight functions w (in the weak formulation) are expressed with the same basic functions as the function u , one may apply a finite element type discretization (and a corresponding interpolation) which will lead to a matrix system of the form $\mathbf{KU} = \mathbf{F} + \mathbf{D}$. Here \mathbf{K} is the global matrix of the system (a symmetrical matrix), \mathbf{U} is the corresponding matrix of the unknowns (the values of the potentials at the nodes), \mathbf{F} is the vector constructed with the integrals $\int_{\Gamma_2} \bar{q} w d\Gamma$ and the vector \mathbf{D} corresponds to the integrals $-\int_{\Omega} b w d\Omega$.

Finally, if the above inverse formulation is used in Ω^2 , we will obtain, on the boundary $\tilde{\Gamma}$ of this subdomain, the integral equation

$$c u + \int_{\tilde{\Gamma}} q^* u d\Gamma = \int_{\tilde{\Gamma}} u^* q d\Gamma + \int_{\Omega^2} b u^* d\Omega.$$

But this integral equation underlies a BEM in Ω^2 and by its application we will come, finally, to a matrix system of the form $\mathbf{HU} = \mathbf{GQ} + \mathbf{B}$ where the unknowns are grouped into the vectors \mathbf{U} (the nodal boundary values of the potential) and \mathbf{Q} (the nodal boundary values of the derivatives of the potential with respect to the outward normal, the fluxes). Concerning the known matrices \mathbf{H} and \mathbf{G} , their form depends on the fundamental solution and the chosen interpolating functions space while the vector \mathbf{B} is constructed starting from the integrals $\int_{\Omega^2} b u^* d\Omega$.

In order to “match” the distinct algebraic systems obtained in Ω^1 , respectively in Ω^2 , to assemble them into a unique system, we will transform the region Ω^2 into an equivalent finite element. In other words, we will try to rewrite the matrix system obtained for Ω^2 in a form identical with that of the system obtained for Ω^1 .

Remarking that the vector \mathbf{F} was obtained by multiplying the given fluxes by the interpolating functions used for the weight, one could always find a matrix N , called the distribution matrix, so that $\mathbf{F} = \mathbf{N}\mathbf{Q}$ where \mathbf{Q} is the vector containing the unknown values at the boundary nodes of the flux (the derivative, with respect to the outward normal, of the potential). If we write then the system $\mathbf{H}\mathbf{U} = \mathbf{G}\mathbf{Q} + \mathbf{R}$ in the form $\mathbf{G}^{-1}(\mathbf{H}\mathbf{U} - \mathbf{R}) = \mathbf{Q}$ and we multiply both sides of this equality by the distribution matrix \mathbf{N} , the result could be written in a form, specific to Ω^1 , that means $\mathbf{K}^1\mathbf{U} = \mathbf{F}^1 + \mathbf{D}^1$ where $\mathbf{K}^1 = \mathbf{N}\mathbf{G}^{-1}\mathbf{H}$, $\mathbf{D}^1 = \mathbf{N}\mathbf{G}^{-1}\mathbf{R}$, $\mathbf{F}^1 = \mathbf{N}\mathbf{Q}$. Unfortunately, as regards the computational efficiency, the matrix \mathbf{K}^1 is no longer symmetrical as the matrices \mathbf{K} associated to the FEM, are. If we choose to “symmetrize” the matrix $\mathbf{K}^1(k_{ij}^1)$ by replacing it with a symmetric matrix $\mathbf{K}^2(k_{ij}^2)$, we could proceed, for instance, by a simple “error diminishing” technique. Thus, let ε_{ij} be the error — due to the asymmetry — measured by the deviation of the nondiagonal coefficients k_{ij}^1 and k_{ji}^1 , versus the corresponding coefficients k_{ij} (yet unknowns)) and equals to $\varepsilon_{ij} = \frac{1}{2} \left[(k_{ij} - k_{ij}^1) + (k_{ij} - k_{ji}^1) \right]$.

Writing the necessary condition for the minimization of $\varepsilon_{ij}^2(k_{ij})$, we get

$$\frac{\partial (\varepsilon_{ij}^2)}{\partial k_{ij}} = 2k_{ij} - k_{ij}^1 - k_{ji}^1 = 0,$$

which means the coefficients k_{ij} are given by $k_{ij} = \frac{1}{2} [k_{ij}^1 + k_{ji}^1]$, so that the symmetric matrix \mathbf{K}^2 is $\mathbf{K}^2 = \frac{1}{2} [\mathbf{K}^1 + \mathbf{K}^{1,T}]$.

Correspondingly, the system $\mathbf{K}^1\mathbf{U} = \mathbf{F}^1\mathbf{U} + \mathbf{D}^1$ is rewritten under the form $\mathbf{K}^2\mathbf{U} = \mathbf{F}^2 + \mathbf{D}^2$ which will be assembled into a usual manner (ensuring the compatibility and equilibrium conditions on the interface Γ_1) with the system, of the same type, from Ω^1 .

A direct procedure to obtain the symmetric matrix \mathbf{K}^2 could be the so-called “energetic onset”. Starting from the expression of the energy in the domain where the BEM for the potential problem is applied, i.e.,

$$\pi = \frac{1}{2} \int_{\Gamma} qu d\Gamma - \int_{\Gamma_2} \bar{q}u d\Gamma - \int_{\Omega} bud\Omega,$$

the equilibrium requirement leads to

$$\delta\pi = \frac{1}{2} \int_{\Gamma} (q\delta u + \delta q u) d\Gamma - \int_{\Gamma_2} \bar{q} \delta u d\Gamma - \int_{\Omega} b\delta u d\Omega = 0.$$

But the integral equation which gives the solution of the problem, once one knows the values of u and q on the boundary, is

$$u + \int_{\bar{\Gamma}} q^* u d\Gamma = \int_{\bar{\Gamma}} u^* q d\Gamma + \int_{\Omega} b u^* d\Omega,$$

and then, by replacing into the above relation,

$$q \left(\equiv \frac{\partial u}{\partial n} \right) = \frac{\partial}{\partial n} \left\{ - \int_{\bar{\Gamma}} q^* u d\Gamma + \int_{\bar{\Gamma}} u^* q d\Gamma + \int_{\Omega} b u^* d\Omega \right\}$$

we get, in a matrix form and after the introduction of the interpolating functions for u and q , the system

$$\delta \mathbf{U}^T \left\{ \frac{1}{2} (\mathbf{N} \mathbf{C} \mathbf{U} + \mathbf{C}^T \mathbf{N}^T \mathbf{U}) - \mathbf{F} - \mathbf{D} \right\} = 0.$$

But this leads necessarily to a system of the form $\mathbf{K} \mathbf{U} = \mathbf{F} + \mathbf{D}$ where \mathbf{K} is a symmetric matrix given by $\mathbf{K} = \frac{1}{2} (\mathbf{N} \mathbf{C} + \mathbf{C}^T \mathbf{N}^T)$, i.e., to a system which could be assembled with that obtained by FEM. Here we denote by \mathbf{N} the matrix formed by integrating the interpolating functions and by \mathbf{C} the matrix that links \mathbf{Q} and \mathbf{U} ($\mathbf{Q} = \mathbf{C} \mathbf{U}$).

Another coupling procedure of FEM and BEM uses the so-called *approximative boundary elements* or the so-called *Sommerfeld relation*. In order to illustrate this technique we will consider, for simplicity, the case of the Laplace equation for the domain Ω with the boundary Γ whose weighted residual expression (in the inverse formulation) is

$$\int_{\Omega} (\nabla^2 u^*) u d\Omega = \int_{\Gamma} u q^* d\Gamma - \int_{\Gamma} q u^* d\Gamma.$$

Assume that the domain Ω is the outside of a body, which means it is an unbounded domain. Due to some known reasons we apply the FEM only in a finite domain, the outside of the body limited by a spherical interface Γ_I while in the exterior of Γ_I , the BEM will be applied. As the fundamental solution of the Laplace equation is $u^* = \frac{1}{4\pi r}$, the above inverse formulation shows that at any point of the interior region we have

$$\int_{\Gamma_1} uq^* d\Gamma = \int_{\Gamma_1} u^* qd\Gamma,$$

which means

$$\int_{\Gamma_1} \left(\frac{1}{r}\right) qd\Gamma = \int_{\Gamma_1} u \frac{\partial}{\partial n} \left(\frac{1}{r}\right) d\Gamma.$$

This last relation establishes the link between the boundary values of \mathbf{u} and \mathbf{q} and it could replace the boundary integral equation within the use of the BEM in the exterior of Γ_I . But the result of the application of this integral relation will be a system of equations with a “non-band” matrix. If we choose the radius R of the interface Γ_I large enough that the above integral relation, written under the form

$$\int_0^{2\pi} \int_0^{2\pi} \left\{ \frac{1}{R} \frac{\partial u}{\partial R} + u \left(\frac{1}{R^2} \right) \right\} R^2 \sin \varphi d\theta d\varphi = 0,$$

could be approximated by $r \frac{\partial u}{\partial r} + u = 0$ on Γ_I , using also the above link between \mathbf{u} and \mathbf{q} at the points of Γ_I , we will come to a system of equations with a “band” matrix as in the FEM.

The above approximate relation, at the points of Γ_I , is a relation of Sommerfeld type. The establishment of such relations is important especially for problems with complicated fundamental solutions. Thus, if we consider, in an unbounded domain Ω , the Helmholtz bidimensional equation $\nabla^2 u + k^2 u = 0$ with the fundamental solution $u^* = \frac{i}{4} H_0^{(1)}(kr)$ (where $H_0^{(1)}$ is the Hankel function), the integral Sommerfeld equation on the boundary is obviously difficult to manage. But, concurrently, the Sommerfeld relation on the boundary Γ with radius R large enough, is $\frac{\partial u}{\partial r} - ik u = 0$ which essentially simplifies the calculations.

Chapter 7

THE FINITE VOLUME METHOD AND THE GENERALIZED DIFFERENCE METHOD

The finite volume method is, probably, the most popular discretization method used in CFD. It is similar, in some aspects, to the finite differences method while the discretization procedure is linked to the finite element method. More precisely, the discretization is performed by transforms joined to the physics of the studied phenomenon and conserving some quantities during numerical computations. For this, one uses often the integral formulation of the conservation laws.

The physical domain is considered divided into cells. Between the time instants t_n and t_{n+1} the variation of some physical quantity, for example of the mass, in a cell C_j , denoted by

$$mass_j = vol(C_j) * density_j$$

is given by the sum of the flow fields $flux_{jk}$ between C_j and the neighboring cells C_k , namely

$$mass_j^{(n+1)} = mass_j^{(n)} + \sum_{k \in V(j)} flux_{jk}.$$

The total mass conservation is ensured by the conditions

$$flux_{jk} = -flux_{kj}$$

The finite differences method allows high order approximation schemes with a reasonable computing effort. However these schemes are difficult to apply on domains with a complicated geometry or complicated boundary conditions.

The finite element method works very well on domains with complex geometry and has a well founded theory but it needs more calculations for the same accuracy.

The *finite volume method* combines the simplicity of the finite differences methods with the local accuracy of the finite element method. It allows the use of a flexible mesh with small geometrical errors. The computational effort is greater than in classical finite difference methods and less than in the finite element method for a similar accuracy. At the same dimension of the discretized problem, the accuracy is higher than with finite differences and nearly the same as with finite elements. The theory is elaborated, the variational form of the problems connects the theory and the algorithms of finite element and finite differences methods.

In 1953 R. H. MacNeal used integral interpolation methods to establish difference schemes on irregular networks. After many years, A. M. Winslow and other researchers (1967, 1973) employed the linear finite elements to construct difference schemes on arbitrary triangulations, using the circumcenter dual grid and also the barycenter dual grid. At the end of 1970 some computational fluid researchers (S. V. Patankar [99] among others) proposed to apply the difference method on irregular networks to the computation of compressible and incompressible fluid flows. Due to its many advantages this method developed rapidly, becoming one of the most efficient methods for fluid computations. The researchers called it the *finite volume method* (FVM) or finite control volume method, indicating that it is a discrete approximation of the control equations in an integral form.

In 1978, R. Li, using finite element spaces and generalized characteristic functions on dual elements, rewrote integral interpolation methods in a form of generalized Galerkin methods and thus obtained the so-called *generalized difference methods* (GDM). This method is basically an extension of the finite volume method (i.e., with piecewise constant and piecewise linear elements the two methods are, in fact equivalent) and provides a useful theoretical basis for it.

1. ENO Finite Volume Schemes

ENO (Essentially Non-Oscillatory) schemes are high order accurate schemes designed for problems with piecewise smooth solutions containing discontinuities. The use of the finite volume method to construct numerical schemes for nonlinear conservative equations allows the generalization of the classical difference schemes to arbitrary grids.

The key idea is to use a nonlinear adaptive procedure to automatically choose the locally smoothest stencil and avoid crossing discontinuities in the interpolation procedure. ENO schemes are quite successful in computational fluid dynamics especially for problems containing shocks.

In the sequel we will shortly present some ENO finite volume schemes, following [137].

These schemes are based on interpolation of discrete data, by using algebraic polynomials. Traditional finite volume methods are based on fixed stencil interpolations. For example, the interpolation for the cell i uses the cells $i-1, i, i+1$ to build a second order interpolation polynomial, i.e., the cell i plus one cell to the right and one cell to the left. This works well for globally smooth problems but it is oscillatory (the Gibbs phenomenon) near a discontinuity and such oscillations do not decay in magnitude when the mesh is refined.

Earlier attempts to eliminate or reduce these spurious oscillations were mainly based on the explicit artificial viscosity and limiters. The artificial viscosity must be large enough near discontinuity to reduce the oscillations but small elsewhere to maintain a high-order accuracy, so it is problem dependent. The limiters eliminate the oscillations by reducing the order of accuracy of the interpolant near the discontinuity but the accuracy degenerates also near smooth extrema.

ENO schemes were first introduced by Harten, Engquist, Osher and Chakravarthy in 1987 [62]. Today their study is very active and most of the problems solved have solutions containing strong shocks and rich smooth region structures, so that lower order methods usually have difficulties.

1.1 ENO Finite Volume Scheme in One Dimension

Let us consider the one-dimensional conservation law

$$\frac{\partial}{\partial t} u(x, t) + \frac{\partial}{\partial x} f(u(x, t)) = 0 \tag{7.1}$$

with suitable initial and boundary conditions. We will discretize only the spatial variable x and will leave the time variable t to be continuous for the moment.

The computational domain is $a \leq x \leq b$. We consider the grid

$$a = x_{1/2} < x_{3/2} < \dots < x_{N-1/2} < x_{N+1/2} = b \tag{7.2}$$

and we define cells, cell centers and cell sizes respectively by

$$I_i = [x_{i-1/2}, x_{i+1/2}], \quad x_i = \frac{1}{2} (x_{i-1/2} + x_{i+1/2}), \quad h_i = x_{i+1/2} - x_{i-1/2}$$

for $i = 1, 2, \dots, N$. We denote the maximum cell size by $h = \max_{1 \leq i \leq N} h_i$.

We assume that the values of the numerical solution are also available outside the computational domain whenever they are needed (this is the case, for example, for periodic or compactly supported problems).

First of all we must solve the following problem (reconstruction):

Problem 7.1. (One-dimensional reconstruction)

Given the cell averages of a function $v(x)$

$$\bar{v}_i = \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} v(\xi) d\xi, \quad i = 1, 2, \dots, N \quad (7.3)$$

find a polynomial $p_i(x)$, of degree at most $k - 1$, for each cell I_i , such that it is a k -th order accurate approximation to the function $v(x)$ inside I_i , i.e.,

$$p_i(x) = v(x) + O(h^k), \quad x \in I_i, \quad i = 1, \dots, N. \quad (7.4)$$

In particular we obtain approximations to the function $v(x)$ at the cell boundaries

$$v_{i+1/2}^- = p_i(x_{i+1/2}), \quad v_{i-1/2}^+ = p_i(x_{i-1/2}), \quad i = 1, \dots, N.$$

In order to solve this problem, we consider a cell I_i and an order of accuracy k . We choose a stencil based on r cells to the left, s cells to the right and I_i itself ($r, s \geq 0, r + s + 1 = k$),

$$S(i) = \{I_{i-r}, \dots, I_{i+s}\}.$$

There is a unique polynomial $p(x)$ of degree at most $k - 1 = r + s$, whose cell average in each of the cells in $S(i)$ agrees with that of $v(x)$,

$$\frac{1}{h_j} \int_{x_{j-1/2}}^{x_{j+1/2}} p(\xi) d\xi = \bar{v}_j, \quad j = i - r, \dots, i + s.$$

This polynomial $p(x)$ is the approximation we are looking for, as long as the function $v(x)$ is smooth in the region covered by the stencil $S(i)$ (see the complete proof in [137]).

Consequently, given the k cell averages

$$\bar{v}_{i-r}, \dots, \bar{v}_{i-r+k-1},$$

there are constants c_{rj} such that the reconstructed value at the cell boundary $x_{i+1/2}$,

$$v_{i+1/2} = \sum_{j=0}^{k-1} c_{rj} \bar{v}_{i-r+j},$$

is k -th order accurate

$$v_{i+1/2} = v(x_{i+1/2}) + O(h^k).$$

For a uniform grid ($h_i = h$), the coefficients c_{rj} do not depend on i or h and we have, for example, for $k = 4$ and $r = 1$,

$$v_{i+1/2} = -\frac{1}{12}\bar{v}_{i-1} + \frac{7}{12}\bar{v}_i + \frac{7}{12}\bar{v}_{i+1} - \frac{1}{12}\bar{v}_{i+2}.$$

The second problem to solve is how to choose the stencils. We are interested in the class of piecewise smooth functions, i.e., functions which have as many derivatives as the scheme calls for, everywhere except for at finitely many isolated discontinuity points, where the function $v(x)$ and its derivatives are assumed to have finite left and right limits.

For piecewise smooth functions $v(x)$, a fixed stencil approximation may not be adequate near discontinuities. If the stencils contain a discontinuous cell for x_i close enough to a discontinuity, the Gibbs phenomenon happens and the approximation property (7.4) is no longer valid. The basic idea is to avoid including the discontinuous cells in the stencil (if possible), by using an adaptive stencil, i.e., the left shift r changes with the location x_i .

Let us consider the primitive function of $v(x)$,

$$V(x) = \int_{-\infty}^x v(\xi) d\xi$$

(where the $-\infty$ limit can be replaced by any fixed number) and we have, obviously,

$$V(x_{i+1/2}) = \sum_{j=-\infty}^i \int_{x_{j-1/2}}^{x_{j+1/2}} v(\xi) d\xi = \sum_{j=-\infty}^i \bar{v}_j h_j. \tag{7.5}$$

Thus we know exactly the primitive function $V(x)$ at the cell boundaries. If we denote by $P(x)$ the unique polynomial of degree at most k , which interpolates $V(x_{j+1/2})$ at the $k + 1$ points

$$x_{i-r-1/2}, \dots, x_{i+s+1/2},$$

then its derivative is the above polynomial $p(x)$.

Let us define the 0-th degree divided difference of the function $V(x)$ by

$$V[x_{i-1/2}] = V(x_{i-1/2}).$$

Then, the j -th degree divided differences, for $j \geq 1$, are defined inductively by

$$V[x_{i-1/2}, \dots, x_{i+j-1/2}] = \frac{V[x_{i+1/2}, \dots, x_{i+j-1/2}] - V[x_{i-1/2}, \dots, x_{i+j-3/2}]}{x_{i+j-1/2} - x_{i-1/2}}.$$

Similarly, the divided differences of the cell averages \bar{v} (7.3) are defined by

$$\bar{v}[x_i] = \bar{v}_i$$

and in general

$$\bar{v}[x_i, \dots, x_{i+j}] = \frac{\bar{v}[x_{i+1}, \dots, x_{i+j}] - \bar{v}[x_i, \dots, x_{i+j-1}]}{x_{i+j} - x_i}.$$

But from (7.5) we have

$$V[x_{i-1/2}, x_{i+1/2}] = \frac{V(x_{i+1/2}) - V(x_{i-1/2})}{x_{i+1/2} - x_{i-1/2}} = \bar{v}_i$$

so that we can write the divided differences of $V(x)$ in terms of \bar{v} and completely avoid the computation of V .

The Newton form of $P(x)$ is

$$P(x) = \sum_{j=0}^k V[x_{i-r-1/2}, \dots, x_{i-r+j-1/2}] \prod_{m=0}^{j-1} (x - x_{i-r+m-1/2})$$

so that

$$p(x) = P'(x) = \sum_{j=0}^k V[x_{i-r-1/2}, \dots, x_{i-r+j-1/2}] \sum_{m=0}^{j-1} \prod_{\substack{l=0 \\ l \neq m}}^{j-1} (x - x_{i-r+l-1/2}). \tag{7.6}$$

Of course, we can express $p(x)$ completely in terms of \bar{v} .

An important property of divided differences is

$$V[x_{i-1/2}, \dots, x_{i+j-1/2}] = \frac{V^{(j)}(\xi)}{j!}$$

for some $\xi \in (x_{i-1/2}, x_{i+j-1/2})$ as long as $V(x)$ is smooth in this stencil. If $V(x)$ is discontinuous at some point inside the stencil, we have

$$V[x_{i-1/2}, \dots, x_{i+j-1/2}] = O\left(\frac{1}{h^j}\right).$$

thus the divided difference is a measurement of the smoothness of V .

Finally, the ENO reconstruction procedure is the following :

Algorithm 7.1. (ENO reconstruction)

Given the cell averages $\{\bar{v}_i\}$ of a function $v(x)$, we obtain a piecewise polynomial reconstruction, of degree at most $k - 1$, by

1. Computing the divided differences of the primitive function $V(x)$, for degrees 1 to k using \bar{v} ;
2. Starting in the cell I_i with a two-point stencil

$$\tilde{S}_2(i) = \{x_{i-1/2}, x_{i+1/2}\}$$

for $V(x)$ (which is equivalent to a one-point stencil $S_1(i) = \{I_i\}$ for \bar{v} ;

3. For $l = 2, \dots, k$, assuming

$$\tilde{S}_l(i) = \{x_{j+1/2}, \dots, x_{j+l-1/2}\}$$

is known, add one of the neighboring points to the stencil following:

–if

$$|V[x_{j-1/2}, \dots, x_{j+l-1/2}]| < |V[x_{j+1/2}, \dots, x_{j+l+1/2}]|$$

add $x_{j-1/2}$ to the stencil $\tilde{S}_l(i)$ to obtain

$$\tilde{S}_{l+1}(i) = \{x_{j-1/2}, \dots, x_{j+l-1/2}\},$$

–otherwise, add $x_{j+l+1/2}$ to the stencil $\tilde{S}_l(i)$ to obtain

$$\tilde{S}_{l+1}(i) = \{x_{j+1/2}, \dots, x_{j+l+1/2}\};$$

4. Use (7.6) to obtain $p_i(x)$ and use it to get the approximations at the cell boundaries

$$v_{i+1/2}^- = p_i(x_{i+1/2}), \quad v_{i-1/2}^+ = p_i(x_{i-1/2}),$$

The finite volume schemes are based on cell averages so we do not solve (7.1) directly but its integrated version. If we integrate over I_i we obtain

$$\frac{d}{dt} \bar{u}(x_i, t) = -\frac{1}{h_i} (f(u(x_{i+1/2}, t)) - f(u(x_{i-1/2}, t))) \quad (7.7)$$

where

$$\bar{u}(x_i, t) = \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u(\xi, t) d\xi$$

is the cell average. We approximate the equation (7.7) by the conservative scheme

$$\frac{d}{dt} \bar{u}_i(t) = -\frac{1}{h_i} (\hat{f}_{i+1/2} - \hat{f}_{i-1/2}) \quad (7.8)$$

where $\bar{u}_i(t)$ is the numerical approximation to the cell average $\bar{u}(x_i, t)$ while the numerical flux $\hat{f}_{i+1/2}$ is defined by

$$\hat{f}_{i+1/2} = h (u_{i+1/2}^-, u_{i+1/2}^+). \quad (7.9)$$

Here the values $u_{i+1/2}^\pm$ are obtained by Algorithm 7.1 (ENO reconstruction).

The above function h is a monotone flux, satisfying:

- $h(a, b)$ is a Lipschitz continuous function in both arguments;
- $h(a, b)$ is a nondecreasing function in a and a nonincreasing function in b ;
- $h(a, b)$ is consistent with the physical flux f , i.e., $h(a, a) = f(a)$.

Some examples of monotone fluxes are:

1. Godunov flux

$$h(a, b) = \begin{cases} \min_{a \leq u \leq b} f(u) & \text{if } a \leq b \\ \max_{b \leq u \leq a} f(u) & \text{if } a > b \end{cases} ; \quad (7.10)$$

2. Engquist–Osher flux

$$h(a, b) = \int_0^a \max(f'(u), 0) du + \int_0^b \min(f'(u), 0) du + f(0); \quad (7.11)$$

3. Lax–Friedrichs flux

$$h(a, b) = \frac{1}{2} [f(a) + f(b) - \alpha (b - a)], \quad (7.12)$$

where $\alpha = \max_u |f'(u)|$ is a constant and the max is taken over the relevant range of u .

Consequently, an ENO finite volume scheme is the following

Algorithm 7.2. (ENO finite volume scheme)

Given the cell averages $\{\bar{u}_i\}$,

1. Follow Algorithm 7.1 to obtain the k -th order reconstructed values $u_{i+1/2}^-$ and $u_{i-1/2}^+$ for all i ;

2. Choose a monotone flux and use (7.9) to compute the flux $\hat{f}_{i+1/2}$ for all i ;

3. Form the scheme (7.8).

The time discretization of an initial value problem for the system (7.8) can be performed by various methods, like Runge-Kutta or multi-step. Another way to discretize the time variable in the equation (7.1) is by the Lax–Wendroff procedure.

We start from the Taylor series expansion in time

$$u(x, t + \Delta t) = u(x, t) + \Delta t \frac{\partial}{\partial t} u(x, t) + \frac{\Delta t^2}{2} \frac{\partial^2}{\partial t^2} u(x, t) + \dots \quad (7.13)$$

Then we use the original equation (7.1) to replace the time derivatives by the spatial derivatives

$$\frac{\partial}{\partial t}u(x, t) = -\frac{\partial}{\partial x}f(u(x, t)) = -f'(u(x, t))\frac{\partial}{\partial x}u(x, t), \tag{7.14}$$

$$\begin{aligned} \frac{\partial^2}{\partial t^2}u(x, t) &= 2f'(u(x, t))f''(u(x, t))\left(\frac{\partial}{\partial x}u(x, t)\right)^2 \\ &\quad + (f'(u(x, t)))^2\frac{\partial^2}{\partial x^2}u(x, t). \end{aligned}$$

We substitute now these derivatives into (7.13) and discretize the spatial derivatives of $u(x, t)$ by an ENO finite volume scheme, for example.

Actually, we first integrate (7.1) in space-time over $[x_{i-1/2}, x_{i+1/2}] \times [t_n, t_{n+1}]$ to obtain

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{1}{h_i} \left(\int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t))dt - \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t))dt \right)$$

Now we use a Gaussian quadrature to discretize the time integrations

$$\frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t))dt \approx \sum_k w_k f(u(x_{i+1/2}, t_n + \alpha_k \Delta t))$$

where w_k and α_k are respectively the Gaussian quadrature nodes and weights. Finally we replace each

$$f(u(x_{i+1/2}, t_n + \alpha_k \Delta t))$$

by a monotone flux

$$h(u(x_{i+1/2}^-, t_n + \alpha_k \Delta t), u(x_{i+1/2}^+, t_n + \alpha_k \Delta t))$$

and use (7.13) and (7.14) to convert

$$u(x_{i+1/2}^\pm, t_n + \alpha_k \Delta t)$$

to $u(x_{i+1/2}^\pm, t_n)$ and its spatial derivatives also at t_n . The derivatives can be obtained by using the reconstructions $p(x)$ inside I_i and I_{i+1} . We remark that each derivative of $p(x)$ is one order lower in accuracy but this is compensated by the presence of Δt in front of it in (7.13).

1.2 ENO Finite Volume Scheme in Multi-Dimensions

In this case we consider the 2D conservation law

$$\frac{\partial}{\partial t}u(x, y, t) + \frac{\partial}{\partial x}f(u(x, y, t)) + \frac{\partial}{\partial y}g(u(x, y, t)) = 0 \quad (7.15)$$

with initial and boundary conditions. Of course, most of the considerations are also valid for higher dimensions.

First we describe how the reconstruction and approximation are generalized to higher dimension spaces. Now we have two cases:

a) structured meshes, where the computational (spatial) domain is a rectangle $[a, b] \times [c, d]$, covered by the cells

$$I_{ij} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}], \quad 1 \leq i \leq N_x, \quad 1 \leq j \leq N_y$$

where

$$\begin{aligned} a &= x_{1/2} < x_{3/2} < \cdots < x_{N_x-1/2} < x_{N_x+1/2} = b, \\ c &= y_{1/2} < y_{3/2} < \cdots < y_{N_y-1/2} < y_{N_y+1/2} = d. \end{aligned}$$

The centers of the cells and the grid sizes are

$$\begin{aligned} (x_i, y_j), \quad x_i &= \frac{1}{2} (x_{i-1/2} + x_{i+1/2}), \quad y_j = \frac{1}{2} (y_{j-1/2} + y_{j+1/2}), \\ \Delta x_i &= x_{i+1/2} - x_{i-1/2}, \quad i = 1, \dots, N_x, \\ \Delta y_j &= y_{j+1/2} - y_{j-1/2}, \quad j = 1, \dots, N_y. \end{aligned}$$

We denote as above the maximum grid sizes by

$$\Delta x = \max_{1 \leq i \leq N_x} \Delta x_i, \quad \Delta y = \max_{1 \leq j \leq N_y} \Delta y_j, \quad \Delta = \max(\Delta x, \Delta y)$$

and assume that Δx and Δy are of the same order of magnitude during refinements.

b) unstructured meshes, where the computational (spatial) domain is covered by a triangulation with N triangles (for example)

$$\{\Delta_0, \Delta_1, \dots, \Delta_N\}$$

where we denote by $|\Delta_i|$ the area of the triangle Δ_i and we use again Δ to denote a typical “length” of the triangles, for example the longest side of the triangle.

The corresponding reconstruction problem in the rectangular case is:
Problem 7.2. (Two dimensional reconstruction for rectangles)

Given the cell averages of a function $v(x, y)$,

$$\bar{v}_{ij} = \frac{1}{\Delta x_i \Delta y_j} \int_{y_{j-1/2}}^{y_{j+1/2}} \int_{x_{i-1/2}}^{x_{i+1/2}} v(\xi, \eta) d\xi d\eta, \quad i = 1, \dots, N_x, \quad j = 1, \dots, N_y,$$

find a polynomial $p_{ij}(x, y)$, preferably of degree at most $k - 1$ (in each variable), for each cell I_{ij} , such that it is a k -th order accurate approximation to the function $v(x, y)$ inside I_{ij} ,

$$p_{ij}(x, y) = v(x, y) + O(\Delta^k), \quad (x, y) \in I_{ij}, \quad \forall i, j.$$

In particular, this gives the approximations to the function $v(x, y)$ at the cell boundaries

$$v_{i+1/2, y}^- = p_{ij}(x_{i+1/2}, y), \quad v_{i-1/2, y}^+ = p_{ij}(x_{i-1/2}, y), \\ i = 1, \dots, N_x, \quad y_{j-1/2} \leq y \leq y_{j+1/2},$$

$$v_{x, j+1/2}^- = p_{ij}(x, y_{j+1/2}), \quad v_{x, j-1/2}^+ = p_{ij}(x, y_{j-1/2}), \\ j = 1, \dots, N_y, \quad x_{i-1/2} \leq x \leq x_{i+1/2},$$

which are k -th order accurate.

In order to solve the problem, if we consider a location I_{ij} and the order of accuracy k , we again choose a stencil $S(i, j)$ based on $k(k+1)/2$ neighboring cells and we try to find a polynomial $p(x, y)$ of degree at most $k - 1$ whose cell average in each of the cells of $S(i, j)$ agrees with that of $v(x, y)$. We remark that in 2D there are many more candidate stencils than in the 1D case and, unfortunately, not all the candidate stencils can be used to obtain the polynomial p (neither existence nor uniqueness automatically holds).

For rectangular meshes, however, we can proceed as in 1D, using the tensor product stencils

$$S_{rs}(i, j) \\ = \{I_{lm} \mid i - r - 1 \leq l \leq i + k - r - 1, \quad j - s - 1 \leq m \leq j - s - 1 + k\}.$$

Then we introduce the primitive

$$V(x, y) = \int_{-\infty}^y \int_{-\infty}^x v(\xi, \eta) d\xi d\eta$$

and, obviously, we have as in the 1D case,

$$V(x_{i+1/2}, y_{j+1/2}) = \sum_{m=-\infty}^j \sum_{l=-\infty}^i \bar{v}_{lm} \Delta x_l \Delta y_m,$$

i.e., with the knowledge of the cell averages \bar{v} we know exactly the primitive function V at cell corners.

Now, on each tensor product stencil

$$\tilde{S}_{rs}(i, j) = \left\{ (x_{l+1/2}, y_{m+1/2}) \left| \begin{array}{l} i - r - 1 \leq l \leq i - r - 1 + k, \\ j - s - 1 \leq m \leq j - s - 1 + k \end{array} \right. \right\},$$

there is a unique polynomial $P(x, y)$ of degree at most k in each variable which interpolates V at every point in $\tilde{S}_{rs}(i, j)$. Finally, we get the solution of Problem 7.2

$$p(x, y) = \frac{\partial^2 P(x, y)}{\partial x \partial y}.$$

Practically, we first perform a 1D reconstruction (Problem 7.1) in the y direction, obtaining one-dimensional cell averages of v in x direction and then we perform the reconstruction also in the x direction. Of course, the cost of this kind of reconstruction is very high. If the cost to perform a 1D reconstruction is c , then for nD reconstruction we need nc per grid point.

The reconstruction problem in the triangular case is

Problem 7.3. (Two dimensional reconstruction for triangles)

Given the cell averages of a function $v(x, y)$,

$$\bar{v}_i = \frac{1}{|\Delta_i|} \int_{\Delta_i} v(\xi, \eta) d\xi d\eta, \quad i = 1, \dots, N,$$

find a polynomial $p_i(x, y)$ of degree at most $k - 1$, for each triangle Δ_i such that it is a k -th order accurate approximation to the function v inside Δ_i .

In particular, p gives approximations to the function v at the triangle boundaries, which are needed in forming the finite volume schemes.

The general procedure to solve this problem is the following. Once given the location Δ_i and the order of accuracy k , we first choose a stencil $S(i)$ based on $m = k(k + 1)/2$ neighboring triangles and then we try to find a polynomial $p(x, y)$ of degree at most $k - 1$, whose cell average in each of the triangle in $S(i)$ agrees with that of $v(x, y)$. If the given $m \times m$ linear system has a unique solution, $S(i)$ is called an *admissible stencil*.

Of course, the reconstruction is performed using only such admissible stencils and this procedure is essentially two dimensional.

In the sequel we describe the ENO finite volume schemes for the 2D conservation law (7.15). Again, we do not solve directly this equation but we focus on its integrated version.

For a structured mesh, we integrate (7.15) over the cell I_{ij} to obtain

$$\begin{aligned} & \frac{d\bar{u}_{ij}(t)}{dt} \\ = & -\frac{1}{\Delta x_i \Delta y_j} \left(\int_{y_{j-1/2}}^{y_{j+1/2}} f(u(x_{i+1/2}, y, t)) dy - \int_{y_{j-1/2}}^{y_{j+1/2}} f(u(x_{i-1/2}, y, t)) dy \right. \\ & \left. + \int_{x_{i-1/2}}^{x_{i+1/2}} g(u(x, y_{j+1/2}, t)) dx - \int_{x_{i-1/2}}^{x_{i+1/2}} g(u(x, y_{j-1/2}, t)) dx \right) \end{aligned} \quad (7.16)$$

where $\bar{u}_{ij}(t)$ is the cell average. We approximate this equation by the conservative scheme

$$\frac{d\bar{u}_{ij}(t)}{dt} = -\frac{1}{\Delta x_i} (\hat{f}_{i+1/2,j} - \hat{f}_{i-1/2,j}) - \frac{1}{\Delta y_j} (\hat{g}_{i,j+1/2} - \hat{g}_{i,j-1/2}).$$

Here, again, the numerical flux $\hat{f}_{i+1/2,j}$ is defined by

$$\hat{f}_{i+1/2,j} = \sum_{\alpha} w_{\alpha} h \left(u_{i+1/2, y_j + \beta_{\alpha} \Delta y_j}^{-}, u_{i+1/2, y_j + \beta_{\alpha} \Delta y_j}^{+} \right) \quad (7.17)$$

where w_{α} and β_{α} are respectively Gaussian quadrature weights and nodes for approximating

$$\frac{1}{\Delta y_j} \int_{y_{j-1/2}}^{y_{j+1/2}} f(u(x_{i+1/2}, y, t)) dy$$

and $u_{i+1/2, y}^{\pm}$ are the k -th order accurate reconstructed values obtained by the following ENO reconstruction.

ENO reconstruction. We use the one-dimensional ENO reconstruction Algorithm 7.1 on the two-dimensional cell averages in the y (or x) direction to obtain one-dimensional cell averages in x (or y). Then, using again the one-dimensional ENO reconstruction in the x (or y) direction, we recover the function itself.

We remark that the superscript $-$ implies the values obtained within the cell I_{ij} and the superscript $+$ implies the cell $I_{i+1,j}$.

The flux $\widehat{g}_{i,j+1/2}$ is defined similarly,

$$\widehat{g}_{i,j+1/2} = \sum_{\alpha} w_{\alpha} h \left(u_{x_i+\beta_{\alpha}\Delta x_i,j+1/2}^{-}, u_{x_i+\beta_{\alpha}\Delta x_i,j+1/2}^{+} \right) \quad (7.18)$$

for approximating

$$\frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} g(u(x, y_{j+1/2}, t)) dx.$$

Here the k -th order accurate ENO reconstruction values $u_{x,j+1/2}^{\pm}$ are obtained as above and h is a one-dimensional monotone flux such as (7.10), (7.11) or (7.12).

Consequently, the ENO finite volume procedure, given the cell averages $\{\bar{u}_{ij}\}$ and the one-dimensional monotone flux h , could be the following:

Algorithm 7.3. (2D ENO finite volume scheme for rectangular mesh)

1. Follow the above ENO reconstruction procedure to obtain the values

$$u_{i+1/2,y_j+\beta_{\alpha}\Delta y_j}^{\pm}, \quad u_{x_i+\beta_{\alpha}\Delta x_i,j+1/2}^{\pm}$$

at the Gaussian nodes;

2. Calculate the flux $\widehat{f}_{i+1/2,j}$ using (7.17) and the flux $\widehat{g}_{i,j+1/2}$ using (7.18);

3. Form the scheme (7.16).

The time discretization works as in the 1D case. If the geometry cannot be covered by a Cartesian grid, the computational domain can be mapped smoothly to a rectangle by the transforms

$$\xi = \xi(x, y), \quad \eta = \eta(x, y)$$

leading to

$$v_x = v_{\xi}\xi_x + v_{\eta}\eta_x$$

for example. The smoothness of ξ_x and η_x guarantees a high order approximation to v_x and the above scheme is still conservative.

Unfortunately, this 2D ENO finite volume scheme for rectangular mesh is very expensive and this is why multidimensional finite volume schemes of order of accuracy higher than 2 are rarely used for a structured mesh. Finite difference versions of such schemes are much more economical for these cases.

One advantage of the ENO finite volume method is that it can be defined on arbitrary meshes, provided that an ENO reconstruction on that

mesh is available. Consequently, adaptive algorithms can be formulated and therefore the cost could be greatly reduced.

Let us discuss now the case of unstructured meshes, using a two-dimensional ENO reconstruction. Taking the triangle Δ_i as a control volume, the semi-discrete finite volume scheme for the equation (7.15) is

$$\frac{d\bar{u}_i(t)}{dt} + \frac{1}{|\Delta_i|} \int_{\partial\Delta_i} \mathbf{F} \cdot \mathbf{n} ds = 0$$

where $\bar{u}_i(t)$ is the cell average, $\mathbf{F} = (f, g)^T$ and \mathbf{n} is the outward unit normal of the triangle boundary $\partial\Delta_i$.

The line integral is discretized by a q -point Gaussian integration formula

$$\int_{\Gamma_k} \mathbf{F} \cdot \mathbf{n} ds \approx |\Gamma_k| \sum_{j=1}^q \omega_j \mathbf{F}(u(G_j, t)) \cdot \mathbf{n}$$

where $\mathbf{F}(u(G_j, t)) \cdot \mathbf{n}$ is replaced by a one-dimensional numerical flux in the \mathbf{n} direction, any of (7.10), (7.11) or (7.12). For example, the Lax–Friedrichs flux yields

$$\int_{\Gamma_k} \mathbf{F} \cdot \mathbf{n} ds \approx \frac{1}{2} [(\mathbf{F}(u^-(G_j, t)) + \mathbf{F}(u^+(G_j, t))) \cdot \mathbf{n} - \alpha (u^+(G_j, t) - u^-(G_j, t))]$$

where α is an upper bound for $|\mathbf{F}'(u) \cdot \mathbf{n}|$. Here u^- and u^+ are the reconstructed values of u inside the triangle and outside the triangle (inside the neighboring triangle) at the Gaussian points, see [1], [138].

2. Generalized Difference Method

In the sequel we will shortly present the GDM, following [83].

2.1 Two-Point Boundary Value Problems

We will illustrate the principle of this generalization of the finite volume method by studying the simple case of a two-point boundary value problem. Consider the problem

$$Lu \equiv -\frac{d}{dx} \left(p \frac{du}{dx} \right) + q \frac{du}{dx} + ru = f, \quad x \in I = (a, b), \tag{7.19}$$

$$u(a) = 0, \quad u'(b) = 0$$

where we have both natural and essential boundary conditions. Here we will suppose $p \in C^1(a, b)$, $p \geq p_0 > 0$ and $q, r, f \in C(a, b)$. As we know,

by multiplying the equation with

$$v \in H_E^1(a, b) = \{v \in H^1(a, b) \mid v(a) = 0\}$$

and by integrating by parts, we obtain the variational problem to find the functions $u \in H_E^1(a, b)$ such that

$$a(u, v) = (f, v), \quad \forall v \in H_E^1(a, b) \quad (7.20)$$

where

$$a(u, v) = \int_a^b (pu'v' + qu'v + ruv) dx, \quad (f, v) = \int_a^b f v dx.$$

If the solution of the variational problem (7.20) $u \in C^1[a, b] \cap C^2(a, b)$, then u is also a classical solution.

The Galerkin method consists in choosing a finite dimensional subspace U_h of $H_E^1(a, b)$ and solving of discrete problem of finding $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in U_h. \quad (7.21)$$

The finite element method constructs U_h as a space of piecewise polynomial functions. For the finite volume method the type of problem (7.20) is generalized as follows.

We discretize $[a, b]$ by the grid

$$a = x_0 < x_1 < \dots < x_n = b$$

and we call, as above, the subintervals $I_i = [x_{i-1}, x_i]$ elements. We will denote $I_i^0 = (x_{i-1}, x_i)$ and $T = \{I_i, 1 \leq i \leq n\}$. Let P_r be again the set of the polynomial functions of order less than or equal to r and

$$S_T^{(r)}(a, b) = \{v \in L_2(a, b) \mid v|_{I_i^0} \in P_r, \quad i = 1, \dots, n\}$$

the set of the piecewise polynomial functions with respect to T . Generally, we denote

$$S^{(r)}(a, b) = \bigcup_T S_T^{(r)}(a, b)$$

the space of the piecewise polynomial functions of order less than or equal to r on (a, b) .

The essential boundary condition $v(a) = 0$ is imposed by choosing the subspaces

$$S_{T,E}^{(r)}(a, b) = \{v \in S_T^{(r)}(a, b) \mid v(a+) = 0\}, \\ S_E^{(r)}(a, b) = \bigcup_T S_{T,E}^{(r)}(a, b) \equiv V.$$

As above, we multiply the original equation by $v \in S_T^{(r)}(a, b)$ and we integrate by parts on (a, b) , obtaining

$$(Lu, v) = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (pu'v' + qu'v + ruv) dx - \sum_{i=1}^n pu'v \Big|_{x_{i-1}^+}^{x_i^-} = (f, v)$$

or, shortly,

$$a_T(u, v) = (f, v), \quad \forall v \in V$$

where

$$a_T(u, v) = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (pu'v' + qu'v + ruv) dx - \sum_{i=1}^{n-1} p(x_i)u'(x_i) [v(x_i^+) - v(x_i^-)].$$

Thus we obtain the variational problem to find $u \in H_E^1(a, b) \cap H^2(a, b)$ such that

$$a_T(u, v) = (f, v), \quad \forall v \in V. \tag{7.22}$$

Of course, if u is a solution of the problem (7.22) and $u \in C^1[a, b] \cap C^2(a, b)$, then u is also a classical solution of the problem (7.19).

In the sequel we will simply denote $a(u, v) \equiv a_T(u, v)$. Also, we denote $\sigma(x)$ the Heaviside function

$$\sigma(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases},$$

and $\delta(x)$ the Dirac distribution, which is also the derivative of σ and we use the “formula”

$$\int_{\alpha}^{\beta} g(x)\delta(x)dx = g(0), \quad \alpha < 0 < \beta$$

for all smooth functions g . The piecewise polynomial function $v \in V$ can be expressed as the sum of a continuous function v_1 and a step function v_2 ,

$$v_2 = \sum_{i=1}^{n-1} [v(x_i^+) - v(x_i^-)] \sigma(x - x_i).$$

Consequently, for the functions $u \in H^2(a, b)$, or for functions u with u' continuous, the above formulas will be interpreted in the sense of distributions. Of course, in the case of a function $v \in H_E^1(a, b)$, a_T is reduced to its original definition.

Let us present now the principle of the finite volume method. We will construct the grid T_h ,

$$a = x_0 < x_1 < \cdots < x_n = b$$

and the dual grid T_h^*

$$a = x_0 < x_{1/2} < x_{3/2} < \cdots < x_{n-1/2} < x_n = b$$

where

$$x_{i-1/2} = \frac{x_{i-1} + x_i}{2}, i = 1, 2, \dots, n.$$

We will denote $I_0^* = [x_0, x_{1/2}]$, $I_i^* = [x_{i-1/2}, x_{i+1/2}]$ and $I_n^* = [x_{n-1/2}, x_n]$ the cells of the dual grid. We will choose the trial functions from the space $U_h \subset U \equiv H_E^1(a, b)$ as the space of finite elements with respect to the grid T_h and the test functions from the space $V_h \subset V \equiv S_E^{(r)}(a, b)$ as piecewise polynomial functions, of low order, with respect to the dual grid T^* , from the space $S_{T^*, E}^{(r)}(a, b)$. The discretized form of the variational problem by the finite volume method is to find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h. \quad (7.23)$$

Different choices of U_h, V_h lead to different schemes. Let us describe some particular cases.

2.1.1 The Linear Case

Let us consider the problem

$$Lu \equiv -\frac{d}{dx}\left(p \frac{du}{dx}\right) = f, \quad x \in (a, b),$$

$$u(a) = 0, \quad u'(b) = 0$$

where $p \in C^1(a, b)$, $p(x) \geq p_0 > 0$ and $f \in L_2(a, b)$.

We will discretize the interval $[a, b]$ by the grid T_h as above, where we will denote $h_i = x_i - x_{i-1}$ and $h = \max h_i$. We suppose that the grid satisfies a requirement of the type $h_i \geq \mu h$, $i = 1, \dots, n$, for a constant $\mu > 0$, which does not allow the generation of very small cells compared with the others.

The space U_h will be chosen as the space of the piecewise linear functions, corresponding to the grid T_h . It consists of the functions u_h which are continuous on $[a, b]$, $u_h(a) = 0$ and u_h is linear on each I_i and thus it is uniquely determined by the values at the ends of the element. Obviously, this is an n -dimensional subspace of $H_E^1(a, b)$.

As for the finite element method, a basis in U_h is formed by the functions

$$\Phi_i(x) = \begin{cases} 1 - \frac{x_i - x}{h_i}, & x_{i-1} \leq x \leq x_i, \\ 1 - \frac{x - x_i}{h_{i+1}}, & x_i \leq x \leq x_{i+1}, \\ 0, & x \leq x_{i-1}, x \geq x_{i+1}. \end{cases}$$

Then $u_h \in U_h$ is expressed as

$$u_h(x) = \sum_{i=1}^n u_i \Phi_i(x)$$

where $u_i = u(x_i)$. On the element I_i we have

$$u_h(x) = u_{i-1} \left(1 - \frac{x - x_{i-1}}{h_i} \right) + u_i \frac{x - x_{i-1}}{h_i},$$

$$u'_h(x) = \frac{u_i - u_{i-1}}{h_i}.$$

We will now construct the dual grid T_h^* , as above, and we will choose V_h as the piecewise constant functions space. It contains all the functions $v_h \in L_2(a, b)$ such that $v_h(x) = 0$ for $x \in I_0^*$ and v_h is a constant on each $I_i^*, i = 2, \dots, n$.

The basis for this space consists of the functions

$$\Psi_j(x) = \begin{cases} 1, & x \in I_j^*, \\ 0, & x \notin I_j^*, \end{cases}$$

and then every $v_h \in V_h$ is expressed as

$$v_h(x) = \sum_{i=1}^n v_i \Psi_i(x)$$

with $v_i = v_h(x_i)$.

We will discretize now the variational equation (7.23). We will look for

$$u_h(x) = \sum_{i=1}^n u_i \Phi_i(x)$$

such that

$$a(u_h, \Psi_j) = (f, \Psi_j), \quad j = 1, \dots, n.$$

In the case of our equation,

$$a(u_h, \Psi_j) = \int_a^b p u'_h [\delta(x - x_{j-1/2}) - \delta(x - x_{j+1/2})] dx$$

where $p, q \in C^1(a, b)$, $p(x) \geq p_0 > 0$, $q \geq 0$ and $f \in L_2(a, b)$.

We will choose now U_h as the subspace of functions that are piecewise polynomial of second order, with respect to the grid T_h . Thus any function $u_h \in U_h$ must be continuous, satisfying $u_h(a) = 0$ and being on each element I_i a second order polynomial, determined by its values at the ends and the midpoint of I_i . Obviously, we obtain a $2n$ -dimensional subspace of $U = H^1_E(a, b)$.

We find a basis if on each element we look for quadratic functions which take at the three nodes associated to the element (the ends and the midpoint) the successive values 1,0,0, respectively 0,1,0, respectively 0,0,1. So that, the basis elements will be

$$\Phi_i(x) = \begin{cases} \left(\frac{2(x_i - x)}{h_i} - 1 \right) ((x_i - x) h_i - 1), & x_{i-1} \leq x \leq x_i, \\ \left(\frac{2(x - x_i)}{h_{i+1}} - 1 \right) ((x - x_i) h_{i+1} - 1), & x_i \leq x \leq x_{i+1}, \\ 0, & x \leq x_{i-1}, x \geq x_{i+1} \end{cases}$$

and

$$\Phi_{i-\frac{1}{2}}(x) = \begin{cases} 4 \left(1 - \frac{x - x_{i-1}}{h_i} \right) \frac{x - x_{i-1}}{h_i}, & x_{i-1} \leq x \leq x_i, \\ 0, & x \leq x_{i-1}, x \geq x_i \end{cases}$$

for $i = 1, \dots, n$.

Then, any $u_h \in U_h$ can be expressed as

$$u_h(x) = \sum_{i=1}^n \left[u_i \Phi_i(x) + u_{i-\frac{1}{2}} \Phi_{i-\frac{1}{2}}(x) \right]$$

where $u_i = u_h(x_i)$ and $u_{i-\frac{1}{2}} = u_h(x_{i-\frac{1}{2}})$. On each element $I_i = [x_{i-1}, x_i]$ we have

$$\begin{aligned} u_h &= u_{i-1}(2\xi - 1)(\xi - 1) + 4u_{i-\frac{1}{2}}\xi(1 - \xi) + u_i(2\xi - 1)\xi \\ &= (\xi^2, \xi, 1) \begin{pmatrix} 2 & -4 & 2 \\ -3 & 4 & -1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_{i-1} \\ u_{i-\frac{1}{2}} \\ u_i \end{pmatrix}, \\ u'_h &= u_{i-1} \frac{4\xi - 3}{h_i} + u_{i-\frac{1}{2}} \frac{4 - 8\xi}{h_i} + u_i \frac{4\xi - 1}{h_i} \\ &= (\xi, 1) \begin{pmatrix} -4 & 4 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} \frac{u_{i-\frac{1}{2}} - u_{i-1}}{h_i} \\ \frac{u_i - u_{i-\frac{1}{2}}}{h_i} \end{pmatrix} \end{aligned}$$

where $\xi = \frac{x-x_{i-1}}{h_i}$.

We will choose the dual grid T_h^* (which generates a space V_h of the same dimension) as

$$a = x_0 < x_{\frac{1}{4}} < x_{\frac{3}{4}} < \dots < x_{n-\frac{3}{4}} < x_{n-\frac{1}{4}} < x_n = b$$

where $x_{i-\frac{k}{4}} = x_i - \frac{k}{4}h_i$, $k = 1, 3$, $i = 1, 2, \dots, n$. The functions v_h will be also chosen piecewise constant, and form a $2n$ -dimensional space spanned by the basis

$$\Psi_i(x) = \begin{cases} 1, & x_{i-\frac{1}{4}} \leq x \leq x_{i+\frac{1}{4}} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\Psi_{i-\frac{1}{2}}(x) = \begin{cases} 1, & x_{i-\frac{3}{4}} \leq x \leq x_{i-\frac{1}{4}} \\ 0, & \text{otherwise.} \end{cases}$$

Then any $v_h \in V_h$ is represented as

$$v_h = \sum_{j=1}^n [v_j \Psi_j(x) + v_{j-\frac{1}{2}} \Psi_{j-\frac{1}{2}}(x)].$$

We can now discretize the problem. The discrete variational problem is now to find the function $u_h \in U_h$ such that

$$\begin{cases} a(u_h, \Psi_j) = (f, \Psi_j), & j = 1, \dots, n \\ a(u_h, \Psi_{j-\frac{1}{2}}) = (f, \Psi_{j-\frac{1}{2}}), & j = 1, \dots, n \end{cases} \quad (7.25)$$

where

$$\begin{aligned} a(u_h, \Psi_j) &= p_{j-\frac{1}{4}} u'_h(x_{j-\frac{1}{4}}) - p_{j+\frac{1}{4}} u'_h(x_{j+\frac{1}{4}}) + \int_{x_{j-\frac{1}{4}}}^{x_{j+\frac{1}{4}}} q u_h dx \\ &= 2p_{j-\frac{1}{4}} \frac{u_j - u_{j-\frac{1}{2}}}{h_j} - 2p_{j+\frac{1}{4}} \frac{u_{j+\frac{1}{2}} - u_j}{h_{j+1}} + \int_{x_{j-\frac{1}{4}}}^{x_{j+\frac{1}{4}}} q u_h dx, \end{aligned}$$

respectively,

$$\begin{aligned} a(u_h, \Psi_{j-\frac{1}{2}}) &= p_{j-\frac{3}{4}} u'_h(x_{j-\frac{3}{4}}) - p_{j-\frac{1}{4}} u'_h(x_{j-\frac{1}{4}}) + \int_{x_{j-\frac{3}{4}}}^{x_{j-\frac{1}{4}}} q u_h dx \\ &= 2p_{j-\frac{3}{4}} \frac{u_{j-\frac{1}{2}} - u_{j-1}}{h_j} - 2p_{j-\frac{1}{4}} \frac{u_j - u_{j-\frac{1}{2}}}{h_j} + \int_{x_{j-\frac{3}{4}}}^{x_{j-\frac{1}{4}}} q u_h dx. \end{aligned}$$

In all these expressions $u_0 = 0$ and for $j = n$ the quantities on the right-hand side of b must be dropped. Similarly, we make the convention that $p_{n+\frac{1}{4}} = 0$ and $x_{n+\frac{1}{4}} = x_n$.

From the quadrature formulas

$$\int_{x_{j-\frac{1}{4}}}^{x_{j+\frac{1}{4}}} qu_h dx = \frac{h_j + h_{j+1}}{4} q_j u_j,$$

$$\int_{x_{j-\frac{3}{4}}}^{x_{j-\frac{1}{4}}} qu_h dx = \frac{h_j}{2} q_{j-\frac{1}{2}} u_{j-\frac{1}{2}}$$

we obtain the system

$$2p_{j-\frac{1}{4}} \frac{u_j - u_{j-\frac{1}{2}}}{h_j} - 2p_{j+\frac{1}{4}} \frac{u_{j+\frac{1}{2}} - u_j}{h_{j+1}} + \frac{h_j + h_{j+1}}{4} q_j u_j = \int_{x_{j-\frac{1}{4}}}^{x_{j+\frac{1}{4}}} f dx,$$

$$2p_{j-\frac{3}{4}} \frac{u_{j-\frac{1}{2}} - u_{j-1}}{h_j} - 2p_{j-\frac{1}{4}} \frac{u_j - u_{j-\frac{1}{2}}}{h_j} + \frac{h_j}{2} q_{j-\frac{1}{2}} u_{j-\frac{1}{2}} = \int_{x_{j-\frac{3}{4}}}^{x_{j-\frac{1}{4}}} f dx.$$

If the unknowns are arranged in the order

$$u_{\frac{1}{2}}, u_1, u_{\frac{3}{2}}, u_2, \dots, u_{n-\frac{1}{2}}, u_n,$$

then the coefficient matrix of the system is also symmetric tridiagonal and of the form

$$\begin{pmatrix} a_{00} & a_{01} & & & \\ a_{10} & a_{11} & a_{12} & & \\ & a_{21} & a_{22} & \ddots & \\ & & & \ddots & \ddots \end{pmatrix}$$

where

$$\begin{aligned} a_{00} &= \frac{2p_{\frac{1}{4}}}{h_1} + \frac{2p_{\frac{3}{4}}}{h_1} + \frac{h_1}{2} q_{\frac{1}{2}}, & a_{01} &= a_{10} = -\frac{2p_{\frac{3}{4}}}{h_1}, \\ a_{11} &= \frac{2p_{\frac{3}{4}}}{h_1} + \frac{2p_{\frac{5}{4}}}{h_2} + \frac{h_1 + h_2}{4} q_1, & a_{12} &= a_{21} = -\frac{2p_{\frac{5}{4}}}{h_2}, \\ a_{22} &= \frac{2p_{\frac{5}{4}}}{h_2} + \frac{2p_{\frac{7}{4}}}{h_2} + \frac{h_2}{2} q_{\frac{3}{2}}, & a_{23} &= a_{32} = -\frac{2p_{\frac{7}{4}}}{h_2}, \\ a_{33} &= \frac{2p_{\frac{7}{4}}}{h_2} + \frac{2p_{\frac{9}{4}}}{h_3} + \frac{h_2 + h_3}{4} q_2, & a_{34} &= a_{43} = -\frac{2p_{\frac{9}{4}}}{h_3}, \end{aligned}$$

and so on.

As regards the convergence, we have the following result.

THEOREM 7.1. *If $u \in H^3(a, b)$ is the solution of the problem (7.24) and u_h is the solution of the discretized by quadratic element problem (7.25), then*

$$|u - u_h|_1 \leq Ch^2 |u|_3.$$

2.1.3 **The Cubic Case**

Let us consider the more general problem

$$Lu \equiv -\frac{d}{dx}\left(p \frac{du}{dx}\right) + r \frac{du}{dx} + qu = f, \quad x \in (a, b), \tag{7.26}$$

$$u(a) = 0, \quad u'(b) = 0$$

where $p \in C^1(a, b)$, $p(x) \geq p_0 > 0$, $q, r \in C(a, b)$ and $f \in L_2(a, b)$.

We will choose now U_h as the space of the piecewise polynomial functions of third order with respect to the grid T_h . Thus any function $u_h \in U_h$ must be continuous and differentiable, satisfying $u_h(a) = 0$ and on each element I_i it is a cubic polynomial determined by its values and derivatives at the ends of I_i . We obtain a $(2n + 1)$ -dimensional subspace of $U = H_E^1(a, b) \cap H^2(a, b)$.

We construct a basis looking for cubic polynomials P which verify

$$P(0) = 1, P'(0) = P(1) = P'(1) = 0$$

respectively

$$P'(0) = 1, P(0) = P(1) = P'(1) = 0.$$

In the first case, $P(s) = (1 - s)^2(2s + 1)$ while in the second case $P(s) = cs(1 - s)^2$ where c will be separately determined on each element. By the changes of variable $s = \frac{x-x_i}{h_{i+1}}$, respectively $s = \frac{x_i-x}{h_i}$, we obtain the corresponding expressions of P on the elements I_i , thus we have

$$\Phi_i^{(0)}(x) = \begin{cases} \left(1 - \frac{x_i - x}{h_i}\right)^2 (2h_i(x_i - x) + 1), & x_{i-1} \leq x \leq x_i \\ \left(1 - \frac{x - x_i}{h_{i+1}}\right)^2 (2h_{i+1}(x - x_i) + 1), & x_i \leq x \leq x_{i+1} \\ 0, & \text{otherwise,} \end{cases}$$

$$\Phi_i^{(1)}(x) = \begin{cases} \left(\frac{x_i - x}{h_i} - 1\right)^2 (x - x_i), & x_{i-1} \leq x \leq x_i \\ \left(\frac{x - x_i}{h_{i+1}} - 1\right)^2 (x - x_i), & x_i \leq x \leq x_{i+1} \\ 0, & \text{otherwise.} \end{cases}$$

Any $u_h \in U_h$ could be represented as

$$u_h(x) = \sum_{i=0}^n \left[u_i \Phi_i^{(0)}(x) + u'_i \Phi_i^{(1)}(x) \right]$$

where $u_0 = 0$, $u_i = u_h(x_i)$, $u'_i = u'_h(x_i)$.

On each element $I_i = [x_{i-1}, x_i]$ we have

$$\begin{aligned} u_h &= u_{i-1}(1 - \xi)^2(2\xi + 1) + u_i \xi^2(3 - 2\xi) + \dots \\ &\quad + u'_{i-1} h_i \xi(1 - \xi)^2 + u'_i h_i \xi^2(\xi - 1) \\ &= (\xi^3, \xi^2, \xi, 1) \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_{i-1} \\ u_i \\ h_i u'_{i-1} \\ h_i u'_i \end{pmatrix}, \end{aligned} \tag{7.27}$$

and

$$\begin{aligned} u'_h &= u_{i-1} \frac{6\xi^2 - 6\xi}{h_i} + u_i \frac{6\xi - 6\xi^2}{h_i} + \dots \\ &\quad + u'_{i-1} (3\xi^2 - 4\xi + 1) + u'_i (3\xi^2 - 2\xi) \\ &= (\xi^2, \xi, 1) \begin{pmatrix} -6 & 3 & 3 \\ 6 & -4 & -2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{u_i - u_{i-1}}{h_i} \\ u'_{i-1} \\ u'_i \end{pmatrix} \end{aligned} \tag{7.28}$$

where $\xi = \frac{x - x_{i-1}}{h_i}$.

We will choose the dual grid T_h^* (which generates a space V_h of the same dimension) as

$$a = x_0 < x_{\frac{1}{2}} < x_{\frac{3}{2}} < \dots < x_{n-\frac{1}{2}} < x_n = b$$

where $x_{i-\frac{1}{2}} = \frac{x_i + x_{i-1}}{2}$. The functions v_h will be chosen now piecewise linear, from a $2n$ -dimensional space with the basis

$$\Psi_i^{(0)}(x) = \begin{cases} 1, & x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\Psi_i^{(1)}(x) = \begin{cases} x - x_i, & x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}} \\ 0, & \text{otherwise.} \end{cases}$$

Then, any $v_h \in V_h$ is represented as

$$v_h = \sum_{i=0}^n \left[v_i \Psi_i^{(0)}(x) + v'_i \Psi_i^{(1)}(x) \right]$$

where $v_0 = 0, v_i = v_h(x_i)$ and $v'_i = v'_h(x_i)$.

Let us discretize the problem. The discrete variational problem is now to find the function $u_h \in U_h$ such that

$$\begin{cases} a(u_h, \Psi_j^{(0)}) = (f, \Psi_j^{(0)}), & j = 1, \dots, n \\ a(u_h, \Psi_j^{(1)}) = (f, \Psi_j^{(1)}), & j = 0, \dots, n. \end{cases} \quad (7.29)$$

Let us study here only the dominant term from $a(u_h, v_h)$, i.e.,

$$b(u_h, v_h) = \int_a^b p \frac{du_h}{dx} \frac{dv_h}{dx} dx.$$

From (7.27) and (7.28) we have

$$\begin{aligned} b(u_h, \Psi_j^{(0)}) &= p_{j-\frac{1}{2}} u'_{j-\frac{1}{2}} - p_{j+\frac{1}{2}} u'_{j+\frac{1}{2}} \\ &= \frac{3}{2} p_{j-\frac{1}{2}} \frac{u_j - u_{j-1}}{h_j} - \frac{3}{2} p_{j+\frac{1}{2}} \frac{u_{j+1} - u_j}{h_{j+1}} \\ &\quad - \frac{1}{4} p_{j-\frac{1}{2}} u'_{j-1} + \frac{1}{4} (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) + \frac{1}{4} p_{j+\frac{1}{2}} u'_{j+1}, \quad j = 1, \dots, n-1 \end{aligned}$$

and

$$b(u_h, \Psi_n^{(0)}) = p_{n-\frac{1}{2}} u'_{j-\frac{1}{2}} = \frac{3}{2} p_{n-\frac{1}{2}} \frac{u_n - u_{n-1}}{h_n} - \frac{1}{4} p_{n-\frac{1}{2}} (u'_{n-1} + u'_n),$$

respectively

$$\begin{aligned} b(u_h, \Psi_j^{(1)}) &= -\frac{h_j}{2} p_{j-\frac{1}{2}} u'_{j-\frac{1}{2}} - \frac{h_{j+1}}{2} p_{j+\frac{1}{2}} u'_{j+\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p u'_h dx \\ &= -\frac{3}{4} p_{j-\frac{1}{2}} (u_j - u_{j-1}) - \frac{3}{4} p_{j+\frac{1}{2}} (u_{j+1} - u_j) + \frac{1}{8} p_{j-\frac{1}{2}} h_j u'_{j-1} \\ &\quad + \frac{1}{8} (p_{j+\frac{1}{2}} h_{j+1} + p_{j-\frac{1}{2}} h_j) u'_j + \frac{1}{8} p_{j+\frac{1}{2}} h_{j+1} u'_{j+1} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p u'_h dx. \end{aligned}$$

We will approximate the integral by

$$\begin{aligned} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p u'_h dx &= p u_h \Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p' u_h dx \\ &\simeq p_{j+\frac{1}{2}} u_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} u_{j-\frac{1}{2}} - u_j (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) \end{aligned}$$

from which we obtain

$$b(u_h, \Psi_j^{(1)}) = -\frac{1}{4} p_{j-\frac{1}{2}} (u_j - u_{j-1}) - \frac{1}{4} p_{j+\frac{1}{2}} (u_{j+1} - u_j)$$

Then for the approximate solution u_h of the problem (7.29) we have the estimation

$$\|\Pi_h u - u_h\|_1 \leq Ch^4 \|u\|_5.$$

This result shows that at certain nodes the accuracy could be increased with respect to the optimal one.

2.2 **Second Order Elliptic Problems**

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with piecewise smooth boundary $\partial\Omega$ and let us consider the boundary value problem

$$\begin{aligned}
 - \left[\frac{\partial}{\partial x} \left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} \right) + \frac{\partial}{\partial y} \left(a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} \right) \right] + qu = f, \text{ in } \Omega, \\
 u|_{\partial\Omega} = 0,
 \end{aligned} \tag{7.30}$$

of elliptic type. This means that $a_{ij}(x, y)$ and $q(x, y)$ are smooth enough and they verify

$$\sum_{i,j=1}^2 a_{ij}(x, y) \xi_i \xi_j \geq \gamma \sum_{i=1}^2 \xi_i^2, \quad q(x, y) \geq 0, \quad \forall (\xi_1, \xi_2) \in \mathbb{R}^2, \quad \forall (x, y) \in \bar{\Omega}$$

where $\gamma > 0$ is a constant. We also suppose $f \in L_2(\Omega)$.

The associated variational problem is to find $u \in U \equiv H_0^1(\Omega)$ such that

$$a(u, v) = (f, v), \quad \forall v \in U$$

where

$$\begin{aligned}
 &= \int_{\Omega} \left[\left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} \right) \frac{\partial v}{\partial x} + \left(a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} \right) \frac{\partial v}{\partial y} + quv \right] dx dy, \\
 &(f, v) = \int_{\Omega} f v dx dy.
 \end{aligned}$$

In order to discretize the problem, let U_h and V_h be finite dimensional spaces (of the same dimension); the discretized problem is to find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h \tag{7.31}$$

The case $U_h = V_h$ leads to the standard Galerkin method. In the finite volume method we choose, generally, $V_h \neq U_h$ and even V_h is not included in U . It is defined by a dual grid and the equation (7.31) is considered in the distribution sense. Different choices of U_h and V_h generate different schemes and the representation of u_h by a basis of U_h leads to a system of algebraic equations for determining the coefficients.

We will present the case of a triangular mesh. The case of quadrilateral elements is treated similarly.

Suppose that Ω is a polygonal domain which will be divided into a finite number of triangles. These have no overlapping internal regions; a vertex of any triangle does not belong to a side of any other triangle, it may coincide only with another vertex. Moreover, each vertex of $\partial\Omega$ is a vertex of a triangle.

Each triangle is called an *element* and each vertex is called a *node*. All these elements constitute a triangulation T_h of Ω where h is the maximum length of all the sides.

Let us construct the dual grid T_h^* . Given a node P_0 let $P_i, i = 1, 2, \dots$ be the neighboring nodes and M_i the midpoints of the sides P_0P_i . Choosing a point Q_i on each element $P_0P_iP_{i+1}$, we will connect successively $M_1Q_1M_2Q_2\dots$ to form the polygonal region $K_{P_0}^*$ (obviously, the polygonal line is closed after a finite number of segments). The polygon $K_{P_0}^*$ is the dual element of P_0 and all the dual elements constitute the dual decomposition of Ω .

For concrete problems the following dual decompositions are the most important. One case is to choose Q_i as the barycenter of the triangle $P_0P_iP_{i+1}$ and the other is when Q_i is the circumcenter of the same triangle.

Of course, the triangulations must be quasi-uniform, corresponding to the relation (6.16), i.e.,

$$C_1h^2 \leq S_Q \leq h^2$$

for any node Q of the dual grid T_h^* . The two above important cases for the choice of the nodes Q_i also implies the relation

$$C_2h^2 \leq S_{P_0}^* \leq C_3h^2$$

for every node P_0 of the grid T_h . Here S_Q and $S_{P_0}^*$ are respectively the area of the element from T_h containing Q and the area of the dual element $K_{P_0}^*$.

The space U_h can be chosen as the space of the piecewise linear functions generated by T_h . Therefore, the functions u_h should be continuous; they satisfy $u_h|_{\partial\Omega} = 0$ and on each element K from T_h , u_h is a linear function with respect to x and y , determined by its values on the vertices of the triangle. Consequently, $U_h \subset U = H_0^1(\Omega)$.

The expression of these elements with respect to a basis in U_h is made as in the formula (6.17).

Concerning the test functions space, these will be chosen piecewise constant with respect to T_h^* . The spatial basis is constructed as follows.

For any interior node P_0 we choose the function

$$\Phi_{P_0}(P) = \begin{cases} 1, & P \in K_{P_0}^* \\ 0, & \text{otherwise.} \end{cases}$$

Then any $v_h \in V_h$ is expressed as

$$v_h = \sum_P v_h(P) \Phi_P$$

where P belongs to the set of the interior nodes. If $w \in U = H_0^1(\Omega)$ and

$$\Pi_h^* w = \sum_P w(P) \Phi_P$$

is the interpolant on V_h , we have the estimation

$$|w - \Pi_h^* w|_0 \leq Ch |w|_1.$$

With these discretizations, the numerical problem is reduced to finding $u_h \in U_h$ for which

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h$$

or

$$a(u_h, \Phi_{P_0}) = (f, \Phi_{P_0}), \quad \forall P_0 \in \dot{\Omega}_h \quad (7.32)$$

where $\dot{\Omega}_h$ is the set of the interior nodes from T_h . Here

$$\begin{aligned} a(u_h, \Phi_{P_0}) &= - \int_{\partial K_{P_0}^*} \left[W_h^{(1)} \cos \langle n, x \rangle + W_h^{(2)} \cos \langle n, y \rangle \right] d\sigma \\ &+ \int_{K_{P_0}^*} q u_h dx dy = - \int_{\partial K_{P_0}^*} W_h^{(1)} dy + \int_{\partial K_{P_0}^*} W_h^{(2)} dx + \int_{K_{P_0}^*} q u_h dx dy \end{aligned}$$

and n is the outward normal to the boundary of the element and

$$W_h^{(1)} = a_{11} \frac{\partial u_h}{\partial x} + a_{12} \frac{\partial u_h}{\partial y}, \quad W_h^{(2)} = a_{21} \frac{\partial u_h}{\partial x} + a_{22} \frac{\partial u_h}{\partial y}.$$

The integrals can be calculated by different quadrature formulas.

Let us illustrate the method on the simplest case of the Poisson equation

$$-\Delta u = f,$$

for which

$$a(u_h, \Phi_{P_0}) = - \int_{\partial K_{P_0}^*} \frac{\partial u_h}{\partial n} d\sigma = - \sum_i \int_{M_i Q_i M_{i+1}} \left(\frac{\partial u_h}{\partial x} dy - \frac{\partial u_h}{\partial y} dx \right).$$

Now $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$ are constant on each element, thus the integrals do not depend on the location of the nodes Q_i . We obtain the system

$$a(u_h, \Phi_{P_0}) \equiv \sum_i \frac{\left[(u_{P_i} - u_{P_0}) \frac{b_i^2 - c_i^2 - a_i^2}{2} + (u_{P_{i+1}} - u_{P_0}) \frac{a_i^2 - b_i^2 - c_i^2}{2} \right]}{4S_{Q_i}} = \int_{K_{P_0}^*} f dx dy$$

where $P_{i+1}P_0 = a_i$, $P_iP_0 = b_i$, $P_{i+1}P_i = c_i$.

In the case of a uniform triangulation (see Figure 7.1) the discrete

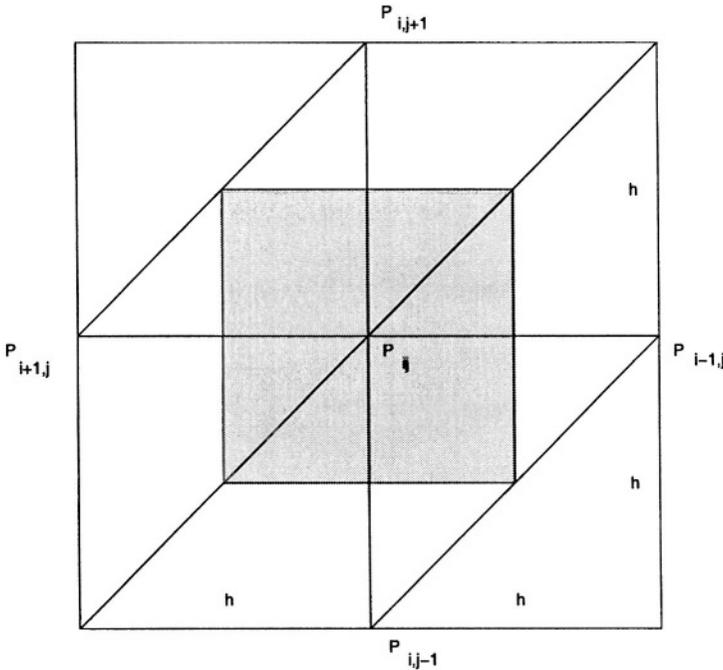


Figure 7.1. Uniform triangulation

system reduces to

$$4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = \int_{K_{P_0}^*} f dx dy$$

where u_{ij} represents $u(P_{ij})$. In fact, this is the standard five points scheme.

But we could consider an equilateral triangulation, with the sides $P_0P_i = h$ from where $Q_iQ_{i+1} = h/\sqrt{3}$ (see Figure 7.2)

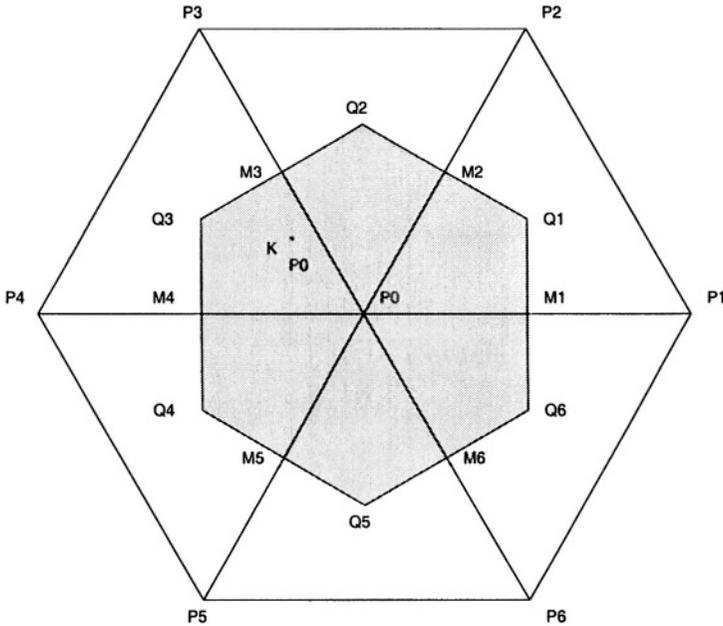


Figure 7.2. Equilateral triangulation

In this case the discrete system becomes

$$\frac{1}{\sqrt{3}} \left(6u_{P_0} - \sum_{i=1}^6 u_{P_i} \right) = \int_{K_{P_0}^*} f dx dy.$$

Of course, in all the above formulas, the last node $i + 1$ coincides with the first (for instance, $P_{6+1} = P_1$).

As regards the errors, we have the following estimation.

THEOREM 7.4. *Let u be the solution of the problem (7.30) and u_h the solution of the problem (7.32). If $u \in H^2(\Omega)$, then*

$$\|u - u_h\|_1 \leq Ch |u|_2.$$

This estimation could be improved by choosing better spaces U_h . The presented methods could be extended for quadrilateral meshes, to high order elliptic equations or to nonlinear equations.

2.3 Parabolic Equations

Let us consider now the mixed problem for a parabolic equation

$$\begin{aligned} u_t + Au &= f(x, t), \quad x \in \Omega, \quad 0 < t \leq T, \\ u &= 0, \quad x \in \partial\Omega, \\ u &= u_0(x), \quad x \in \Omega, \quad t = 0, \end{aligned}$$

where Ω is a bounded domain in \mathbb{R}^n with a Lipschitz continuous boundary and A is a second order elliptic differential operator,

$$Au \equiv - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) + \sum_{j=1}^n b_j \frac{\partial u}{\partial x_j} + cu.$$

The corresponding variational problem is to find a function $u = u(\cdot, t) \in H_0^1(\Omega)$, $0 \leq t \leq T$ such that

$$\begin{aligned} (u_t, v) + a(u, v) &= (f, v), \quad \forall v \in H_0^1(\Omega), \quad t > 0, \\ u(x, 0) &= u_0(x), \quad x \in \Omega. \end{aligned} \tag{7.33}$$

Here

$$a(u, v) = \int_{\Omega} \left(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sum_{j=1}^n b_j \frac{\partial u}{\partial x_j} v + cuv \right) dx$$

and we suppose

$$a(u, u) \geq \alpha \|u\|_1^2, \quad \forall u \in H_0^1(\Omega).$$

In order to discretize this problem we will construct a quasi-uniform grid and dual grid on Ω , together with the spaces $U_h \subset H_0^1(\Omega)$ and $V_h \subset L_2(\Omega)$. The discrete problem is to seek a function $u_h = u_h(\cdot, t) \in U_h$, for $0 \leq t \leq T$, such that

$$\left(\frac{\partial}{\partial t} u_h, v_h \right) + a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h, \quad t > 0,$$

$$u_h(x, 0) = u_{0h}(x), \quad x \in \Omega$$

where u_{0h} is the interpolation projection of u_0 in U_h .

If $\Phi_j(x)$, $j = 1, \dots, m$, respectively $\Psi_j(x)$, $j = 1, \dots, m$, are bases of U_h and V_h , the above problem could be expressed in the following form:

find $u_h = \sum_{j=1}^m c_j(t)\Phi_j(x)$ such that the coefficients $c_j(t)$ verify

$$\sum_{j=1}^m \left[\frac{dc_j(t)}{dt} (\Phi_j, \Psi_i) + c_j(t)a(\Phi_j, \Psi_i) \right] = (f, \Psi_i), \quad i = 1, \dots, m, \quad t > 0,$$

$$c_j(0) = \alpha_j, \quad j = 1, \dots, m$$

where α_j are the coefficients of u_{0h} ,

$$u_{0h} = \sum_{j=1}^m \alpha_j \Phi_j.$$

If we denote the matrices

$$M = (m_{ij}) = ((\Phi_j, \Psi_i)), \quad K = (k_{ij}) = (a(\Phi_j, \Psi_i))$$

and the vectors

$$\mathbf{u} = (c_j(t)), \quad F = ((f, \Psi_j)), \quad \alpha = (\alpha_j)$$

for $i, j = 1, \dots, m$, the above system can be rewritten in matrix form

$$\begin{aligned} M \frac{d\mathbf{u}}{dt} + K\mathbf{u} &= F, \\ \mathbf{u}(0) &= \alpha \end{aligned} \tag{7.34}$$

which can be solved by specific methods. We remark that M is nonsingular thus the differential system has a unique solution for any $f \in L_2(\Omega)$.

As regards the error estimation we have

THEOREM 7.5. *If u , respectively u_h , are the solutions of the problems (7.33) and (7.34), then*

$$\begin{aligned} \|u - u_h\|_0 &\leq C \left\{ \|u_0 - u_{0h}\|_0 + h^2 \left[\|u_0\|_3 + \int_0^t \|u_\tau\|_3 d\tau \right] \right\}, \\ \|u - u_h\|_1 &\leq C \left\{ \|u_0 - u_{0h}\|_1 + h \left[\|u_0\|_2 + \int_0^t \|u_\tau\|_2 d\tau + \left(\int_0^t \|u_\tau\|_2^2 d\tau \right)^{\frac{1}{2}} \right] \right\}. \end{aligned}$$

In order to obtain numerical solutions for the problem (7.33) we must also discretize the differential system (7.34).

Let us denote by Δt the time stepsize and by $t_n = n\Delta t$, $u_h^n = u_h(t_n)$. At the moment t_n , we will discretize the time derivative by a backward finite differences formula

$$\frac{d}{dt} u_h^n \simeq \frac{u_h^n - u_h^{n-1}}{\Delta t}.$$

Thus we obtain a fully-discrete scheme (backward Euler):

find $u_h^n \in U_h, n = 1, 2, \dots$ such that $\forall v_h \in V_h,$

$$(u_h^n, v_h) + \Delta t a(u_h^n, v_h) = (u_h^{n-1} + \Delta t f(t_n), v_h),$$

$$u_h^0 = u_{0h}.$$

If we choose other discretization type for the time derivative, for instance

$$\frac{d}{dt} u_h^{n-\frac{1}{2}} \simeq \frac{u_h^n - u_h^{n-1}}{2\Delta t},$$

then we obtain also a fully-discrete scheme (Crank–Nicolson):

find $u_h^n \in U_h, n = 1, 2, \dots$ such that $\forall v_h \in V_h,$

$$\left(\frac{u_h^n - u_h^{n-1}}{\Delta t}, v_h \right) + a\left(\frac{u_h^n + u_h^{n-1}}{2}, v_h \right) = \left(\frac{f(t_n) + f(t_{n-1})}{2}, v_h \right),$$

$$u_h^0 = u_{0h}.$$

Both methods are implicit and the coerciveness of a guarantees the existence and the uniqueness of the solutions u_h^n for a given u_h^{n-1} .

As regards the error estimation of the fully-discretized schemes, we have the following results:

THEOREM 7.6. *Let u and u_h^n be the solutions of the problem (7.33) respectively of the backward Euler scheme. Then*

$$\begin{aligned} & \|u(t_n) - u_h^n\|_0 \\ & \leq C \left\{ \|u_0 - u_{0h}\|_0 + h^2 \left[\|u_0\|_3 + \int_0^{t_n} \|u_t\|_3 dt \right] + \Delta t \int_0^{t_n} \|u_{tt}\|_0 dt \right\}, \end{aligned}$$

$$\begin{aligned} & \|u(t_n) - u_h^n\|_1 \leq C \|u_0 - u_{0h}\|_1 \\ & + Ch \left[\|u_0\|_2 + \int_0^{t_n} \|u_t\|_2 dt + \left(\int_0^{t_n} \|u_t\|_2^2 dt \right)^{\frac{1}{2}} \right] \\ & + C\Delta t \left(\int_0^{t_n} \|u_{tt}\|_0^2 dt \right)^{\frac{1}{2}}, \end{aligned}$$

for $n = 1, 2, \dots$

THEOREM 7.7. *Let u and u_h^n be the solutions of the problem (7.33), respectively of the Crank–Nicolson scheme. Then*

$$\begin{aligned} & \|u(t_n) - u_h^n\|_0 \\ & \leq C \left\{ \|u_0 - u_{0h}\|_0 + h^2 \left[\|u_0\|_3 + \int_0^{t_n} \|u_t\|_3 dt \right] + \Delta t^2 \int_0^{t_n} \|u_{ttt}\|_0 dt \right\}, \\ & \|u(t_n) - u_h^n\|_1 \leq C \|u_0 - u_{0h}\|_1 \\ & + Ch \left[\|u_0\|_2 + \int_0^{t_n} \|u_t\|_2 dt + \left(\int_0^{t_n} \|u_t\|_2^2 dt \right)^{\frac{1}{2}} \right] \\ & + C\Delta t^2 \left(\int_0^{t_n} \|u_{ttt}\|_0^2 dt \right)^{\frac{1}{2}}, \end{aligned}$$

for $n = 1, 2, \dots$

The above schemes were the most simple, with an accuracy of order h . In the sequel we will discuss a high order scheme.

Let us consider the mixed problem

$$\begin{aligned} & \frac{\partial u}{\partial t} + Lu = f(x, t), \quad x \in (a, b), \quad 0 < t \leq T, \\ & u(a, t) = 0, \quad \frac{\partial u(b, t)}{\partial x} = 0, \\ & u(x, 0) = u_0(x), \end{aligned} \tag{7.35}$$

where

$$Lu \equiv -\frac{d}{dx} \left(p \frac{du}{dx} \right) + r \frac{\partial u}{\partial x} + qu$$

with $p \in C^1(a, b)$, $p \geq p_0 > 0$, $q, r \in C(a, b)$, $f \in L_2(a, b)$ with respect to x .

We will choose the grid T_h and the (barycenter) dual grid T_h^* on $[a, b]$. We take as U_h the space of the piecewise cubic polynomial functions related to T_h , belonging to C^1 and satisfying the boundary conditions, and as V_h the space of piecewise linear functions related to T_h^* , belonging to C .

The semi-discrete problem is to find $u_h(\cdot, t) \in U_h$ such that

$$\left(\frac{\partial u_h}{\partial t}, v_h \right) + (Lu_h, v_h) = (f, v_h), \quad \forall v_h \in V_h, \quad 0 < t \leq T, \tag{7.36}$$

$$u_h(x, 0) = u_{0h}(x)$$

and we have the following error estimation result.

THEOREM 7.8. *If u , respectively u_h , are the solutions of the problems (7.35) and (7.36), then*

$$\|u - u_h\|_1 \leq C \left\{ \|u_0 - u_{0h}\|_1 + h^3 \left[\|u_0\|_4 + \int_0^t \|u_\tau\|_4 d\tau + h \left(\int_0^t \|u_\tau\|_4^2 d\tau \right)^{\frac{1}{2}} \right] \right\}.$$

If we consider a fully-discrete Crank–Nicolson scheme,

$$\left(\frac{u_h^n - u_h^{n-1}}{\tau}, v_h \right) + \left(L \left(\frac{u_h^n + u_h^{n-1}}{2} \right), v_h \right) = \left(\frac{f(t_n) + f(t_{n-1})}{2}, v_h \right),$$

$$u_h^0 = u_{0h},$$

$\forall v_h \in V_h$, we obtain

THEOREM 7.9. *Let u and u_h^n be the solutions of the problems (7.35) respectively Crank–Nicolson scheme. Then*

$$\|u(t_n) - u_h^n\|_1 \leq C \left\{ \|u_0 - u_{0h}\|_1 + h^3 \left[\|u_0\|_4 + \int_0^{t_n} \|u_t\|_4 dt + h \left(\int_0^{t_n} \|u_t\|_4^2 dt \right)^{\frac{1}{2}} \right] \right\} + C \left[\Delta t^2 \left(\int_0^{t_n} \|u_{ttt}\|_0^2 dt \right)^{\frac{1}{2}} \right]$$

for $n = 1, 2, \dots$.

Similar results could be obtained for hyperbolic problems.

As a conclusion, we can see that for accuracy and robustness similar to the finite element method, the finite volume method is more efficient, with less computing effort.

2.4 Application

Let us consider now as an application of the GDM the numerical simulation of underground water pollution. Underground water is often contaminated by the chemical fertilizer and pesticide in agriculture, for example, which seep into the ground with rain or irrigation. These solutes in the water perform a convective motion (with respect to the underground water) and a diffusive motion due to the density diffusion of the water molecules.

A mathematical model describing the contaminated water (or the water with any chemical solute), is the following equation of the solute density C :

$$\frac{\partial (mC)}{\partial t} = \text{div} (mD \text{grad} C) - \text{div} (\mathbf{V}mC) - \frac{C'W}{n}, \quad \text{on } \Omega \subset \mathbb{R}^2. \quad (7.37)$$

Here m is the saturation thickness (depending on x, y), V is the known velocity of water, $D = \begin{pmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{pmatrix}$ is the diffusion coefficient tensor, W is the amount of the water flooded into (positive) or pumped off (negative) from a unit area of water-bearing formation. In particular, if the water goes in or out through a well $P_0(x_0, y_0)$, i.e., P_0 is either a source or a sink, then $W = Q\delta(x - x_0, y - y_0)$ where Q is the amount of water and δ the Dirac distribution. Finally, C' is the density of the solute, known for a source and unknown for a sink. Obviously, initial and boundary conditions are also considered.

Let us consider now a triangulation $T_h = \{K\}$ and its barycenter dual grid $T_h^* = \{K^*\}$, see for example Figure 7.2, where a node P_0 together with its neighboring nodes and its dual element are depicted. The sources and the sinks must be taken as nodes and, moreover, if the coefficient of the diffusion term is discontinuous on a line L , then L should be cut into several line segments by some nodes and such that each segment is a side of an element.

We assume that C is continuous when crossing such an L ($C^+ = C^-$) and the flow of the solute is also supposed to be continuous,

$$(m(D \operatorname{grad} C) \cdot \mathbf{n})^+ = (m(D \operatorname{grad} C) \cdot \mathbf{n})^-$$

for \mathbf{n} the unit outer normal vector to L .

The trial function space U_h is the piecewise linear function space related to T_h with the vertices of the elements as the nodes while the test function space is the piecewise constant space corresponding to T_h^* .

Let us integrate the equation (7.37) on $K_{P_0}^*$. We obtain

$$\begin{aligned} \int_{K_{P_0}^*} \frac{\partial(mC)}{\partial t} dx dy &= \int_{K_{P_0}^*} \operatorname{div}(mD \operatorname{grad} C) dx dy \\ &\quad - \int_{K_{P_0}^*} \operatorname{div}(\mathbf{V}mC) dx dy - \int_{K_{P_0}^*} \frac{C'W}{n} dx dy. \end{aligned}$$

By Green's formula and using $C_h \in U_h$ instead of C , we have

$$\int_{K_{P_0}^*} \operatorname{div}(mD \operatorname{grad} C_h) dx dy = \int_{\partial K_{P_0}^*} m(D \operatorname{grad} C_h) \cdot \mathbf{n} ds, \quad (7.38)$$

$$\int_{K_{P_0}^*} \operatorname{div}(\mathbf{V}mC_h) dx dy = \int_{\partial K_{P_0}^*} mC_h(\mathbf{V} \cdot \mathbf{n}) ds.$$

Working now in the triangle Δ_{Q_i} with a barycenter Q_i (see again Figure 7.2) and using the linearity of $u_h \in U_h$, we evaluate the above line integrals piecewise on the fold line segments obtained by intersecting the integral line with Δ_{Q_i} .

For example, in Δ_{Q_1} we have

$$\begin{aligned} & \int_{\overline{M_1 Q_1 M_2}} m (D \text{grad } C_h) \cdot \mathbf{n} \, ds \tag{7.39} \\ &= \int_{\overline{M_1 Q_1}} m \text{grad } C_h \cdot D\mathbf{n}_1 \, ds + \int_{\overline{Q_1 M_2}} m \text{grad } C_h \cdot D\mathbf{n}_2 \, ds \end{aligned}$$

where, denoting by (x_P, y_P) the coordinates of a point P ,

$$\begin{aligned} \text{grad } C_h &= \frac{1}{2\Delta_{Q_1}} ((y_{P_1} - y_{P_2}) C_0 + (y_{P_2} - y_{P_0}) C_1 \tag{7.40} \\ &\quad + (y_{P_0} - y_{P_1}) C_2 + (x_{P_2} - x_{P_1}) C_0 \\ &\quad + (x_{P_0} - x_{P_2}) C_1 + (x_{P_1} - x_{P_0}) C_2), \end{aligned}$$

and

$$\begin{aligned} \mathbf{n}_1 &= \frac{(y_{Q_1} - y_{M_1}, -(x_{Q_1} - x_{M_1}))}{|\overline{M_1 Q_1}|}, \tag{7.41} \\ \mathbf{n}_2 &= \frac{(y_{M_2} - y_{Q_1}, -(x_{M_2} - x_{Q_1}))}{|\overline{Q_1 M_2}|}. \end{aligned}$$

Similarly,

$$\begin{aligned} & \int_{\overline{M_1 Q_1 M_2}} m C_h (\mathbf{V} \cdot \mathbf{n}) \, ds \\ &= \int_{\overline{M_1 Q_1}} m C_h (\mathbf{V} \cdot \mathbf{n}_1) \, ds + \int_{\overline{Q_1 M_2}} m C_h (\mathbf{V} \cdot \mathbf{n}_2) \, ds. \end{aligned}$$

Here, on the line segment $\overline{M_1 Q_1}$ of equation

$$y = y_{M_1} + \frac{y_{Q_1} - y_{M_1}}{x_{Q_1} - x_{M_1}} (x - x_{M_1})$$

we have

$$C_h = \frac{x_{Q_1} - x}{x_{Q_1} - x_{M_1}} C_{M_1} + \frac{y - y_{M_1}}{y_{Q_1} - y_{M_1}} C_{Q_1},$$

while on $\overline{Q_1M_2}$ of equation

$$y = y_{Q_1} + \frac{y_{M_2} - y_{Q_1}}{x_{M_2} - x_{Q_1}} (x - x_{Q_1})$$

we have

$$C_h = \frac{x_{M_2} - x}{x_{M_2} - x_{M_1}} C_{Q_1} + \frac{y - y_{Q_1}}{y_{M_2} - y_{Q_1}} C_{M_2}.$$

Moreover,

$$C_{Q_1} = \frac{1}{3} (C_{P_0} + C_{P_1} + C_{P_2}), \quad C_{M_i} = \frac{1}{2} (C_{P_0} + C_{P_i}).$$

Concerning the source term of the equation (7.37), if P_0 is not a well, then it is directly computed, while if P_0 is a well, then

$$W = Q\delta(x - x_{P_0}, y - y_{P_0})$$

and, in this case,

$$\int_{K_{P_0}^*} \frac{C'_h Q}{n} \delta(x - x_{P_0}, y - y_{P_0}) dx dy = \frac{C'(P_0) Q(P_0)}{n(P_0)}. \quad (7.42)$$

Finally, we discretize the derivative with respect to t on the left-hand side of the equation (7.37). Let Δt be the time step size and let us take the nodes $t_k = k\Delta t$, $k = 0, 1, \dots, K$. Using, for example, the Crank–Nicolson method, we obtain

$$\begin{aligned} & \frac{1}{\Delta t} \int_{K_{P_0}^*} (m^{k+1} C_h^{k+1} - m^k C_h^k) dx dy \\ &= \frac{1}{2} \int_{K_{P_0}^*} \operatorname{div} \left(m^{k+1/2} D \operatorname{grad} (C_h^k + C_h^{k+1}) \right) dx dy \\ & \quad - \frac{1}{2} \int_{K_{P_0}^*} \operatorname{div} \left(\mathbf{V}^{k+1/2} m^{k+1/2} (C_h^k + C_h^{k+1}) \right) dx dy \\ & \quad - \frac{1}{2} \int_{K_{P_0}^*} \frac{W^k}{n} C_h^{k'} dx dy. \end{aligned}$$

The above equations with initial and boundary conditions give the GDM for the problem (7.37). This scheme could be easily extended to tetrahedral, cuboid or triangular prismatic grids on a 3D field.

Unfortunately, in the computation of contaminated underground water, one often encounters problems where the diffusion coefficient is much less than the convection speed. In such a case, the above method fails to approximate accurately the transitional band that results from the diffusion and undesirable oscillations appear. Upwind schemes are often used to eliminate these oscillations.

Let us consider a simpler two-dimensional solute transfer equation

$$\frac{\partial C}{\partial t} = \text{div}(D \text{grad} C) - \text{div}(\mathbf{V} C) + I \tag{7.43}$$

where the diffusion tensor D and the convection speed \mathbf{V} are known. The source term is $I = C'Q\delta(x - x_0, y - y_0)$ at a well and we have also initial and boundary conditions.

As above, let $T_h = \{K\}$ and $T_h^* = \{K^*\}$ be a triangulation and its barycenter dual grid. Let U_h be the piecewise linear, globally continuous function space and V_h the piecewise constant function space. Denote by Π_h^* the interpolation projection operator from U_h to V_h , i.e., for given $C_h \in U_h$, we have $\Pi_h^* C_h \in V_h$ and $(\Pi_h^* C_h)(P_i) = C_h(P_i)$.

If we integrate the equation (7.43) on $K_{P_0}^*$ for example, taking $C = C_h \in U_h$ and replacing C_h in the convection term by $\Pi_h^* C_h$, then we obtain

$$\begin{aligned} \int_{K_{P_0}^*} \frac{\partial C_h}{\partial t} dx dy &= \int_{K_{P_0}^*} \text{div}(D \text{grad} C_h) dx dy \\ &- \int_{K_{P_0}^*} \text{div}(\mathbf{V} \Pi_h^* C_h) dx dy + \int_{K_{P_0}^*} I dx dy. \end{aligned}$$

Now, the diffusion term is calculated according to (7.38), (7.39), (7.40) and (7.41) with $m = 1$, while the source term is calculated according to (7.42) with $n = 1$. The convection term is treated as follows.

We apply Green's formula

$$\int_{K_{P_0}^*} \text{div}(\mathbf{V} \Pi_h^* C_h) dx dy = \int_{\partial K_{P_0}^*} (\mathbf{V} \cdot \mathbf{n}) \Pi_h^* C_h ds,$$

denoting by $\Gamma_{0l\delta} = \overline{Q_l M_{l+\delta}}$, where $l = 1, 2, \dots, 5$ and $\delta = 0, 1$ (see Figure 7.2) and we define

$$\beta_{0l\delta} = \int_{\Gamma_{0l\delta}} (\mathbf{V} \cdot \mathbf{n}) ds,$$

$$(\partial K_{P_0}^*)^- = \left\{ \bigcup_{\beta_{0l\delta} \leq 0} \Gamma_{0l\delta} \mid 1 \leq l \leq 5, \delta = 0, 1 \right\}, \text{ i.e., flow in,}$$

$$(\partial K_{P_0}^*)^+ = \left\{ \bigcup_{\beta_{0l\delta} > 0} \Gamma_{0l\delta} \mid 1 \leq l \leq 5, \delta = 0, 1 \right\}, \text{ i.e., flow out,}$$

$$\beta_{0l\delta}^+ = \max \{ \beta_{0l\delta}, 0 \}, \quad \beta_{0l\delta}^- = \max \{ -\beta_{0l\delta}, 0 \} .$$

Then we have the approximation

$$\int_{\partial K_{P_0}^*} (\mathbf{V} \cdot \mathbf{n}) \Pi_h^* C_h ds \approx \sum_{\substack{1 \leq l \leq 5 \\ \delta = 0, 1}} \{ \beta_{0l\delta}^+ C_h(P_0) - \beta_{0l\delta}^- C_h(P_l) \} .$$

Finally, using the Crank–Nicolson method (for example) to discretize the time we obtain the scheme

$$\begin{aligned} & \frac{1}{\Delta t} \int_{K_{P_0}^*} (C_h^{k+1} - C_h^k) dx dy \\ &= \frac{1}{2} \int_{K_{P_0}^*} \operatorname{div} \left(D \operatorname{grad} (C_h^k + C_h^{k+1}) \right) dx dy \\ & - \frac{1}{4} \int_{K_{P_0}^*} \operatorname{div} [(\mathbf{V}^{k+1} + \mathbf{V}^k) \Pi_h^* (C_h^k + C_h^{k+1})] dx dy \\ & \quad + \frac{1}{2} \int_{K_{P_0}^*} (I^{k+1} + I^k) dx dy, \end{aligned}$$

where the right-hand side is calculated as above. Applying it to the equation (7.43) the oscillations disappear and the density front becomes narrower with a more accurate position [83].

Chapter 8

SPECTRAL METHODS

The spectral methods approximate the unknown functions by truncated series of orthogonal functions e_k , for example Fourier series for periodic problems or Chebyshev or Legendre polynomials for nonperiodic problems, that is

$$u_N(x) = \sum_{k=0}^N \hat{u}_k e_k$$

where the values \hat{u}_k are the unknowns. The specific way to determine these unknowns, characterizes the spectral method.

For example, in the case of the problem

$$\begin{aligned} Lu &= f, x \in (a, b), \\ u(a) &= u(b) = 0, \end{aligned}$$

the Galerkin method consists of the vanishing of the residue $R_N = Lu_N - f$ “in a weak sense”, i.e.,

$$\int_a^b w R_N e_k dx = 0, \quad k = 0, 1, \dots, N,$$

where w is a weight function associated to the orthogonality of the functions e_k .

The Galerkin method works when the functions e_k satisfies homogeneous boundary conditions. This happens for the trigonometric systems where periodic conditions appear but it does not for the orthogonal polynomial systems. For these cases the Galerkin method is modified by reducing the weak vanishing of the residue R_N equations only for

$k = 1, \dots, N - 1$ and by adding the boundary conditions

$$\sum_{k=0}^N \hat{u}_k e_k(a) = 0, \quad \sum_{k=0}^N \hat{u}_k e_k(b) = 0$$

thus obtaining the *tau method*.

Another possibility to calculate the unknown coefficients is to require that the given equation is satisfied at a certain grid, together with the boundary conditions, i.e.,

$$\begin{aligned} Lu_N(x_k) - f(x_k) &= 0, \quad k = 1, \dots, N - 1, \\ u_N(a) &= 0, \quad u_N(b) = 0, \end{aligned}$$

obtaining the *collocation method*. From the interpretation of the expression of $u_N(x)$ as a Lagrange interpolation polynomial at the nodes x_k ,

$$u_N(x) = \sum_{k=0}^N u_N(x_k) L_k(x)$$

where $L_k(x_j) = \delta_{kj}$, the unknowns to be determined are, in fact, the values $u_N(x_k)$. These methods use also relations which express the derivatives of u_N at the nodes implying the values of u_N at the same nodes, relations deduced by differentiation of the above relation.

The spectral methods are very attractive, due to the fact that the distance between the exact solution u and the approximative solution u_N is of order $1/N^s$, that is

$$\|u - u_N\| \leq \frac{C}{N^s}$$

where s depends on the regularity of $u(x)$ (the highest derivative order that $u(x)$ admits). Therefore, for a sufficiently large number of grid points, the accuracy is determined by the regularity of the exact solution. Particularly, if the solution $u(x)$ is infinitely differentiable, the error tends towards zero faster than any power of $1/N$, which means a *spectral accuracy*. This behaviour is better than that of the finite differences or finite element methods where the accuracy is fixed, of order $O(1/N^p)$, depending on the approximation scheme.

In the previous sections two principal numerical methods were presented. The first one, the finite differences method (and its variants), replaces the function u by its values

$$u_1 = u(x_1), \dots, u_m = u(x_m)$$

on a given grid. The derivatives of different orders of the function are then approximated on the same grid, by processing the discrete values. The second one is the finite element method (and its variants) which replaces the function u by the coefficients of its development with respect to a given function's basis,

$$u(x) = \sum_{i=1}^m \hat{u}_i \Phi_i(x).$$

Its derivatives are calculated directly from the above expression and then they are rediscritized with respect to the same basis, so that the derivative coefficients are obtained as functions of the original coefficients \hat{u}_i .

The great advantage of the finite differences method consists in the simplicity of the relations which discretize a differential problem, for a required (possibly high) accuracy. But if the computational domain has a more complicated geometry, this advantage is lost.

The finite element method adapts very well to computational domains of any admissible form, it allows the local refinement of the mesh depending on the gradient of the approximated solution, it allows an increasing accuracy depending on the complexity of the discretization formulas. However, this accuracy is limited by the qualities of the basis functions used for the discretization.

Both above discretization methods lead to solving of algebraic systems. Another of their advantages is the fact that the obtained linear (or linearized) algebraic systems have sparse matrices, which requires a reasonable computing effort even for a very large dimension of the systems.

The discretization by developing the function with respect to a properly chosen orthogonal system of basis functions has, moreover, the great advantage that the approximation accuracy depends on the smoothness of the function to be approximated: the higher smoothness (the function has higher order derivatives), the faster decaying of the coefficients sequence u_i . This means that smooth functions could be very well approximated by a very small number of (development) coefficients. Of course, the matrices of the systems obtained by this type of discretization are now "full", but their small dimension could compensate this drawback.

This section, following [13], presents such a type of discretizations. They induce a linear transformation between u and its coefficients sequence $(\hat{u}_i)_{i=1,\dots}$, between the physical and the transforms space, called *the finite transform* of u . If the basis system is complete, this transform could be inverted and the function u can be described either through its

values in the physical space or through its coefficients in the transforms space.

The coefficients \hat{u}_i depend on all the values of u in physical space. But a finite number m of coefficients could be calculated, with an accuracy depending on the smoothness of the function u , from a finite number of values of u on a properly selected grid. This defines a *discrete transform* between the set of the respective values of u and the set of respective approximate (discrete) coefficients. It is important to remark that this discrete transform could be performed in many cases by fast procedures with a number of operations of order $m \log_2 m$ instead of m^2 usually required by the matrix-vector multiplications.

1. **Fourier Series**

1.1 **The Discretization**

It is known that the set of functions

$$\phi_k(x) = e^{ikx}, \quad k = 0, \pm 1, \pm 2, \dots$$

is an orthogonal system in $L_2(0, 2\pi)$, i.e.,

$$\int_0^{2\pi} \phi_j(x) \overline{\phi_k(x)} dx = \begin{cases} 0, & j \neq k \\ 2\pi, & j = k. \end{cases}$$

The Fourier series of the function $u \in L_2(0, 2\pi)$ is

$$u = \sum_{k=-\infty}^{\infty} \hat{u}_k \phi_k$$

where

$$\hat{u}_k = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ikx} dx$$

are the Fourier coefficients of u , the series being convergent in $L_2(0, 2\pi)$.

An important problem is to approximate u by the truncated Fourier series

$$P_m u(x) = \sum_{k=-m/2}^{m/2-1} \hat{u}_k e^{ikx}.$$

Since from the Parseval identity we have

$$\|u - P_m u\| = \left(2\pi \sum_{k < -m/2, k \geq m/2} |\hat{u}_k|^2 \right)^{1/2},$$

a result is that the approximation error depends upon how fast the Fourier coefficients decay to zero when $|k| \rightarrow \infty$, which means that it depends on the smoothness and the periodicity of u .

Indeed, if $u \in C^1(0, 2\pi)$, we have

$$2\pi\hat{u}_k = \frac{-1}{ik} [u(2\pi-) - u(0+)] + \frac{1}{ik} \int_0^{2\pi} u'(x)e^{-ikx} dx,$$

thus $\hat{u}_k = O(k^{-1})$. If $u' \in C^1(0, 2\pi)$ and $u(2\pi-) = u(0+)$, then $\hat{u}_k = O(k^{-2})$. Iterating, if $u \in C^p(0, 2\pi)$ and $u^{(j)}$ is periodic for $j \leq p-2$, then $\hat{u}_k = O(k^{-m})$, $k = 0, \pm 1, \pm 2, \dots$. Particularly, if u is infinitely differentiable and periodic with all its derivatives, then its Fourier coefficients \hat{u}_k decay faster than any power of $1/k$, a property called *spectral accuracy*. Of course, this property can be only “asymptotically seen”, i.e., for $|k| > k_0$ large enough.

For applications, the truncation of the Fourier series is not sufficient. Another adjacent problem will be to approximate the remaining Fourier coefficients.

For an even $N > 0$, let us consider the nodes

$$x_j = \frac{2\pi j}{N}, \quad j = 0, 1, \dots, N-1. \quad (8.1)$$

The discrete Fourier coefficients of the function u corresponding to this grid are

$$\tilde{u}_k = \frac{1}{N} \sum_{j=0}^{N-1} u(x_j)e^{-ikx_j}, \quad -\frac{N}{2} \leq k \leq \frac{N}{2} - 1. \quad (8.2)$$

The above relation can be inverted and we have

$$u(x_j) = \sum_{k=-N/2}^{N/2-1} \tilde{u}_j e^{ikx_j}, \quad j = 0, 1, \dots, N-1. \quad (8.3)$$

Consequently, the trigonometric polynomial

$$I_N(x) = \sum_{k=-N/2}^{N/2-1} \tilde{u}_j e^{ikx} \quad (8.4)$$

interpolates u at the nodes (8.1) and it is called *the discrete Fourier series* of u .

We remark that the N coefficients \tilde{u}_j correspond by a one-to-one mapping with the N values $u(x_j)$ of u on the grid, a mapping which is called *the discrete Fourier transform* and is described by the relations

(8.2) and (8.3). The calculations can be accomplished by *the fast Fourier transform* (FFT).

The relation between the discrete Fourier coefficients \tilde{u}_k and the exact ones \hat{u}_k is given by the formula

$$\tilde{u}_k = \hat{u}_k + \sum_{m=-\infty, m \neq 0}^{\infty} \hat{u}_{k+Nm}, \quad k = -\frac{N}{2}, \dots, \frac{N}{2} - 1,$$

which shows that the Fourier terms with the frequencies $k + Nm$, behave on the grid (8.1) similarly with the terms corresponding to the frequency k and they are indistinguishable at the considered nodes. Therefore we have the formula

$$I_N u = P_N u + R_N u$$

where R_N represents the *aliasing error*. Its influence on the accuracy of a spectral method is of the same order as that of the truncation error.

Another phenomenon which could deteriorate the approximating qualities of the method is the oscillatory behaviour of the truncated or discrete Fourier series in a neighborhood of a discontinuity point (the ends of the interval in the case of a non-periodic function, are also included here). One remarks that $P_N u$ has oscillations of order $O(1)$ in a neighborhood of order $O(\frac{1}{N})$ of the discontinuity point. The convergence speed of $P_N u$ towards u is also reduced to an order $O(\frac{1}{N})$ even when u is smooth, excepting this discontinuity point. A similar behaviour can be observed also for the interpolant $I_N u$.

This phenomenon is called the *Gibbs phenomenon* and its reduction is very important for both theoretical and practical considerations. Its source is the slow decay of the Fourier coefficients in the case of discontinuous (or non-periodic) functions, thus its attenuation can be obtained by damping the high order modes. Of course, all the Fourier coefficients carry information about the discontinuity so that this damping must be carefully done.

Concluding, a practical mode to attenuate the Gibbs phenomenon is to replace $P_N u$ (or $I_N u$) with the smoothed series

$$S_N u = \sum_{k=-N/2}^{N/2-1} \sigma_k \hat{u}_k e^{ikx}$$

where σ_k must be real non-negative numbers and $\sigma_{|k|}$ is a decreasing function of $|k|$.

Some usual choices are:

- Cesaro smoothing,

$$\sigma_k = 1 - \frac{|k|}{\frac{N}{2} + 1}$$

which eliminates the Gibbs phenomenon, preserves the bounded variation quality of the function but generates a heavy smearing of u , modifying its values outside of the neighborhood of the discontinuity;

- Lanczos smoothing

$$\sigma_k = \frac{\sin \frac{2\pi k}{N}}{\frac{2\pi k}{N}};$$

- raised cosine smoothing

$$\sigma_k = \frac{1 + \cos \frac{2\pi k}{N}}{2}.$$

The last two types of smoothing attenuate the Gibbs phenomenon and approximate well the function u outside of the neighborhood of the discontinuity.

Figure 8.1 shows these types of smoothings for the function

$$u(x) = \begin{cases} 1, & \pi/2 < x < 3\pi/2 \\ 0, & \text{otherwise.} \end{cases}$$

Here

$$P_N u(x) = 0.5 + 2 \sum_{k=1}^{N/2} \sigma_k \hat{u}_k \cos kx$$

where

$$\hat{u}_k = \frac{(\cos(k\pi - 1)) \sin \frac{k\pi}{2}}{2k\pi}, \quad k = 1, 2, \dots, N/2.$$

1.2 Approximation of the Derivatives

The most important problem for the discretization of differential and partial differential equations is the approximation of the derivatives of the unknown function. This depends upon the representation of the function in the physical or transforms spaces.

The differentiation in the transforms space is very simple. If

$$u = \sum_{k=-\infty}^{\infty} \hat{u}_k \phi_k$$

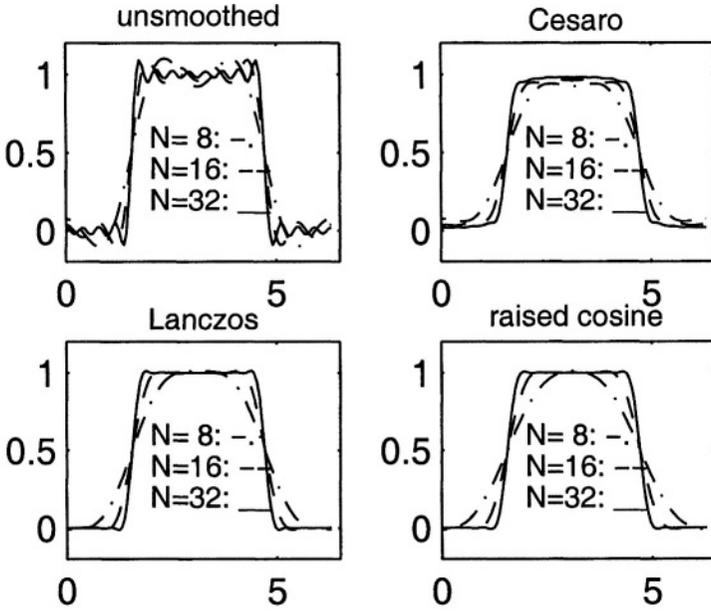


Figure 8.1. The attenuation of the Gibbs phenomenon

is the Fourier series of u , then $Su' = \sum_{k=-\infty}^{\infty} ik\hat{u}_k\phi_k$ is the Fourier series of the derivative u' . Shortly, $(P_Nu)' = P_Nu'$ and it is called the *Fourier–Galerkin derivative* of u . If $u \in H^1(0, 2\pi)$ both the series are convergent in $L_2(0, 2\pi)$. The differentiation and the truncation commute.

The differentiation in the physical space which starts from the values of u on the grid (8.1), evaluates the discrete Fourier coefficients by the formula (8.2), then these coefficients are multiplied by ik in order to obtain the discrete Fourier coefficients of the derivative and, finally, the values of the derivative on the grid are obtained using the corresponding formula (8.3). This differentiation procedure leads to the *Fourier collocation derivative* D_Nu of u .

So that we have $D_Nu = (I_Nu)'$, $D_Nu \neq P_Nu'$ and, generally, the differentiation and the interpolation do not commute.

This transform of the values of u on the grid to the (approximate) values of the derivative u' on the same grid could be performed using a derivative matrix D_N , i.e.,

$$(D_Nu)_i = \sum_{j=0}^{N-1} (D_N)_{ij} u_j,$$

where

$$(D_N)_{lj} = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} ike^{2ik(l-j)\pi/N} = \begin{cases} \frac{1}{2}(-1)^{l+j}ctg \left[\frac{(l-j)\pi}{N} \right], & l \neq j \\ 0, & l = j. \end{cases} \tag{8.5}$$

If $N \geq 32$, the following formula is recommended:

$$(D_N u)_l = \sum_{k=-N/2}^{N/2-1} a_k e^{2ikl\pi/N}, \quad l = 0, 1, \dots, N - 1$$

where

$$a_k = \frac{ik}{N} \sum_{j=0}^{N-1} u_j e^{-2ikj\pi/N}, \quad k = -\frac{N}{2}, \dots, \frac{N}{2} - 1$$

and the calculations could be performed using FFT.

Concerning the truncation and interpolation errors, we have the result

$$\|u - P_N u\|_{H^l(0,2\pi)} \leq CN^{l-m} \left\| u^{(m)} \right\|_{L_2(0,2\pi)}$$

for all $u \in H_p^m(0, 2\pi)$ where $m \geq 0$ and $0 \leq l \leq m$. Here $H_p^m(0, 2\pi)$ is the subspace of the functions belonging to H^m with the first $m-1$ derivatives being periodical. A similar formula exists also for the interpolant $I_N u$. Particularly,

$$\|u' - (I_N u)'\|_{L_2(0,2\pi)} \leq CN^{1-m} \left\| u^{(m)} \right\|_{L_2(0,2\pi)}.$$

2. Orthogonal Polynomials

2.1 Discrete Polynomial Transforms

Let us denote by P_N the space of the polynomials of at most N degree. Let us choose a system of polynomials p_k with degree equals to k , for $k = 0, 1, \dots$ and orthogonal with respect to the weight w over $(-1, 1)$, i.e.,

$$\int_{-1}^1 p_k(x)p_n(x)w(x)dx = 0 \text{ for } n \neq k$$

The Weierstrass theorem implies that this system of polynomials is complete in $L_{2,w}(-1,1)$. Then, any function u of this space could be expanded in a series with respect to the system p_k , that is

$$u = \sum_{k=0}^{\infty} \hat{u}_k p_k$$

where the coefficients are given by the relations

$$\hat{u}_k = \frac{1}{\|p_k\|_w^2} \int_{-1}^1 u(x) p_k(x) w(x) dx$$

and

$$\|v\|_w = \left(\int_{-1}^1 |v(x)|^2 w(x) dx \right)^{1/2}.$$

We also define, in this case, the truncated series

$$P_N u = \sum_{k=0}^N \hat{u}_k p_k.$$

In the sequel the nodes of different quadrature formulas will be important. We have the following results:

Gauss integration. Let x_0, \dots, x_N be the roots of p_{N+1} and let w_0, \dots, w_N be the solution of the linear system

$$\sum_{j=0}^N (x_j)^k w_j = \int_{-1}^1 x^k w(x) dx, \quad 0 \leq k \leq N.$$

Then $w_j > 0$ for $j = 0, \dots, N$ and

$$\sum_{j=0}^N p(x_j) w_j = \int_{-1}^1 p(x) w(x) dx, \quad \forall p \in P_{2N+1}.$$

In this case the roots are all inside of $(-1, 1)$. In order to include one end point, we will consider the polynomial $q(x) = p_{N+1}(x) + ap_N(x)$ where a is calculated such that $q(-1) = 0$.

Gauss–Radau integration. Let $-1 = x_0, x_1, \dots, x_N$ be the roots of q and let w_0, \dots, w_N be the solution of the linear system

$$\sum_{j=0}^N (x_j)^k w_j = \int_{-1}^1 x^k w(x) dx, \quad 0 \leq k \leq N.$$

Then

$$\sum_{j=0}^N p(x_j) w_j = \int_{-1}^1 p(x) w(x) dx, \quad \forall p \in P_{2N}.$$

Similarly, in order to include both ends of the interval in the grid, consider $r(x) = p_{N+1}(x) + ap_N(x) + bp_{N-1}(x)$ where a and b are calculated now such that $r(-1) = r(1) = 0$.

Gauss-Lobatto integration. Let $-1 = x_0, x_1, \dots, x_N = 1$ be the roots of r and let w_0, \dots, w_N be the solution of the linear system

$$\sum_{j=0}^N (x_j)^k w_j = \int_{-1}^1 x^k w(x) dx, \quad 0 \leq k \leq N.$$

Then

$$\sum_{j=0}^N p(x_j) w_j = \int_{-1}^1 p(x) w(x) dx, \quad \forall p \in P_{2N-1}.$$

We will suppose that the weight function w is given together with the corresponding orthogonal polynomials p_k . For a given N , let x_0, x_1, \dots, x_N be the nodes of the above quadrature formulas and let w_0, \dots, w_N be the corresponding weights.

Let us consider now a smooth function u on $(-1, 1)$ and let u_j be its values at the above grid points, $u_j = u(x_j)$. Let $I_N(u)$ be the interpolating polynomial on these nodes, i.e., $I_N(u) \in P_N$ and $I_N u(x_j) = u(x_j)$, $j = 0, \dots, N$. Since it is a polynomial, it could be represented as

$$I_N u(x) = \sum_{k=0}^N \tilde{u}_k p_k(x)$$

and then

$$u(x_j) = \sum_{k=0}^N \tilde{u}_k p_k(x_j). \quad (8.6)$$

We have also

$$\tilde{u}_k = \frac{1}{\gamma_k} \sum_{j=0}^N u(x_j) p_k(x_j) w_j \quad (8.7)$$

where

$$\gamma_k = \sum_{j=0}^N p_k^2(x_j) w_j.$$

The relations (8.6) and (8.7) relate the physical space of $\{u(x_j)\}$ with the transforms space of $\{\tilde{u}\}$, a transformation which is similar to that for the Fourier series and which is called *the discrete polynomial transform* associated with the weight w and the nodes x_0, \dots, x_N .

The relation between the discrete and continuous polynomial coefficients is

$$\tilde{u}_k = \hat{u}_k + \frac{1}{\gamma_k} \sum_{l>n}^N (p_l, p_k)_N \hat{u}_l$$

where

$$(u, v)_N = \sum_{j=0}^N u(x_j)v(x_j)w_j$$

is the discrete inner product. Therefore, $I_N u = P_N u + R_N u$ where $R_N u$ is the aliasing error.

We will present in the sequel some details about two types of polynomials, much used in CFD. For more informations see [13] and [144].

2.2 Legendre Polynomials

The Legendre polynomials $L_k(x)$, $k = 0, 1, \dots$ are the eigenfunctions of the Sturm–Liouville problem

$$((1-x^2)L'_k(x))' + k(k+1)L_k(x) = 0$$

on the interval $(-1, 1)$ with the weight $w = 1$. Usually they are normalized such that $L_k(1) = 1$.

The expansion of a function $u \in L_2(-1, 1)$ with respect to L_k is

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k L_k(x),$$

where the expansion coefficients are

$$\hat{u}_k = \left(k + \frac{1}{2}\right) \int_{-1}^1 u(x)L_k(x)dx.$$

Concerning the discrete expansions, the three types of grids and the corresponding weights are:

Legendre–Gauss. x_j , $j = 0, \dots, N$, are the roots of L_{N+1} and the weights are

$$w_j = \frac{2}{(1-x_j^2)[L'_{N+1}(x_j)]^2}, \quad j = 0, \dots, N,$$

$$\gamma_k = \frac{1}{k + \frac{1}{2}}, \quad k \leq N.$$

Legendre–Gauss–Radau. x_j , $j = 0, \dots, N$, are the roots of $L_{N+1} + L_N$ and the weights are

$$w_0 = \frac{1}{(N+1)^2}, \quad w_j = \frac{1}{(N+1)^2} \frac{1-x_j}{[L_N(x_j)]^2}, \quad j = 1, \dots, N,$$

$$\gamma_k = \frac{1}{k + \frac{1}{2}}, \quad k \leq N.$$

Legendre–Gauss–Lobatto. $x_0 = -1, x_j, j = 1, \dots, N-1,$ are the roots of $L'_N, x_N = 1$ and the weights are

$$w_j = \frac{2}{N(N+1)} \frac{1}{[L_N(x_j)]^2}, \quad j = 0, \dots, N,$$

$$\gamma_k = \frac{1}{k + \frac{1}{2}}, \quad k < N, \quad \gamma_N = \frac{2}{N}.$$

The differentiation in the transforms space consists in calculation of the derivative coefficients with respect to the given function coefficients. If $u = \sum_{k=0}^{\infty} \hat{u}_k L_k$ is smooth enough, then u' could be represented as $u' = \sum_{k=0}^{\infty} \hat{u}_k^{(1)} L_k(x)$ where the derivative coefficients are

$$\hat{u}_k^{(1)} = (2k+1) \sum_{p=k+1, p+k \text{ impar}}^{\infty} \hat{u}_p.$$

For the second derivative we have

$$u'' = \sum_{k=0}^{\infty} \hat{u}_k^{(2)} L_k(x), \quad \hat{u}_k^{(2)} = (k + \frac{1}{2}) \sum_{p=k+2, p+k \text{ par}}^{\infty} [p(p+1) - k(k+1)] \hat{u}_p.$$

Here, unlike for the Fourier series, the differentiation and the truncation do not commute, $(P_N u)' \neq P_{N-1}(u')$. The result of this type of differentiation is called *the Legendre–Galerkin derivative*.

The differentiation in the physical space is performed starting from the values of u of one of the above grids, then constructing the interpolating polynomial $I_N u$ and evaluating its derivative on that grid. The result, $D_N u = (I_N u)'$, is called *the Legendre-collocation derivative* of u and generally it is different from the Galerkin derivative $(P_N u)'$.

The calculation could be performed by multiplication of the vector of the values of u on the grid by a derivative matrix,

$$(D_N u)(x_l) = \sum_{j=0}^N (D_N)_{lj} u(x_j), \quad l = 0, \dots, N,$$

where, for the Gauss–Lobatto nodes (for example), we have

$$(D_N)_{lj} = \begin{cases} \frac{L_N(x_l)}{L_N(x_j)} \frac{1}{x_l - x_j}, & l \neq j, \\ \frac{(N+1)N}{4}, & l = j = 0, \\ -\frac{(N+1)N}{4}, & l = j = N, \\ 0, & \text{otherwise.} \end{cases}$$

For the differentiation by Legendre polynomials we have also some estimations for the truncation and for the interpolation errors, precisely

$$\|u - P_N u\|_{H^l(-1,1)} \leq CN^{-1/2} N^{2l-m} \|u\|_{H^m(-1,1)},$$

for all $u \in H^m(-1,1)$, where $m \geq 1$ and $1 \leq l \leq m$. A similar formula also exists for the interpolant $I_N u$. Particularly,

$$\|u' - (I_N u)'\|_{L_2(-1,1)} \leq CN^{5/2-m} \|u\|_{H^m(-1,1)}.$$

2.3 **Chebyshev Polynomials**

The Chebyshev polynomials $T_k(x)$, $k = 0, 1, \dots$ are the eigenfunctions of the Sturm–Liouville problem

$$\left(\sqrt{1-x^2} T_k'(x)\right)' + \frac{k^2}{\sqrt{1-x^2}} T_k(x) = 0.$$

The weight function is now $w(x) = \frac{1}{\sqrt{1-x^2}}$. If we normalize, as usual, by the relation $T_k(1) = 1$, these polynomials become

$$T_k(x) = \cos(k \arccos x), \quad k = 0, 1, \dots .$$

Therefore, by the transform $x = \cos \theta$ many results (and, implicitly, fast computing possibilities) from the theory of Fourier series could be adapted for Chebyshev polynomials.

The expansion of a function $u \in L_{2,w}(-1,1)$ with respect to T_k is

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k T_k(x)$$

where the coefficients of the expansion are

$$\hat{u}_k = \frac{2}{\pi c_k} \int_{-1}^1 u(x) T_k(x) w(x) dx$$

for

$$c_k = \begin{cases} 2, & k = 0 \\ 1, & k \geq 1. \end{cases}$$

It is interesting to remark that by the change of function $\tilde{u}(\theta) = u(\cos \theta)$, the above series becomes a cosine Fourier series

$$\tilde{u}(\theta) = \sum_{k=0}^{\infty} \hat{u}_k \cos k\theta.$$

If $u(x)$ is infinitely differentiable, then $\tilde{u}(\theta)$ is also infinitely differentiable and periodical together with all its derivatives. In this case, the Chebyshev coefficients \hat{u}_k decay to zero faster than every power of $1/k$.

For the discrete Chebyshev series we have the following nodes and weights:

Chebyshev–Gauss

$$x_j = \cos \frac{(2j+1)\pi}{2N+2}, \quad w_j = \frac{\pi}{N+1}, \quad j = 0, \dots, N.$$

Chebyshev–Gauss–Radau

$$x_j = \cos \frac{2\pi j}{2N+1}, \quad w_j = \begin{cases} \frac{\pi}{2N+1}, & j = 0, \\ \frac{2\pi}{2N+2}, & 1 \leq j \leq N. \end{cases}$$

Chebyshev–Gauss–Lobatto

$$x_j = \cos \frac{\pi j}{N}, \quad w_j = \begin{cases} \frac{\pi}{2N}, & j = 0, N, \\ \frac{\pi}{N}, & 1 \leq j \leq N-1. \end{cases}$$

Taking into account also the boundary conditions, the most used are the Gauss–Lobatto nodes. The transformation from the physical space to the Chebyshev transforms space (8.7) could be performed by multiplication by the matrix

$$C_{kj} = \frac{2}{N\bar{c}_j\bar{c}_k} \cos \frac{\pi jk}{N}$$

where

$$\bar{c}_j = \begin{cases} 2, & j = 0, N \\ 1, & 1 \leq j \leq N-1, \end{cases}$$

while the inverse transformation (8.6) is performed by multiplication by the matrix

$$(C^{-1})_{jk} = \cos \frac{\pi jk}{N}.$$

We remark that both transforms could be efficiently performed by the FFT.

We have again the aliasing error

$$\tilde{u}_k = \hat{u}_k + \sum_{j=2mN \pm k, j > N} \hat{u}_j.$$

The differentiation in the transforms space is represented by

$$u' = \sum_{k=0}^{\infty} \hat{u}_k^{(1)} T_k$$

where

$$\hat{u}_k^{(1)} = \frac{2}{c_k} \sum_{p=k+1, p+k \text{ impar}}^{\infty} p \hat{u}_p$$

while for the second derivative it is

$$\hat{u}_k^{(2)} = \frac{1}{c_k} \sum_{p=k+2, p+k \text{ par}}^{\infty} p(p^2 - k^2) \hat{u}_p.$$

The above coefficients could be also iteratively obtained, by using the relations

$$c_k \hat{u}_k^{(q)} = \hat{u}_{k+2}^{(q)} + 2(k+1) \hat{u}_{k+1}^{(q-1)}, \quad 0 \leq k \leq N-1, \quad q = 1, 2, \dots \quad (8.8)$$

where $\hat{u}_k^{(1)} = 0$ for $k \geq N$.

The above derivative is the *Chebyshev–Galerkin derivative*, $(P_N u)'$. The *Chebyshev-collocation derivative* $D_N u = (I_N u)'$, in the physical space is efficiently obtained starting from the values of u on the Gauss–Lobatto nodes and calculating the discrete Chebyshev coefficients from the relation (8.7). Then the differentiation in the transforms space is made by the iterative formulas (8.8) and finally we transform it back to physical space with the values of the derivative on the grid. All these calculations could be performed by FFT, so that for orders $N \geq 32$ this way is much faster.

Of course, the Chebyshev–collocation could be described also in a matrix form, as in the above section. We have

$$(D_N u)(x_l) = \sum_{j=0}^N (D_N)_{lj} u(x_j), \quad l = 0, \dots, N$$

where, for the Gauss-Lobatto nodes we have

$$(D_N)_{lj} = \begin{cases} \frac{\bar{c}_l (-1)^{l+j}}{\bar{c}_j x_l - x_j}, & l \neq j, \\ \frac{-x_j}{2(1-x_j^2)}, & 1 \leq l = j \leq N-1, \\ \frac{2N^2+1}{6}, & l = j = 0, \\ -\frac{2N^2+1}{6}, & l = j = N. \end{cases} \quad (8.9)$$

We have also the following estimations of the truncation and interpolation errors for the discretization by Chebyshev polynomials,

$$\|u - P_N u\|_{H_w^l(-1,1)} \leq C N^{-1/2} N^{2l-m} \|u\|_{H_w^m(-1,1)}$$

for all $u \in H_w^m(-1,1)$, where $m \geq 1$ and $1 \leq l \leq m$. A similar formula takes place also for the interpolant $I_N u$. Particularly,

$$\|u' - (I_N u)'\|_{L_{2,w}(-1,1)} \leq C N^{2-m} \|u\|_{H_w^m(-1,1)}.$$

3. Spectral Methods for PDE

We will illustrate the spectral methods on some classical problems. Consider, first, the Burgers equation

$$\begin{aligned} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - v \frac{\partial^2 u}{\partial x^2} &= 0, \\ u(x, 0) &= u_0(x) \end{aligned} \quad (8.10)$$

with a corresponding boundary condition. We should define the trial space X_N where the discrete solution will be looked for, the test space Y_N , where “the best” satisfaction of the partial derivatives equation is demanded and, obviously, the discretization scheme for this equation.

3.1 Fourier–Galerkin Method

We will look for periodical solutions on the interval $(0, 2\pi)$. The space X_N will be chosen as the space S_N of the trigonometric polynomials of degree at most $N/2$ and the approximate solution of the problem will be in the form of a truncated Fourier series

$$u^N(x, t) = \sum_{k=-N/2}^{N/2-1} \hat{u}_k(t) e^{ikx}.$$

If we require that the residue of the equation (8.10) be orthogonal to any test function from $Y_N = S_N$, we obtain

$$\int_0^{2\pi} \left(\frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - v \frac{\partial^2 u^N}{\partial x^2} \right) e^{-ikx} dx = 0, \quad k = -\frac{N}{2}, \dots, \frac{N}{2} - 1$$

thus, the coefficients $\widehat{u}_k(t)$ must verify the differential system

$$\frac{\partial \widehat{u}_k}{\partial t} + \left(u^N \frac{\partial u^N}{\partial x} \right)_k + k^2 v \widehat{u}_k = 0, \quad k = -\frac{N}{2}, \dots, \frac{N}{2} - 1$$

where

$$\left(u^N \frac{\partial u^N}{\partial x} \right)_k = \frac{1}{2\pi} \int_0^{2\pi} u^N \frac{\partial u^N}{\partial x} e^{-ikx} dx \tag{8.11}$$

and with the initial condition

$$\widehat{u}_k(0) = \frac{1}{2\pi} \int_0^{2\pi} u_0(x) e^{-ikx} dx.$$

The formula (8.11) is a particular case of a nonlinear term which could be treated in different ways. For instance,

$$\widehat{(uv)}_k = \frac{1}{2\pi} \int_0^{2\pi} uv e^{-ikx} dx = \sum_{j+l=k} \widehat{u}_j \widehat{v}_l$$

which is a convolution sum.

3.2 Fourier-Collocation

Again within the periodicity on $(0, 2\pi)$ hypothesis, we consider now that u^N is represented by its values on the grid $x_j = 2\pi j/N$, $j = 0, \dots, N - 1$. We will require that the equation (8.10) is satisfied at the grid points, i.e.,

$$\left. \frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - v \frac{\partial^2 u^N}{\partial x^2} \right|_{x=x_j} = 0, \quad j = 0, 1, \dots, N - 1$$

The unknowns are the functions of t , $u^N(x_j, t)$, which verify the above system and the initial conditions

$$u^N(x_j, 0) = u_0(x_j).$$

If we denote by U the vector of these unknown functions, the system could be written as

$$\frac{\partial U}{\partial t} + U \odot D_N U - v D_N^2 U = 0 \tag{8.12}$$

where D_N is the Fourier-collocation derivative matrix (8.5) and \odot represents the pointwise product of the two vectors.

We remark that the Burgers equation could also be written in the conservative form

$$\frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial}{\partial x} (u^2) - \nu \frac{\partial^2 u}{\partial x^2} = 0.$$

Applying to this form the Fourier-collocation method, we find the differential system

$$\frac{\partial U}{\partial t} + \frac{1}{2} D_N (U \odot U) - \nu D_N^2 U = 0 \quad (8.13)$$

which is *not equivalent* to (8.12). We also remark that for the Fourier-Galerkin method there is no difference between the two discrete systems.

3.3 Chebyshev-Tau Method

Let us look for the solution of the problem (8.10) on $(-1, 1)$ with the boundary conditions $u(-1, t) = u(1, t) = 0$. We seek the discrete solution as the series

$$u^N(x, t) = \sum_{k=0}^N \hat{u}_k(t) T_k(x)$$

and again we require that the residue is orthogonal to polynomials of at most $N - 2$ degree

$$\int_{-1}^1 \left(\frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - \nu \frac{\partial^2 u^N}{\partial x^2} \right) T_k(x) (1-x^2)^{-\frac{1}{2}} dx = 0,$$

$k = 0, \dots, N - 2$. This leads to

$$\frac{\partial \hat{u}_k}{\partial t} + \left(u^N \frac{\partial u^N}{\partial x} \right)_k - \nu \hat{u}_k^{(2)} = 0, \quad k = 0, 1, \dots, N - 2 \quad (8.14)$$

where

$$\left(u^N \frac{\partial u^N}{\partial x} \right)_k = \frac{2}{\pi c_k} \int_{-1}^1 u^N \frac{\partial u^N}{\partial x} T_k(x) (1-x^2)^{-\frac{1}{2}} dx$$

and it could be calculated by the formula

$$\widehat{(uv)}_k = \frac{2}{\pi c_k} \int_{-1}^1 uv T_k(x) (1-x^2)^{-\frac{1}{2}} dx = \frac{1}{2} \sum_{j+l=k} \hat{u}_j \hat{v}_l + \sum_{|j-l|=k} \hat{u}_j \hat{v}_l.$$

To these relations we join also the boundary conditions $u^N(-1, t) = u^N(1, t) = 0$ which are transformed to

$$\sum_{k=0}^N \hat{u}_k = 0, \quad \sum_{k=0}^N (-1)^k \hat{u}_k = 0. \quad (8.15)$$

The relations (8.14) and (8.15) give a system of $N + 1$ differential equations for the functions $\hat{u}_k(t)$ with the initial conditions

$$\hat{u}_k(0) = \frac{2}{\pi c_k} \int_{-1}^1 u_0(x) T_k(x) (1-x^2)^{-\frac{1}{2}} dx, \quad k = 0, \dots, N.$$

3.4 Chebyshev-Collocation Method

Now the discrete solution u^N is represented by its values at the grid points $x_j = \cos \frac{\pi j}{N}$, $j = 0, \dots, N$ and the equation should be satisfied at the same points, i.e.,

$$\left. \frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - v \frac{\partial^2 u^N}{\partial x^2} \right|_{x=x_j} = 0, \quad j = 1, \dots, N-1$$

To these relations we also join the boundary conditions

$$u^N(-1, t) = u^N(1, t) = 0$$

and the initial conditions

$$u^N(x_j, 0) = u_0(x_j), \quad j = 0, \dots, N.$$

In this case the vector of the unknowns is

$$U = (u^N(x_1, t), \dots, u^N(x_{N-1}, t))^T.$$

The Chebyshev-collocation derivative matrix D_N (8.9) applies to a vector of $N + 1$ dimension (components), with the first and the last component zero. This means, in fact, the deletion of the first and the last column of D_N . But the partial differential equation is discretized with respect to x only at the interior nodes x_1, \dots, x_{N-1} , the values of the derivative at the first and the last node being not used. This means the deletion of the first and the last row from the derivative matrix.

Concluding, in the presence of the boundary conditions $u^N(1) = u^N(-1) = 0$, we could work with the matrix \tilde{D}_N given by

$$\left(\tilde{D}_N \right)_{i,j} = (D_N)_{i,j}, \quad i, j = 1, \dots, N-1$$

which performs the first order differentiation at the interior nodes and with

$$\left(\tilde{D}_N^{(2)}\right)_{i,j} = \left(D_N^2\right)_{i,j}, \quad i, j = 1, \dots, N-1 \quad (8.16)$$

which perform the second order differentiation at the interior nodes. The matrix form of the discrete system is thus

$$\frac{\partial U}{\partial t} + U \odot \tilde{D}_N U - \nu \tilde{D}_N^{(2)} U = 0.$$

3.5 The Calculation of the Convolution Sums

In the Burgers equation and in other equations from fluid dynamics we should discretize also some nonlinear quadratic terms of the form $w(x) = u(x)v(x)$. In the physical space this reduces to a simple multiplication of the values at the nodes, while in the transforms space this leads to the calculation of a convolution sum

$$\hat{w}_k = \sum_{j+l=k, |j|, |l| \leq N/2} \hat{u}_j \hat{v}_l, \quad |k| \leq \frac{N}{2}.$$

The direct calculation requires $O(N^2)$ operations (and much more in the spaces of higher dimension). This computational effort can not be accepted, taking into account that in the physical space only $O(N)$ operations are needed. We have seen that in some cases the direct or converse passing from the transforms space to the physical space could be performed by FFT which needs usually (in similar conditions) only $O(N \log_2 N)$ operations.

The idea is to pass from the coefficients of u and v from the transforms space to their values at the nodes, U_j respectively V_j , in physical space, to make the required sum in the physical space, i.e.,

$$W_j = U_j V_j, \quad j = 0, \dots, N$$

and then to calculate the coefficients of w in the transforms space. The total operations amount in this way is of order $O(N \log_2 N)$.

However we must remark that these transforms between the physical and the transforms spaces introduce aliasing errors too, so that the methods using such type of evaluation of the convolution sum are not genuine spectral methods. They are called *pseudospectral methods* and there are some techniques to decrease these aliasing errors.

3.6 Complete Discretization

In the above examples, only the spatial discretization was performed, leading to a semi-discretized form of the given problem

$$\begin{aligned}\frac{\partial U}{\partial t} &= F(U), \quad t > 0, \\ U(0) &= U_0\end{aligned}$$

a procedure which is called the *method of lines*.

In the study of the stability of these methods the linearized system

$$\frac{\partial U}{\partial t} = Lu$$

interferes, where L is the Jacobian of F at the respective point.

If L is a diagonalizable matrix, by a change of variable, the linear system could be decoupled into independent equations of the form

$$\frac{\partial v_j}{\partial t} = \lambda_j v_j, \quad j = 1, \dots, N$$

where λ_j are the eigenvalues of L .

The numerical integration method for the differential system is asymptotically stable if for small enough time stepsize Δt , the product of Δt by any eigenvalue (possibly complex) belongs to the stability region of the respective numerical method. Thus, it is important to know the eigenvalues of the derivative matrices of first and second orders.

For the operator $Lu = \frac{du}{dx}$ on the interval $(-1, 1)$, in the presence of the boundary condition $u(1) = 0$, the Chebyshev-collocation discretization becomes to the multiplication of the vector $U = (u(x_1), \dots, u(x_N))^T$ by the matrix

$$\left(\widehat{D}_N\right)_{i,j} = (D_N)_{i,j}, \quad i, j = 1, \dots, N$$

where D_N is the derivative matrix (8.9) and $x_j = \cos \frac{\pi j}{N}$ are the Gauss-Lobatto nodes. The spectrum of this matrix, for different values of N , is represented in Figure 8.2.

The MATLAB program is

```
for p=3:6 n=2^p;
k=0:n;x=cos(pi.*k./n);c=ones(n+1);c(1)=2;c(n+1)=2;
for i=1:(n+1) for j=1:(n+1)
if i~=j d(i,j)=c(i)./c(j).*(-1).^(i+j)./(x(i)-x(j));
elseif i==1, d(1,1)=(2*n*n+1)/6;
elseif i==n+1, d(n+1,n+1)=-(2*n*n+1)/6;
else d(i,i)=-x(i)./2./(1-x(i)^2); end end end
d1=d(2:n+1,2:n+1); l=eig(d1); subplot(2,2,p-2);
```

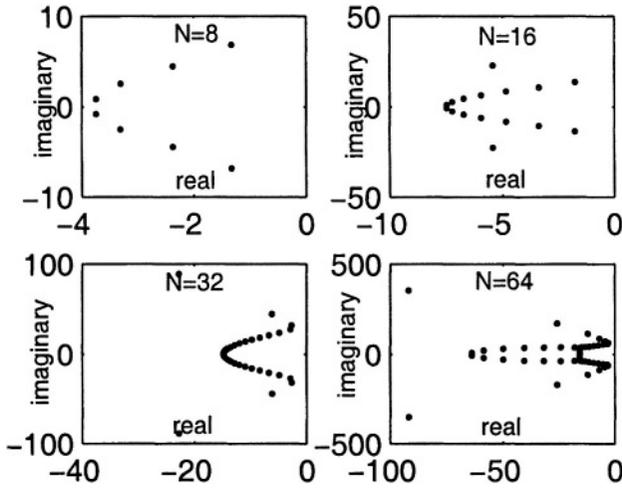


Figure 8.2. The spectrum of the Chebyshev-collocation first derivative matrix

```
plot(real(1),imag(1),'.');title(['N=',num2str(n)]);
xlabel('real');ylabel('imaginary');end
```

We remark computationally that every eigenvalue has a negative real part and their magnitudes satisfy $|\lambda| = O(N^2)$. Moreover, the first derivative matrices are very sensitive to round-off errors as we can see in the cases $N = 32$, respectively $N = 64$.

Concerning the second derivative matrix, for the operator $Lu = \frac{d^2u}{dx^2}$ on the interval $(-1, 1)$, in the presence of the boundary conditions $u(1) = 0$, $u(-1) = 0$, the Chebyshev-collocation discretization leads to the multiplication of the vector $U = (u(x_1), \dots, u(x_{N-1}))^T$ by the matrix $\tilde{D}_N^{(2)}$ given by (8.16). The eigenvalues are real negative and it can be shown (theoretically and numerically) that there exist positive constants c_1 and c_2 , independent of N , for which

$$0 < c_1 \leq -\lambda \leq c_2 N^4.$$

The numerical calculations show that about two-thirds of the eigenvalues approximate very well the eigenvalues of the second order derivative operator with the prescribed boundary conditions. Only the upper-third of the discrete eigenvalues show a very strong growth together with N . This fact influences the stability of the spectral numerical methods for differential systems and imposes the use of some unconditionally stable procedures which often are implicit. However, the very good accuracy

of spectral methods allows the use of some coarser grids than for other methods and this fact reduces essentially the computing effort.

4. **Liapunov–Schmidt (LS) Methods**

An efficient method, which can be used for different types of boundary value problems is the *Liapunov–Schmidt* (LS) method, elaborated in the years 1906–1908 and reformulated in a modern language by L. Cesari after 1963 [17]. This method applies to some nonlinear equations of the type

$$Lu = Nu, \tag{8.17}$$

for instance $u''(x) = f(x, u(x), u'(x))$, in the presence of some boundary conditions, considered on the domain of the linear operator L .

Let X and Y be real Banach spaces and let F be an application

$$F : X \times \mathbb{R} \rightarrow Y$$

satisfying

$$F(0, \lambda) = 0, \quad \forall \lambda \in \mathbb{R} \text{ and } F \in C^2.$$

We are looking for nontrivial solutions of the equation $F(u, \lambda) = 0$.

The value λ_0 is a *bifurcation value* (or $(0, \lambda_0)$ is a *bifurcation point*) for the above equation if every neighborhood of $(0, \lambda_0)$ in $X \times \mathbb{R}$ contains nontrivial solutions of it. The following important result holds.

THEOREM 8.1. *If the point $(0, \lambda_0)$ is a bifurcation point for the equation $F(u, \lambda) = 0$ then the Fréchet derivative $F_u(0, \lambda_0)$ cannot be a linear homeomorphism of X to Y .*

In the sequel we will consider so-called Fredholm operators. A linear operator $L : X \rightarrow Y$ is called a *Fredholm operator* if the kernel of L , $\ker L$, is finite dimensional, the range of L , $\text{im}L$, is closed in Y and the co-kernel of L , $\text{coker}L$, is also finite dimensional. Concerning $F_u(0, \lambda_0)$ we have:

THEOREM 8.2. *Let $F_u(0, \lambda_0)$ be a Fredholm operator with kernel V and co-kernel Z . Then there exists a closed subspace W of X and a closed subspace T of Y such that*

$$X = V \oplus W, \quad Y = Z \oplus T.$$

The operator $F_u(0, \lambda_0)|_W : W \rightarrow T$ is bijective and has a continuous inverse, hence it is a linear homeomorphism of W onto T .

We may decompose now every $u \in X$ and $F : X \rightarrow Y$ uniquely

$$\begin{aligned} u &= u_1 + u_2, & u_1 &\in V, & u_2 &\in W, \\ F &= F_1 + F_2, & F_1 &: X \rightarrow Z, & F_2 &: X \rightarrow T, \end{aligned}$$

hence the equation $F(u, \lambda) = 0$ is equivalent to the system of equations

$$\begin{aligned} F_1(u_1, u_2, \lambda) &= 0, \\ F_2(u_1, u_2, \lambda) &= 0. \end{aligned}$$

If we denote by $L = F_u(0, \lambda_0)$, using a Taylor expansion we have

$$F(u, \lambda) = F(0, \lambda_0) + Lu + N(u, \lambda)$$

and, consequently, the considered equation becomes

$$Lu + N(u, \lambda) = 0$$

or

$$Lu_2 + N(u_1 + u_2, \lambda) = 0.$$

Let now $Q : Y \rightarrow Z$ and $I - Q : Y \rightarrow T$ be projections determined by the decomposition. Then, the above equation leads to

$$QN(u, \lambda) = 0, \tag{8.18}$$

$$u_2 + L^{-1}(I - Q)N(u_1 + u_2, \lambda) = 0. \tag{8.19}$$

The equation (8.19) is a fixed point equation. If u_2 can be determined as a function of u_1 and λ , the equation (8.18) becomes an equation in a finite dimensional space for the finite dimensional u_1 .

Although used mainly for the theoretical demonstration of existence of the solutions of such a problem, including the branching of Navier–Stokes solutions for example, the above LS method (or the *alternative method*, following Cesari) is also very useful for the effective approximation of these solutions. We will present, shortly, a constructive variant of the LS method, illustrated by some examples, following [145].

Let S be a real, separable Hilbert space, $L : D(L) \subset S \rightarrow S$ a linear operator and $N : D(N) \subset S \rightarrow S$ a nonlinear operator. We impose the following assumptions:

a) L is a closed operator (i.e. $x_n \rightarrow x$ and $Lx_n \rightarrow y$ imply $x \in D(L)$ and $Lx = y$), self-adjoint, $D(L)$ is dense in S and the $\dim Ker(L) = p > 0$ is finite,

b) L has the eigenvalues $\lambda_1 = \dots = \lambda_p = 0, \lambda_{p+1} > 0, \dots$ such that $\lambda_{i+1} \geq \lambda_i$ and $\lambda_i \rightarrow \infty$ when $i \rightarrow \infty$; the corresponding eigenfunctions Φ_1, Φ_2, \dots determine an orthonormal complete system in S ,

c) there is a subspace S' of S which is complete with regard to a norm μ and $D(L) \subset S'$; for every $x \in D(L)$ its Fourier series $\sum_{k=1}^{\infty} (x, \Phi_k) \Phi_k$ converges in the norm μ too and $\{\mu(\Phi_k) / \lambda_k\}_{k > p} \in l^2$. Additionally we admit that there is an $\alpha > 0$ such that for every $x \in S'$ we have $\|x\| \leq \alpha\mu(x)$,

d) $D(N) \cap D(L) \neq \emptyset, D(N) \subset S', D(N)$ is closed vs. the norm μ ,

e) for every $R > 0$ there is $\beta_R > 0$ and $b_R > 0$ such that for all $x, y \in D(N)$ with $\mu(x) \leq R, \mu(y) \leq R$ we have $\mu(Nx - Ny) \leq \beta_R \mu(x - y)$ and $\mu(Nx) \leq b_R$.

Our purpose is the study of the existence of the solutions of the equation (8.17) in $D(L) \cap D(N)$, their numerical approximation and the evaluation of the errors.

Let $m \geq p$ and

$$S_m = sp\{\Phi_1, \dots, \Phi_m\}, S_0 = \{0\}.$$

Obviously, $S_m \subset D(L)$. We define the operators $P_m : S \rightarrow S_m$ and $H_m : S \rightarrow S$ by the following:

If

$$u \in S, \quad u = \sum_{k=1}^{\infty} (u, \Phi_k) \Phi_k,$$

then

$$P_m u = \sum_{k=1}^m (u, \Phi_k) \Phi_k, \quad H_m u = \sum_{k=m+1}^{\infty} (u, \Phi_k) \Phi_k.$$

It may be proved that H_m is well defined and for all $u \in S$ we have $H_m u \in D(L)$ while $H_m = (L|_{S_m^\perp})^{-1}$. Further, from $\|H_m\| = 1/\lambda_{m+1}$ and $\mu(H_m) \leq \alpha\sigma(m)$, where

$$\sigma(m) = \left[\sum_{k=m+1}^{\infty} \left(\frac{\mu(\Phi_k)}{\lambda_k} \right)^2 \right]^{1/2}$$

and $\mu(H_m) = \sup_{\mu(u)=1} \mu(H_m u)$, we have $\lim_{m \rightarrow \infty} \mu(H_m) = 0$. At the same time we have $H_m L u = (I - P_m)u$.

Let us suppose, additionally, that

f) $R(H_m) \subset D(N), S_m \subset D(N)$ and $D(N)$ is a subspace of S' .

Let now $\bar{u} \in D(L) \cap D(N)$ be a solution of the equation $Lu = Nu$. By applying the operators H_m and P_m to this equation, we find

$$\bar{u} = P_m \bar{u} + H_m N \bar{u}, \tag{8.20}$$

called the *auxiliary equation* of the problem, and

$$P_m (L\bar{u} - N\bar{u}) = 0, \tag{8.21}$$

called the *bifurcation equation* of the problem. Conversely, every solution of the system of (8.20) and (8.21), belonging to $D(L) \cap D(N)$, is also a solution of (8.17).

The auxiliary equation is, in fact, a fixed point problem. For its study, let $a > 0, b > 0$ and u_0 be an “approximating” solution of the equation $Lu = Nu$. Let $u^* \in S_m, u^* = \sum_{k=1}^m c_k \Phi_k$ be such that $\mu(u^* - u_0) \leq a$. We denote

$$S_{u^*}^b = \{u \in D(N) | P_m u = u^*, \mu(I - P_m)u \leq b\}.$$

We also define the operator $T_{u^*}^b : S_{u^*}^b \rightarrow S$ by

$$T_{u^*}^b(u) = u^* + H_m N u$$

for all $u \in S_{u^*}^b$. We can show that for a sufficiently large $m, T_{u^*}^b$ becomes a contraction with respect to the metric space $S_{u^*}^b$ so, according to the Banach fixed point theorem, the operator $T_{u^*}^b$ admits a unique fixed point $y(u^*)$, called the *associate element for u^** and which can be got by the method of successive approximations. So, we define another operator, $\tau_m^b : S_m \rightarrow D(L) \cap S_m^\perp$ by $\tau_m^b u^* = H_m N(u^* + \tau_m^b u^*)$. Consequently, for every $u^* \in S_m$, the associate element, $y(u^*) = u^* + \tau_m^b u^*$ i.e., it fulfils the auxiliary equation.

This element also satisfies the bifurcation equation if

$$P_m(Ly(u^*) - Ny(u^*)) = 0$$

i.e., if u^* fulfils the system

$$(\lambda_k u^* - N(u^* + \tau_m^b u^*), \Phi_k) = 0, \quad k = 1, 2, \dots, m \tag{8.22}$$

called the *system of determining equations*. This is a system on \mathbf{R}^m for the coefficients $c_k, k = 1, \dots, m$ of u^* . So, we have the theorem:

THEOREM 8.3. *If b, m are sufficiently large, then the equation (8.17) admits a solution \bar{u} if and only if the system of determining equations (8.22) admits a solution u^* and then $\bar{u} = u^* + \tau_m^b u^*$.*

Consequently, the study of the existence of the solutions of the equation $Lu = Nu$ can be reduced to the study of the existence of the solutions of the determining equations and, more, their approximation into \mathbf{R}^m leads to the approximation of the solutions of the equation $Lu = Nu$ into S' . Summarizing, the approximating algorithm is:

a) We are looking for an approximative solution of the equation $Lu = Nu$ of the form

$$u = \sum_{k=1}^m c_k \Phi_k + \sum_{k=m+1}^N c_k \Phi_k$$

where $0 \leq m \leq N$.

b) By fixing $u^* = \sum_{k=1}^m c_k \Phi_k$, we generate the associate function $y(u^*)$ performing the iterations

$$y^0 = u^*, \quad y^{s+1} = u^* + H_m N y^s = u^* + \sum_{k=m+1}^N C_k^s \Phi_k, \quad s = 0, 1, \dots, S.$$

c) With $y = y^{S+1}$ as an approximation of the associated function, we can write the system $Lu^* = P_m N y$ of the determining equations, with the unknowns c_1, \dots, c_m . This system of the form $F(c_1, \dots, c_m) = 0$ is then numerically solved, by a suitable method, for instance by Newton's method. Every evaluation of the function F means reiteration of the b) step. Finally, thus determined u^* generates, also by the b) iterations, an approximation of the solution of the equation $Lu = Nu$.

We remark that in the case of *Galerkin's method*, the approximating solutions are looked for in the form $u^* = \sum_{k=1}^N c_k \Phi_k$, where the coefficients $c_k, k = 1, \dots, N$ are determined from the equations $(Lu^* - Nu^*, \Phi_k) = 0, k = 1, \dots, N$ i.e.,

$$(\lambda_k u^* - Nu^*, \Phi_k) = 0, \quad k = 1, \dots, N.$$

These equations are derived from the determining equations for $m = N$. If $m = 0$ the system of the determining equations disappears. The associate function to a certain u^* satisfies the equation $y = L^{-1} N y$, so the algorithm reduces, in this case, to the transformation of the equation $Lu = Nu$ into a fixed point problem. Obviously, this case arises only when there exists the inverse L^{-1} and $L^{-1} N$ is a contraction.

In the case of the Navier–Stokes equations, the *nonlinear Galerkin method* [37], as a variant of the LS method, is based on the decomposition of the velocity \mathbf{u} into low, respectively high, frequency components

$$\mathbf{u} = \mathbf{y} + \mathbf{z}.$$

If we can express the pressure p with respect to u , the Navier–Stokes system becomes

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \nabla^2 \mathbf{u} + B(\mathbf{u}, \mathbf{u}) = \mathbf{f}$$

where

$$B(\mathbf{u}, \mathbf{u}) = (\mathbf{u} \cdot \nabla) \mathbf{u} + \text{grad } p(\mathbf{u})$$

is a bilinear form. Projecting this equation on the modes y and z we obtain the equivalent system

$$\frac{\partial \mathbf{y}}{\partial t} - \nu \nabla^2 \mathbf{y} + PB(\mathbf{y} + \mathbf{z}, \mathbf{y} + \mathbf{z}) = P\mathbf{f},$$

$$\frac{\partial \mathbf{z}}{\partial t} - \nu \nabla^2 \mathbf{z} + QB(\mathbf{y} + \mathbf{z}, \mathbf{y} + \mathbf{z}) = Q\mathbf{f}$$

where P, Q are the corresponding projectors.

If \mathbf{z} is small compared with \mathbf{y} , the second equation reduces to

$$\frac{\partial \mathbf{z}}{\partial t} - \nu \nabla^2 \mathbf{z} + QB(\mathbf{y}, \mathbf{y}) = Q\mathbf{f}$$

which appears as an interaction between the low and high frequency components. We deduce

$$\mathbf{z} = \varphi(\mathbf{y}) \equiv (\nu \nabla^2)^{-1} Q(B(\mathbf{y}, \mathbf{y}) - \mathbf{f})$$

and replacing this expression into the first equation of the system, we find

$$\frac{\partial \mathbf{y}}{\partial t} - \nu \nabla^2 \mathbf{y} + PB(\mathbf{y} + \varphi(\mathbf{y}), \mathbf{y} + \varphi(\mathbf{y})) = P\mathbf{f}.$$

This equation appears as a bifurcation equation

$$PL\mathbf{u}^* - PN(\mathbf{u}^* + H\mathbf{N}\mathbf{y}(\mathbf{u}^*)) = 0$$

for the NavierStokes system, if in the \mathbf{z} equation we neglect the terms $B(\mathbf{z}, \mathbf{z}), B(\mathbf{z}, \mathbf{y}), B(\mathbf{y}, \mathbf{z})$, i.e., here we approximate $\mathbf{y}(\mathbf{u}^*) \approx \mathbf{u}^*$.

The advantage of the LS method consists of the important reduction of the dimension of the nonlinear system to be solved (m is, generally, small) together with the possibility to oversee the approximating errors. This advantage can be remarked in the following example [116], [117], which presents an application of the LS method for the Burgers equation, which means for

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}. \quad (8.23)$$

First, we will analyze the steady state case

$$\begin{aligned} u_{xx} &= uu_x - f, \quad x \in (-1, 1), \\ u(-1) &= u(1) = 0 \end{aligned} \quad (8.24)$$

where, for instance, $f(x) = 2x^2 - 2x + 2$ is chosen such that $u(x) = 1 - x^2$ is an exact solution of this problem.

Using the above notation, in our case $Lu = u_{xx}$, $Nu = uu_x - f$, $S = L^2(-1, 1)$, $D(L) = D(N) = \{u \in C^2(-1, 1) \cap C[-1, 1], u(-1) = u(1) = 0\}$ endowed with the uniform norm μ . The spectral problem

$$\begin{aligned} u_{xx} &= \lambda u, \\ u(-1) &= u(1) = 0 \end{aligned}$$

admits the eigenfunctions $\Phi_k(x) = \sin \frac{k\pi}{2}(x + 1)$ and the eigenvalues $\lambda_k = -(k\pi/2)^2$ satisfying the needed conditions.

We look for the solution of the problem (8.23) as a truncated Fourier series

$$u(x) = \sum_{k=1}^N c_k \sin \frac{k\pi}{2}(x + 1).$$

Let $1 \leq m \leq N$. Then

$$u^* = P_m u = \sum_{k=1}^m c_k \sin \frac{k\pi}{2}(x + 1)$$

and

$$H_m u = - \sum_{k=m+1}^{\infty} c_k \frac{4}{k^2 \pi^2} \sin \frac{k\pi}{2}(x + 1).$$

The coefficients of the development of f with respect to this system are

$$f_k = \int_{-1}^1 (2x^2 - 2x + 2) \sin \frac{k\pi}{2}(x + 1) dx = \begin{cases} \frac{192}{k^3 \pi^3}, & k \text{ even} \\ \frac{8}{k\pi}, & k \text{ odd.} \end{cases}$$

In this case we have

$$\begin{aligned} Nu &= \sum_{i,j=1}^N c_i \sin \frac{i\pi}{2}(x + 1) c_j \frac{j\pi}{2} \cos \frac{j\pi}{2}(x + 1) \\ &\quad - \sum_{k=1}^N f_k \sin \frac{\pi k}{2}(x + 1) \end{aligned}$$

from which the Fourier coefficients for Nu can be obtained at once:

$$C_k = \sum_{i,j=1}^N c_i c_j \frac{j\pi}{2} \int_{-1}^1 \sin \frac{i\pi}{2}(x + 1) \cos \frac{j\pi}{2}(x + 1) \sin \frac{k\pi}{2}(x + 1) dx - f_k$$

$$= \frac{\pi}{4} \left(\sum_{j=1}^{k-1} j c_j c_{k-j} - k \sum_{j=1}^{N-k} c_j c_{k+j} \right).$$

Consequently, the iterations which lead to the associated element for u^* are

$$c_k^{s+1} = -\frac{1}{k^2 \pi} \left(\sum_{j=1}^{k-1} j c_j^s c_{k-j}^s - k \sum_{j=1}^{N-k} c_j^s c_{k+j}^s \right) + \frac{4f_k}{k^2 \pi^2} \quad (8.25)$$

for $k = m + 1, \dots, N$. Obviously, the elements c_1, \dots, c_m are fixed. Reiterating for $s = 0, 1, \dots, S$, when the desired accuracy is obtained (and for sufficiently large m it necessarily happens),

$$u^{S+1} = \sum_{k=1}^m c_k \Phi_k + \sum_{k=m+1}^N C_k^S \Phi_k$$

represents an approximation of the associated function $U(u^*)$.

The determining equations become

$$-c_k \left(\frac{k\pi}{2} \right)^2 = - \left(\frac{k\pi}{2} \right)^2 C_k^{S+1}, \quad k = 1, \dots, m,$$

which means

$$g_k \equiv c_k - C_k^{S+1} = 0, \quad k = 1, \dots, m. \quad (8.26)$$

When the equation (8.24) has a solution then, for sufficiently large m , the system (8.26) will also have a solution which can be approximated. Such a procedure using the data c_1, \dots, c_m computes g_1, \dots, g_m by the above iterative process (8.25), and on the base of the obtained results it will improve the initial data c_1, \dots, c_m . The cycle is recomputed until the requested accuracy is achieved and the associated function of the final iteration represents an approximation of the solution of the problem (8.24).

In the sequel we will consider the unsteady problem

$$\begin{aligned} u_{xx} &= u_t + uu_x - f, & x \in (-1, 1), t > 0, \\ u(x, 0) &= u_0(x), & x \in (-1, 1), \\ u(-1, t) &= u(1, t) = 0, & t > 0, \end{aligned} \quad (8.27)$$

where, for numerical computations, we will take

$$f(x, t) = (1 - x^2) \cos t + 2 \sin t - 2x \sin^2 t + 2x^3 \sin^2 t$$

so that the problem has the exact solution $u(x, t) = (1 - x^2) \sin t$.

The main difference with respect to the previous case consists in the structure of the operator N where now the term u_t is involved. Supposing we have calculated the solution u_j at the time level t_j , the auxiliary equation becomes at the time level t_{j+1} ,

$$u_{j+1}^0 = u_j, u_{j+1}^{s+1} = u_{j+1}^* + H_m \left(Nu_{j+1}^s + \frac{u_{j+1}^s - u_j}{\delta t} \right), s = 0, 1, \dots$$

where δt is the time step. If m is sufficiently large, the above iterations converge towards the associated function $U(u_{j+1}^*)$. This would be a solution of the problem (8.27) if

$$\frac{\partial^2 u_{j+1}^*}{\partial x^2} = P_m N U(u_{j+1}^*) + \frac{\partial u^*}{\partial t} \Big|_{t=t_{j+1}}. \tag{8.28}$$

In this case the coefficients of $f(x, t)$ are

$$f_k(t) = \begin{cases} \frac{32}{k^3 \pi^3} \cos t + \frac{8}{k\pi} \sin t & k \text{ odd} \\ \frac{192}{k^3 \pi^3} \sin^2 t & k \text{ even.} \end{cases}$$

Since

$$u^*(x, t) = \sum_{k=1}^m c_k^*(t) \sin \frac{k\pi}{2}(x + 1),$$

the equations (8.28) represent a system of differential equations with respect to the unknown functions $c_1(t), \dots, c_m(t)$, with the initial conditions $c_1(0) = 0, \dots, c_m(0) = 0$.

To the system (8.28) of the form $u' = F(t, u)$, one could apply different numerical procedures in order to get an approximate solution. For instance, a predictor-corrector procedure involves

$$\begin{aligned} \bar{u}_{j+1} &= u_j + \delta t F(t_j, u_j), \\ u_{j+1} &= u_j + \frac{\delta t}{2} [F(t_j, u_j) + F(t_{j+1}, \bar{u}_{j+1})] \end{aligned}$$

where the corrector can be recalculated. The result of the numerical integration represents u^* at the time level t_{j+1} . The associated function for u_{j+1}^* is then an approximation of the solution of the problem (8.27) at the time level t_{j+1} .

The algorithm of this procedure is then the following:

One knows the approximative solution at the time level t_j , its coefficients being B_1, \dots, B_N ,

1. We evaluate

$$F_k(t_j, B_1, \dots, B_m) = - \left(\frac{k\pi}{2} \right)^2 B_k - \int_{-1}^1 uu_x \Phi_k dx + f_k(t_j);$$

2. We calculate the predictor

$$\bar{c}_k = B_k + \delta t F_k(t_j, B_1, \dots, B_m), \quad k = 1, \dots, m;$$

3. We calculate the associated function $U(\bar{u}^*)$ for $\bar{c}_1, \dots, \bar{c}_m$ as the limit of the sequence

$$u^{s+1} = \bar{u}^* + H_m(Nu^s - f(t_{j+1})) - \sum_{k=m+1}^N \frac{4}{k^2\pi^2} \frac{\bar{c}_k^s - B_k}{\delta t} \Phi_k$$

where the iterations stop at a convenient rank S ;

4. We evaluate

$$F_k(t_{j+1}, \bar{c}_1, \dots, \bar{c}_m) = - \left(\frac{k\pi}{2} \right)^2 \bar{c}_k - \int_{-1}^1 UU_x \Phi_k dx + f_k(t_{j+1})$$

for $k = 1, \dots, m$;

5. We calculate the ‘‘corrected’’ c_1, \dots, c_m ,

$$c_k = B_k + \frac{\delta t}{2} [F_k(t_j, B) + F_k(t_{j+1}, \bar{c})], \quad k = 1, \dots, m.$$

The steps 3,4,5 are repeated if necessary;

6. We calculate the associated function for c_1, \dots, c_m as the limit of the sequence

$$u^{s+1} = u^* + H_m(Nu^s - f(t_{j+1})) - \sum_{k=m+1}^N \frac{4}{k^2\pi^2} \frac{c_k^s - B_k}{\delta t} \Phi_k$$

where, again, the iterations are stopped at a convenient rank S . In c_1, \dots, c_N we now have the coefficients of the approximate solution of the problem (8.27) at the next time level t_{j+1} ;

7. The values obtained through the new c_k enter into $B_k, k = 1, \dots, N$ and step 1 is repeated for a new level of time.

In what follows we will present some numerical results, taken from [116]. The problem (8.27) was also solved, for comparison, by the finite difference method (Crank–Nicolson for u_{xx} and forward Euler for uu_x), that is

$$\frac{u_k^{j+1}}{\delta t} - \frac{u_{k+1}^{j+1}}{2\delta x^2} + \frac{u_k^{j+1}}{\delta x^2} - \frac{u_{k-1}^{j+1}}{2\delta x^2} = \frac{u_{k+1}^j - 2u_k^j + u_{k-1}^j}{2\delta x^2} - u_k^j \frac{u_{k+1}^j - u_k^j}{\delta x} + f_k^j$$

Table 8.1. The errors for the Burgers equation

| j | finite diff | LS Euler | LS pred-corr |
|----------|--------------------|-----------------|---------------------|
| 1 | 0,009 | 0,00027 | -0,00011 |
| 2 | 0,017 | 0,00088 | -0,00025 |
| 3 | 0,023 | 0,0016 | -0,00042 |
| 4 | 0,027 | 0,0028 | -0,00061 |
| 5 | 0,031 | 0,0040 | -0,00082 |
| 6 | 0,035 | 0,0053 | -0,00100 |

and by the LS method using the Euler method in (8.28) and predictor-corrector as described in the algorithm.

Table 8.1 contains the maximal errors with respect to the exact solution, for the above three algorithms at time levels $t_j = j\pi/32, j = 1, \dots, 6$ and for $m = 2, N = 16$. While growing m in order to accelerate the convergence of iterations the phenomenon of the instability of the numerical calculation is remarked, but by diminishing the time step size δt the stability is kept up.

5. Examples

5.1 Stokes' Problem

We will shortly present in this section, following [14], the particular treatment of the Stokes problem by a spectral method. We will use the representation in primitive variables — the velocity and the pressure — due to their capability of extending towards 3D problems for which the other formulations are less applicable. So, let us consider the problem

$$\begin{aligned}
 \gamma \mathbf{u} + \Delta \mathbf{u} + \nabla p &= \mathbf{f}, & \text{in } \Omega, \\
 \nabla \cdot \mathbf{u} &= 0, & \text{in } \Omega, \\
 \mathbf{u}|_{\partial\Omega} &= 0,
 \end{aligned}
 \tag{8.29}$$

where Ω is a bounded domain in \mathbb{R}^2 . The term $\gamma \mathbf{u}$ from the first equation may be derived, for instance, from a time discretization of a unsteady Stokes problem and in this case $\gamma \geq 0$.

We are looking for a solution in the spaces

$$\begin{aligned}
 \mathbf{u} \in V &= \{ (u_1, u_2) \mid u_i \in H_0^1(\Omega), i = 1, 2 \}, \\
 p \in Q &= \left\{ p \in L^2(\Omega) \mid \int_{\Omega} p dx = 0 \right\}
 \end{aligned}$$

based on a variational formulation. For that, we will introduce the bilinear forms

$$e(\mathbf{u}, \mathbf{v}) = \int_{\Omega} (\gamma \mathbf{u} \mathbf{v} + \nabla \mathbf{u} \nabla \mathbf{v}) dx : V \times V \rightarrow \mathbb{R},$$

$$d(\mathbf{v}, q) = - \int_{\Omega} (\nabla \cdot \mathbf{v}) q dx : V \times Q \rightarrow \mathbb{R}$$

and the linear form

$$(f, \mathbf{v}) = \int_{\Omega} f \mathbf{v} dx : V \rightarrow \mathbb{R}$$

and the problem (8.29) may be reformulated as

$$e(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, p) = (f, \mathbf{v}), \quad \forall v \in V,$$

$$d(\mathbf{u}, q) = 0, \quad \forall q \in Q$$

with the unknowns $\mathbf{u} \in V$ and $p \in Q$.

The discretization of this problem may be made by decomposition of Ω into spectral elements and by using a space of polynomial on spectral elements functions

$$\mathcal{P}_N = \{q|q \in P_N \text{ on each element from } \Omega\}.$$

We will use the finite dimensional subspaces $V_N \subset V$ respectively $Q_N \subset Q$ and the (linear) and bilinear approximative forms

$$e_N : V_N \times V_N \rightarrow \mathbb{R},$$

$$d_n : V \times Q \rightarrow \mathbb{R},$$

$$(f, \mathbf{v})_N : V_N \rightarrow \mathbb{R}$$

based on the Gauss quadrature formulas. The discrete Stokes problem becomes

$$e_N(\mathbf{u}_N, \mathbf{v}) + d_N(\mathbf{v}, p_N) = (f, \mathbf{v})_N, \quad \forall v \in V_N, \quad (8.30)$$

$$d_N(\mathbf{u}_N, q) = 0, \quad \forall q \in Q_N$$

with the unknowns $\mathbf{u}_N \in V_N$ and $p_N \in Q_N$.

In a matrix form, the above system can be written

$$E_N \underline{u} + G_N \underline{p} = \underline{f}_N, \quad (8.31)$$

$$D_N \underline{u} = 0,$$

where G_N is the discrete gradient matrix and $D_N = G_N^T$ is the discrete divergence matrix.

The sufficient conditions for stability and consistency of the above scheme are (*Inf-Sup conditions*)

$$\begin{aligned} \exists \alpha_N : e_N(\mathbf{v}, \mathbf{v}) &\geq \alpha_N \|\mathbf{v}\|_V^2, & \forall \mathbf{v} \in V_N, \\ \exists \beta_N : \sup_{\mathbf{v} \in V_N} \frac{d_N(\mathbf{v}, q)}{\|\mathbf{v}\|_V} &\geq \beta_N \|q\|_Q, & \forall q \in Q_N. \end{aligned} \tag{8.32}$$

These conditions are also of practical interest, as they influence the convergence of the discrete solutions toward the exact solutions as $N \rightarrow \infty$ or the solvability of the Stokes discrete problem for a fixed N .

Indeed, one may show that

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_N\|_V &\leq \left(1 + \frac{C}{\alpha_N}\right) \inf_{\mathbf{w}_N \in K_N} \|\mathbf{u} - \mathbf{w}_N\|_V + \frac{C}{\alpha_N} \inf_{q_N \in Q_N} \|p - q_N\|_Q, \\ \|p - p_N\|_Q &\leq \left(1 + \frac{C}{\alpha_N}\right) \frac{C}{\beta_N} \inf_{\mathbf{w}_N \in K_N} \|\mathbf{u} - \mathbf{w}_N\|_V \\ &\quad + \frac{C}{\alpha_N \beta_N} \inf_{q_N \in Q_N} \|p - q_N\|_Q \end{aligned}$$

where

$$K_N = \{\mathbf{v} \in V_N \mid d_N(\mathbf{v}, q) = 0 \ \forall q \in Q_N\}.$$

The constants α_N, β_N indicate, if they are small, the non-optimality of the discrete method.

But the system (8.31) may be reduced to a scalar equation for the pressure (*Uzawa method*)

$$D_N E_N^{-1} G_N p = D_N E_N^{-1} \underline{f}_N. \tag{8.33}$$

The proper eigenvalues of the discrete Uzawa operator $D_N E_N^{-1} G_N : Q_N \rightarrow Q'_N$ satisfy

$$\frac{\lambda_{\max}}{\lambda_{\min}} \cong \frac{C}{\alpha_N \beta_N^2}$$

and small values for α_N, β_N again indicate a poorly conditioned operator, hence a slow convergence of the iterative methods used to solve the problem (8.33).

We remark that for most of the existing schemes, α_N is independent of N but β_N depends on N by the interaction of the pressure modes with the velocity field. If we fix $q \in \mathcal{P}_N$ and require

$$\beta_N(q) \equiv \sup_{\mathbf{v} \in V_N} \frac{d_N(\mathbf{v}, q)}{\|\mathbf{v}\|_V \|q\|_Q} > 0,$$

then $\beta_N = \inf_{q \in Q_N} \beta_N(q)$. We have three possible situations:

a) $\beta_N(q) = 0$. Such a type of pressure, called *spurious*, must be eliminated from Q_N in order to satisfy the condition (8.32).

b) $\beta_N(q) \cong CN^{-s}$ as $N \rightarrow \infty$, for some $s > 0$. These are *weakly-spurious* modes and yield $\beta_N \rightarrow 0$ as $N \rightarrow \infty$.

c) $\beta_N(q) = O(1)$ as $N \rightarrow \infty$. These are the *essential* pressure modes, the “good” ones, with respect to the conditions (8.32).

Concluding, we can state that

$$\mathcal{P}_N = \mathcal{S}_N \oplus \mathcal{WS}_N \oplus \mathcal{E}_N$$

and the discrete pressure p_N belongs to a subspace $Q_N \subset \mathcal{P}_N$ so that $Q_N \cap \mathcal{S}_N = 0$.

We remark that the above variational formulation may be extended also to the cases in which $G_N \neq D_N^T$, by using Inf-Sup generalized conditions.

Let us illustrate the above numerical algorithms on the computational domain $(-1, 1) \times (-1, 1)$ and look for $\mathbf{u}_N \in (P_N(\Omega))^2$ and $p_N \in P_N(\Omega)$. Both the momentum and the continuity equation are collocated on a Gauss-Lobatto grid in each direction $GL_N \otimes GL_N$. This method, called (P_N, P_N) leads to $\nabla \cdot \mathbf{u}_N = 0$ but also leads to the appearance of the pressure spurious modes

$$\mathcal{S}_N = \text{span} \{L_N(x), L_N(y), L_N(x)L_N(y), L'_N(x)(1 \pm x)L'_N(y)(1 \pm y)\}$$

within a seven-dimensional space, together with the constant mode. For 3D problems, we have $\dim \mathcal{S}_N = 12N + 3$.

These spurious modes must be filtered. One of the possibilities is the reduction of the dimension of the space where the pressure is approximated i.e., $p_N \in P_{N-2}(\Omega)$. The momentum equation is collocated as above but the continuity equation is collocated on $GL_{N-2} \otimes GL_{N-2}$, obtaining the so-called (P_N, P_{N-2}) method. It is possible to use a single grid, where the continuity equation is collocated on $GL_N^* \otimes GL_N^*$, with $GL_N^* = GL_N \setminus \{\pm 1\}$, see [6]. Now there are no spurious modes for the pressure but $\nabla \cdot \mathbf{u}_N \neq 0$. In this case, the numerical calculations indicate $\beta_N \cong O(N^{-1/8})$.

Let us use the (P_N, P_{N-2}) method for the unsteady Stokes problem

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} - \nabla^2 \mathbf{u} + \nabla p &= \mathbf{f}, & \text{in } \Omega &= (-1, 1)^2, \\ \nabla \cdot \mathbf{u} &= 0, & \text{in } \Omega, \\ \mathbf{u}|_{\partial\Omega} &= 0, \end{aligned}$$

where $\mathbf{u} = (u_1, u_2)$. By time discretization using the backward Euler method (for simplicity), we find

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} - \nabla^2 \mathbf{u}^{n+1} + \nabla p^{n+1} &= \mathbf{f}^{n+1}, & \text{in } \Omega = (-1, 1)^2, \\ \nabla \cdot \mathbf{u}^{n+1} &= 0, & \text{in } \Omega, \\ \mathbf{u}^{n+1}|_{\partial\Omega} &= 0, \end{aligned} \tag{8.34}$$

for $n = 0, 1, \dots$

Let N be an even natural number and P_N^0 the space of the polynomials of degree N satisfying the homogeneous Dirichlet conditions on the boundary ± 1 . We will approximate the velocity components by polynomials from P_N^0 and the pressure by polynomials from P_{N-2} . We will use the Chebyshev-collocation discretization, whose derivative matrix $D \in \mathbb{R}^{N+1, N+1}$ is defined in MATLAB.

By eliminating the boundary conditions, we obtain

$$u'(x_j) = (D^0 u)_j, \quad j = 1, \dots, N - 1$$

where $u = (u(x_1), \dots, u(x_{N-1}))^T$, $x_j = \cos \frac{j\pi}{N}$, $j = 1, \dots, N - 1$. The matrix $D^0 \in \mathbb{R}^{N-1, N-1}$ is obtained from the matrix D by eliminating the first and the last rows and columns. Similarly we discretize the second derivative by the matrix D^2 and, considering the boundary conditions, we also obtain

$$u''(x_j) = (D^{2,0} u)_j, \quad j = 1, \dots, N - 1.$$

In order to avoid the interpolation between different grids, we will use for the pressure discretization the same nodes x_j . The derivative operator is now defined by constructing the interpolation polynomial p for these $N - 1$ nodes, then by differentiating and taking the derivative values on these nodes. So we obtain a new derivative matrix D_p for which

$$p'(x_j) = (D_p p)_j, \quad j = 1, \dots, N - 1$$

and this does not use boundary conditions at ± 1 . The elements of D_p are

$$(D_p)_{i,j} = \frac{1 - x_j^2}{1 - x_i^2} D_{i,j} + \frac{2x_i}{1 - x_i^2} \delta_{i,j}, \quad i, j = 1, \dots, N - 1.$$

In the two-dimensional case, the discrete derivative operators may be expressed by tensorial products. Let us consider the mesh (x_i, y_j) , $i, j = 1, \dots, N - 1$ where $x_i = \cos \frac{i\pi}{N}$, $y_j = \cos \frac{j\pi}{N}$ and let us represent $u(x_i, y_j)$

by the matrix $(u)_{i,j}$. If we reassemble the matrix u into the column vector \underline{u} , built by the columns of u written one by one, the derivative matrices for the components of the velocity become

$$\begin{aligned} D_x^0 &= \text{kron}(I, D^0), D_y^0 = \text{kron}(D^0, I), \\ D_{xx}^0 &= \text{kron}(I, D^{2,0}), D_{yy}^0 = \text{kron}(D^{2,0}, I), \\ D_{\Delta}^0 &= D_{xx}^0 + D_{yy}^0 \end{aligned}$$

of dimensions $(N-1)^2 \times (N-1)^2$, where I is the unit $(N-1)$ matrix. Here $\text{kron}(A, B)$ is the Kronecker tensorial product of the matrices A and B i.e., a matrix built by taking all the possible products of the elements of A and B . For instance, if A is a 2×3 matrix, then $\text{kron}(A, B)$ will be the matrix

$$\begin{pmatrix} A(1,1) * B & A(1,2) * B & A(1,3) * B \\ A(2,1) * B & A(2,2) * B & A(2,3) * B \end{pmatrix}.$$

For the pressure we will have

$$D_x = \text{kron}(I, D_p), D_y = \text{kron}(D_p, I).$$

The discretization of the Stokes system (8.30) is

$$\begin{aligned} \left(-D_{\Delta}^0 + \frac{1}{\Delta t} I\right) \underline{u}_1^{n+1} + D_x \underline{p}^{n+1} &= \underline{f}_1^{n+1} + \frac{1}{\Delta t} u_1^n \equiv \widehat{f}_1^{n+1}, \\ \left(-D_{\Delta}^0 + \frac{1}{\Delta t} I\right) \underline{u}_2^{n+1} + D_y \underline{p}^{n+1} &= \underline{f}_2^{n+1} + \frac{1}{\Delta t} u_2^n \equiv \widehat{f}_2^{n+1}, \\ D_x u_1^{n+1} + D_y u_2^{n+1} &= 0, \end{aligned}$$

where $u_1^{n+1} = u_2^{n+1} = 0$ at $\partial\Omega$. From the Uzawa decoupling, by expressing the components of the velocity from the first two equations and replacing them into the last equation, we obtain the following equation for the pressure

$$A \underline{p}^{n+1} = \widehat{f}^{n+1} \quad (8.35)$$

where

$$A = D_x^0 \left(-D_{\Delta}^0 + \frac{1}{\Delta t} I\right)^{-1} D_x + D_y^0 \left(-D_{\Delta}^0 + \frac{1}{\Delta t} I\right)^{-1} D_y$$

and

$$\widehat{f}^{n+1} = D_x^0 \left(-D_{\Delta}^0 + \frac{1}{\Delta t} I\right)^{-1} \widehat{f}_1^{n+1} + D_y^0 \left(-D_{\Delta}^0 + \frac{1}{\Delta t} I\right)^{-1} \widehat{f}_2^{n+1}.$$

After the calculation of the pressure \underline{p}^{n+1} , from the first two equations we may compute the components of the velocity. We remark that if A has a single zero eigenvalue, the discrete problem does not allow spurious modes for the pressure. The constant mode, which is present in the continuous case too, where the pressure may be calculated within an additive constant, may be eliminated by imposing either $\int p^{n+1} dx = 0$ or, more practically, the vanishing of \underline{p}^{n+1} at a mesh node.

The case of the steady state Stokes problem may be obtained as the above by considering $\Delta t \rightarrow \infty$ and then

$$A_{stat} = -D_x^0 (D_\Delta^0)^{-1} D_x - D_y^0 (D_\Delta^0)^{-1} D_y.$$

This matrix has, besides one zero eigenvalue, only real, positive eigenvalues and it is very well conditioned. This allows us to solve the pressure equation (8.35) by direct methods (for a small N) or iterative, like the conjugate gradient method (for large N).

The numerical solution of a particular problem, such as

$$\begin{aligned} p_x - \Delta u &= 2 + \pi \cos(\pi x) \sin(\pi y), & \text{in } \Omega = (-1, 1) \times (-1, 1), \\ p_y - \Delta v &= \pi \sin(\pi x) \cos(\pi y), & \text{in } \Omega, \\ u_x + v_y &= 0, & \text{in } \Omega, \\ u|_{\partial\Omega} = v|_{\partial\Omega} &= 0, \end{aligned}$$

whose exact solution is $u = v = 0$, $p = 2x + \sin(\pi x) \sin(\pi y)$, may be performed by the MATLAB code

```
n=16;
[x,d,dd,dp]=derceb(n);
f1=2+pi*cos(pi*(x(2:n)))'*sin(pi*(x(2:n)));
f2=pi*sin(pi*(x(2:n)))'*cos(pi*(x(2:n)));
I=speye(n-1);
dx0=kron(I,d(2:n,2:n));dy0=kron(d(2:n,2:n),I);
ddel0=kron(dd(2:n,2:n),I)+kron(I,dd(2:n,2:n));
dx=kron(I,dp);dy=kron(dp,I);
dm1=inv(-ddel0);
A=zeros((n-1)^2+1,(n-1)^2);
A=dx0*dm1*dx+dy0*dm1*dy;A((n-1)^2+1,((n-1)^2+1)/2)=1;
F1=reshape(f1,(n-1)^2,1);F2=reshape(f2,(n-1)^2,1);
F=dx0*dm1*F1+dy0*dm1*F2;F((n-1)^2+1)=0;
P=A\F;
U=dm1*(F1-dx*P);V=dm1*(F2-dy*P);
u=reshape(U,n-1,n-1);v=reshape(V,n-1,n-1);
p=reshape(P,n-1,n-1);
uex=zeros(n-1);vex=zeros(n-1);
```

```

pex=sin(pi*(x(2:n)))'*sin(pi*(x(2:n)))+2*x(2:n)'. . .
ones(1,n-1);
surf(x(2:n),x(2:n),abs(u-uex));pause;
surf(x(2:n),x(2:n),abs(v-vex));pause;
surf(x(2:n),x(2:n),abs(p-pex));

```

which uses the following subprogram to calculate the derivative matrices

```

function [x,d,dd,dn,dp]=derceb(n)
k=0:n;x=cos(pi.*k./n);c=ones(n+1);c(1)=2;c(n+1)=2;
d=gallery('chebspec',n+1,0);
dd=d*d;
dp=zeros(n-1,n-1);
for i=1:n-1
  for j=1:n-1
    dp(i,j)=(1-x(j+1)^2)/(1-x(i+1)^2)*d(i+1,j+1);
  end;
  dp(i,i)=dp(i,i)+2*x(i+1)/(1-x(i+1)^2);
end;

```

As N is small, the algebraic system may be directly solved and the computational errors for the velocity are of the order 0.8×10^{-10} and for the pressure of the order 3×10^{-10} .

5.2 Correction in the Dominant Space

We will now present, following [146], [118], [119], an improved algorithm for the numerical calculation of the solutions of some differential systems, coming from the spectral discretization of certain fluid dynamics equations. We will consider equations of the type

$$\frac{\partial u}{\partial t} = Lu + f, \text{ in } \Omega \times [0, 1], \quad \Omega \subset \mathbb{R}^d$$

with joined suitable boundary and initial conditions. Here L is a differential operator with respect to spatial variables; the given function f and the unknown function u are assumed to be sufficiently smooth for the following calculations.

By spectral discretization with respect to the spatial variables, with N_i nodes on each dimension, we obtain a differential system of the form

$$\frac{\partial V}{\partial t} + F(t, V), \quad V \in \mathbb{R}^N,$$

where $F : \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R}^N$, $N = N_1 \times \cdots \times N_d$. That system is large and stiff, so a numerical integration by particular implicit methods should be necessary, in order to avoid the strong restraint on the time step size

imposed by the stability of the calculations which arises in the case of many explicit methods. The improvement of the algorithm consists in the use of implicit methods only for the dominant directions (associated to eigenvalues of the largest magnitude), while the system is explicitly integrated.

We will describe the method for the bidimensional diffusion equation

$$u_t = u_{xx} + u_{yy}, \quad \text{in } \Omega = (-1, 1) \times (-1, 1), \quad t > 0,$$

$$u|_{\partial\Omega} = 0, \quad u(x, y, 0) = u_0(x, y).$$

Using for the spatial discretization the Chebyshev-collocation spectral method, with the Gauss–Lobatto nodes $x_i = \cos(i\pi/N)$, $y_j = \cos(j\pi/N)$, $i, j = 0, \dots, N$, we obtain

$$U' = D_0^2 U + U D_0^{2T}, \quad U(0) = U_0 \quad (8.36)$$

where U is the matrix $u(x_i, y_j, t)$, $i, j = 1, \dots, N - 1$, D_0^2 is the second order derivative matrix in the presence of the homogeneous boundary conditions on Gauss–Lobatto nodes and D_0^{2T} is its transposed matrix.

The exact solution, in matrix form, is

$$U(t_n + h) = e^{D_0^2 h} U(t_n) e^{D_0^{2T} h}.$$

Let λ_k, v_k, w_k for $k = 1, \dots, N - 1$ be the eigenvalues, respectively the right and left eigenvectors of the matrix D_0^2 . As we know, λ_k are real, distinct, negative, the largest in magnitude being of order $O(N^4)$. Then

$$e^{D_0^2 h} U(t_n) = \sum_{k=1}^{N-1} e^{\lambda_k h} v_k w_k^T U(t_n).$$

The explicit numerical methods replace the above matrix by a truncated sum of the exponential matrix

$$\sum_{k=0}^M \frac{(D_0^2)^k h^k}{k!} U(t_n)$$

from which obviously results the need for a very small step, in order to ensure stability. In the method of the dominant space correction, the solution is approximated by an explicit method, followed by a correction on the dominant directions such that the coefficients of the explicit solution corresponding to eigenvalues of large magnitude are replaced by the coefficients of the exact solution.

In the case of the modified Euler method, for instance, the algorithm is

$$T = I + D_0^2 h + \frac{(D_0^2)^2 h^2}{2} + \sum_{k=1}^M \left(e^{\lambda_k h} - 1 - \lambda_k h - \frac{\lambda_k^2 h^2}{2} \right) v_k w_k^T,$$

$$U_{n+1} = T U_n T^T, \quad n = 0, 1, 2, \dots$$

where $\lambda_1, \dots, \lambda_M$ are the dominant eigenvalues (of the largest magnitude). We observe that the matrix T may be precomputed and consequently, the whole calculation is explicit. The constraints on the time step size are those to be imposed if the dominant values do not exist, so the choice of the time step may be made only from the accuracy requirements.

We will present, following [119], an application for a bidimensional fluid through a grooved channel. The equations are

$$\mathbf{v}_t = \mathbf{v} \times \boldsymbol{\omega} - \nabla p + \frac{1}{R} \nabla^2 v,$$

$$\nabla \cdot v = 0,$$

where \mathbf{v} is the velocity, $\boldsymbol{\omega} = \nabla \times \mathbf{v}$, p is the pressure and the computational domain D is the reunion of the rectangles $A = [0, 2] \times [0, 2]$, $B = [-3, 0] \times [0, 2]$, $C = [-3, 0] \times [-1.68, 0]$. The fluid enters through $\{-3\} \times [0, 2]$ and exits through $\{2\} \times [0, 2]$, the periodical boundary ∂D_P of the domain D , the other being the solid boundary ∂D_S . The spatial discretization is made by the spectral element method (see [3]). The time marching is performed by a fractional step scheme; starting from v^n we perform:

1. The nonlinear step:

$$\frac{\bar{\mathbf{v}}^{n+1} - \mathbf{v}^n}{h} = \sum_{i=0}^2 \beta_i (\mathbf{v}^{n-1} \times \boldsymbol{\omega}^{n-1}) \text{ on } D,$$

where $\beta_0 = 23/12$, $\beta_1 = -16/12$, $\beta_2 = 5/12$ (the third order Adams–Bashforth scheme).

2. The pressure step:

$$\frac{\bar{\mathbf{v}}^{n+1} - \bar{\mathbf{v}}^n}{h} = -\nabla p^n \text{ in } D,$$

$$\nabla \cdot \bar{\mathbf{v}}^{n+1} = 0 \text{ in } D,$$

$$\bar{\mathbf{v}}^{n+1} \cdot \mathbf{n} = 0 \text{ on } \partial D.$$

Taking the divergence of the first equation and using also the second equation, we find

$$\begin{aligned} \nabla^2 p^n &= \frac{\nabla \cdot \bar{\mathbf{v}}^{n+1}}{h} \text{ in } D, \\ \nabla p^n \cdot \mathbf{n} &= \frac{\bar{\mathbf{v}}^{n+1} \cdot \mathbf{n}}{h} \text{ on } \partial D_S. \end{aligned}$$

This problem is solved by a $P_N \times P_{N-2}$ method, which avoids the spurious solutions. The discretized system comes from the discretization on A, B, C by adding the smoothness conditions on the inside boundaries (see [119]).

3. The viscous step:

$$\begin{aligned} \frac{\mathbf{v}^{n+1} - \bar{\mathbf{v}}^{n+1}}{h} &= \frac{1}{R} \nabla^2 \mathbf{v}^{n+1} \text{ in } D, \\ \mathbf{v}^{n+1} &= 0 \text{ on } \partial D_S, \quad \mathbf{v}^{n+1}(x+L, y) = \mathbf{v}^{n+1}(x, y) \text{ on } \partial D_P. \end{aligned}$$

In this step the dominant space correction is used. The rectangles A, B, C are mapped into the standard domain $[-1, 1] \times [-1, 1]$ by affine transformations. The x derivative takes into account the periodicity conditions on ∂D_P and the matching conditions on the interface $A - B$, thus resulting in the derivative matrix D_P^2 of order $2(N - 1)$ on A, B and yet using D_0^2 of order $N - 1$ on C . Similarly, the y derivative uses the Dirichlet conditions on ∂D_S and the matching conditions on the interface $B - C$, so resulting in the derivative matrix D_D^2 of order $2(N - 1)$ on B, C and keeping D_0^2 of order $N - 1$ on A .

The system (8.36), written with block matrices, becomes

$$\begin{aligned} \mathbf{v}_{At} &= \frac{1}{R} [D_{P11}^2 \mathbf{v}_A + D_{P12}^2 \mathbf{v}_B + \mathbf{v}_A D_0^{2T}] \text{ in } A, \\ \mathbf{v}_{Bt} &= \frac{1}{R} [D_{P21}^2 \mathbf{v}_A + D_{P22}^2 \mathbf{v}_B + \mathbf{v}_B D_{D11}^{2T} + \mathbf{v}_C D_{D12}^{2T}] \text{ in } B, \\ \mathbf{v}_{Ct} &= \frac{1}{R} [D_0^2 \mathbf{v}_C + \mathbf{v}_B D_{D21}^{2T} + \mathbf{v}_C D_{D22}^{2T}] \text{ in } C. \end{aligned}$$

We remark that we should calculate by iterative methods and record only the dominant eigenvectors of the matrices D_P^2, D_D^2, D_0^2 , respectively the corresponding eigenvalues, in order to apply the correction in the dominant space (generated by the right-hand side of the above system) method. So, the time step size restrictions to ensure stability will be imposed only by the remaining eigenvalues.

Appendix A

Vectorial-Tensorial Formulas

In what follows we intend to give a brief overview on some basic concepts and results which have been used throughout this book. Most of these results represent some vectorial-tensorial relations and they can be established by direct calculation.

First of all we will present a summary of some properties joined to the fundamental concept of a tensor (in general) and of a Cartesian tensor of order 2 in E_3 (in particular).

A. The natural way to define a Cartesian tensor of order 2 in the (Euclidian) vectorial space E_3 is to consider it as an element of the dyadic product $E_3 \otimes E_3$, i.e., an entity of the form¹

$$[\mathbf{T}] = \tau_{ik} \mathbf{i}_i \otimes \mathbf{i}_k.$$

At the same time the tensor $[\mathbf{T}]$ could be seen as a linear application (mapping) of the Euclidean space E_3 onto itself. If this linear application has the components (coordinates, matrix) τ_{ij} , (that is $[\mathbf{T}]\mathbf{i}_j = \tau_{ij}\mathbf{i}_i$, \mathbf{i}_k being an orthonormal basis in E_3) through a transformation of coordinates (change of basis) defined by $\mathbf{i}' = Q_{ik}\mathbf{i}_k$, all these components will change according to the rule $\tau'_{ij} = Q_{ik}Q_{jl}\tau_{kl}$ which represents also a criterion to define such a tensor.

As regards the sum of two tensors and the multiplication by a scalar (tensor of order zero) as well, they could be defined by the corresponding operations on the matrix associated to the linear application (τ_{ij}). The tensor $[\mathbf{0}]$ is the “zero tensor” which maps any vector on the zero vector of E_3 having also the matrix (0) while $[\mathbf{I}]$ is the “unit tensor” which applies any vector to itself, having as components (δ_{ij}). By the “product” of two tensors $[\mathbf{T}]$ and $[\mathbf{S}]$ of matrices (τ_{ij}) and (s_{ij}) respectively, we understand that tensor $[\mathbf{TS}]$ which has the matrix (components) $\tau_{ik}s_{kj}$. Obviously,

¹The dyadic or tensorial product of two vectors $\mathbf{u}(u_i)$ and $\mathbf{v}(v_i)$ from E_3 , denoted by $\mathbf{u} \otimes \mathbf{v}$ is the linear application (mapping) of components $u_i v_j$, that is the application defined by

$$(\mathbf{u} \otimes \mathbf{v}) \mathbf{w} = (\mathbf{w} \cdot \mathbf{v}) \mathbf{u}, \forall \mathbf{w} \in E_3.$$

this product doesn't commute (in general). We will say that a tensor $[\mathbf{T}]$ is invertible if there is a tensor $[\mathbf{T}^{-1}]$ such that $[\mathbf{T}\mathbf{T}^{-1}] = [\mathbf{T}^{-1}\mathbf{T}] = [\mathbf{I}]$.

The successive powers (exponents) of a tensor will be $[\mathbf{T}^0] = [\mathbf{I}]$, $[\mathbf{T}^1] = [\mathbf{T}]$, $\dots, [\mathbf{T}^k] = [\mathbf{T}^{k-1}\mathbf{T}]$.

The transpose of $[\mathbf{T}]$, denoted by $[\mathbf{T}^T]$, is the tensor whose matrix is the transpose matrix of (τ_{ij}) , that is $(\tau_{ij})^T$ while a tensor is symmetric or skew-symmetric if $[\mathbf{S}] = [\mathbf{S}^T]$ or $[\mathbf{\Omega}] = -[\mathbf{\Omega}^T]$ respectively.

The transpose tensor could be also defined by the equality

$$\mathbf{u} \cdot [\mathbf{T}] \mathbf{v} = [\mathbf{T}^T] \mathbf{u} \cdot \mathbf{v}, \forall \mathbf{u}, \mathbf{v} \in E_3.$$

Any tensor $[\mathbf{T}]$ can be uniquely decomposed into the sum of two tensors, one of them $[\mathbf{S}]$ being symmetric while the other $[\mathbf{\Omega}]$ is skew-symmetric, that is $[\mathbf{T}] = [\mathbf{S}] + [\mathbf{\Omega}]$.

A tensor $[\mathbf{T}]$ is called *orthogonal* if and only if $[\mathbf{T}^T] = [\mathbf{T}^{-1}]$ or, equivalently, if it "conserves" the inner (dot) product in the way that $[\mathbf{T}]\mathbf{u} \cdot [\mathbf{T}]\mathbf{v} = \mathbf{u} \cdot \mathbf{v}$, $\forall \mathbf{u}, \mathbf{v} \in E_3$.

The *scalar product of two tensors* of second order $[\mathbf{T}]$ and $[\mathbf{S}]$, denoted by $[\mathbf{T}] \cdot [\mathbf{S}]$, is the scalar $[\mathbf{T}] \cdot [\mathbf{S}] = \tau_{ij} s_{ji}$, where τ_{ij} and s_{ij} are, respectively, the components of the two tensors. Once defined, the inner product, the pre-Hilbertian space of the tensors of order 2 can be also normed by introducing the norm $\|[\mathbf{T}]\| = \sqrt{[\mathbf{T}] \cdot [\mathbf{T}]}$.

The *trace* of a tensor $[\mathbf{T}]$, denoted by $tr[\mathbf{T}]$, is the scalar τ_{ii} , which means it is the sum of the main diagonal components of the associated matrix.

A symmetric tensor is said to be *positively defined (semidefined)* if $\mathbf{u} \cdot [\mathbf{T}]\mathbf{u} > 0$, ($\mathbf{u} \cdot [\mathbf{T}]\mathbf{u} \geq 0$), for $\forall \mathbf{u} \neq 0 \in E_3$ ($\forall \mathbf{u} \in E_3$), that is the attached quadratic form $\tau_{ij} u_i u_j$ is positively defined (semidefined).

By an *eigenvector* \mathbf{u} of the tensor $[\mathbf{T}]$ we understand any vector \mathbf{u} satisfying the equation $[\mathbf{T}]\mathbf{u} = \lambda \mathbf{u}$, the corresponding scalar λ being the associated *eigenvalue*. Within a certain basis the above equation comes to

$$(\tau_{ij} - \lambda \delta_{ij}) u_j = 0, i = 1, 2, 3$$

while the condition on the nontrivial solvability of this homogeneous system (i.e. the system yields also nontrivial solutions) is

$$P_n(\lambda) \equiv \det(\tau_{ij} - \lambda \delta_{ij}) = 0.$$

The polynomial $P_n(\lambda)$ is called the *characteristic polynomial* and the equation $P_n(\lambda) = 0$ which gives the eigenvalues, is known as the *characteristic equation*.

The coefficients I_1, I_2, I_3 of the characteristic polynomial are the *invariants* of $[\mathbf{T}]$ and they are given by

$$I_1 = tr[\mathbf{T}] = \tau_{ii}, 2I_2 = I_1^2 - tr[\mathbf{T}^2], I_3 = \det[\mathbf{T}].$$

Concerning the eigenvalues, they will be real if and only if the tensor $[\mathbf{T}]$ is symmetric and they will be positive if and only if the tensor $[\mathbf{T}]$ is positively defined.

If the roots of the characteristic equation are distinct, the corresponding eigenvectors will form an orthogonal basis. Such orthogonal eigenvectors could be determined even in the case of multiple roots.

The following decomposition theorems hold:

POLAR DECOMPOSITION THEOREM (Cauchy): *Any nonsingular tensor of second order $[\mathbf{T}]$ ($\det(\tau_{ij}) \neq 0$), can be written in the form $[\mathbf{T}] = [\mathbf{R}\mathbf{S}_d] = [\mathbf{S}\mathbf{R}]$, where $[\mathbf{R}]$*

is an orthogonal tensor while $[S_a]$ and $[S]$ are positively defined symmetric tensors with the same eigenvectors, this triplet of tensors being uniquely determined.

THEOREM. Any symmetric tensor $[T]$ of second order from E_3 can be uniquely decomposed as $[T] = (\frac{1}{3})[I] + [D]$, where $[D]$ is a symmetric tensor with the first invariant (trace) zero, which is also called the deviator (tensor), while $(\frac{1}{3})[I]$ is a spheric tensor (which means of the type $\alpha[I]$, α being a scalar).

B. On the analogy of the definitions from the classical field theory one could also define:

- the gradient of a vector $\mathbf{v}(v_i)$ as the second order tensor $[T] = grad \mathbf{v}$ whose components are $\tau_{ij} = v_{i,j}$;
- the gradient of a second order tensor $[T]$ of components (τ_{ij}) as the third order tensor $[S] = grad[T]$ whose components are $s_{ijk} = \tau_{ij,k}$;
- the divergence of a second order tensor $[T]$ as the vector \mathbf{a} which satisfies²

$$div[T] \cdot \mathbf{a} = div[T^T \mathbf{a}], \quad \forall \mathbf{a} \in E_3;$$

- the curl (rot) of a second order tensor $[T]$ is again a second order tensor, denoted by $rot[T]$, which is defined as

$$[rot[T]]\mathbf{a} = rot([T]\mathbf{a}), \quad \forall \mathbf{a} \in E_3;$$

- the Laplacian of a second order tensor $[T]$ is that second order tensor defined by

$$\Delta[T] = div(grad[T]).$$

The extension of the Green–Gauss (–Ostrogradski –Ampère) or the flux-divergence theorem also holds, i.e. we have

$$\int_D div[T]dv = \int_{\partial D} [T]nds,$$

obviously under the conditions of the differentiable tensor fields on D .

If \mathbf{i}_k form an orthonormal basis, which means $\mathbf{i}_j \cdot \mathbf{i}_k = \delta_{jk}$, then by accepting that the Cartesian systems are right-handed we will also have $\mathbf{i}_k \times \mathbf{i}_m = e_{kmn}\mathbf{i}_n$, where $e_{123} = e_{231} = e_{312} = 1$, $e_{132} = e_{321} = e_{213} = -1$, e_{kmn} being zero otherwise.

The relations $e_{ijk}e_{imn} = \delta_{jm}\delta_{kn} - \delta_{jn}\delta_{km}$ are still valid.

Then the following formulas have been stated without proof or derivation (they could be verified with the help of techniques developed so far):

- a) The triple vector product is

$$\begin{aligned} (\mathbf{u} \times \mathbf{v}) \times \mathbf{w} &= \mathbf{v}(\mathbf{u} \cdot \mathbf{w}) - \mathbf{u}(\mathbf{v} \cdot \mathbf{w}) = u_p v_q w_r e_{pqn} e_{nrs} \mathbf{i}_s \\ &= u_p v_q w_r (\delta_{pr} \delta_{qs} - \delta_{ps} \delta_{qr}) \mathbf{i}_s. \end{aligned}$$

- b) For any four arbitrary vectors \mathbf{a} , \mathbf{b} , \mathbf{c} and \mathbf{d} we have (Lagrange)

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c});$$

²If $[T]$ is a higher order tensor the result of applying the divergence will be a tensor of lower order with unity.

For two arbitrary vectors \mathbf{u} and \mathbf{v} we also have

c)

$$\operatorname{div}(\mathbf{u} \otimes \mathbf{v}) = (\operatorname{grad} \mathbf{u}) \mathbf{v} + (\operatorname{div} \mathbf{v}) \mathbf{u};$$

d)

$$\Delta \mathbf{v} \equiv \nabla^2 \mathbf{v} \equiv \operatorname{div}[\operatorname{grad} \mathbf{v}] = \operatorname{grad}(\operatorname{div} \mathbf{v}) - \operatorname{rot}(\operatorname{rot} \mathbf{v});$$

e)

$$(\operatorname{grad} \mathbf{u}) \mathbf{u} = \operatorname{grad} \left(\frac{1}{2} |\mathbf{u}|^2 \right) + \operatorname{rot} \mathbf{u} \times \mathbf{u};$$

f)

$$\operatorname{div}(\operatorname{grad} \mathbf{u})^T = \operatorname{grad}(\operatorname{div} \mathbf{u});$$

and

$$\operatorname{div}[(\operatorname{grad} \mathbf{u}) - (\operatorname{grad} \mathbf{u})^T] = -\operatorname{rot}(\operatorname{rot} \mathbf{u});$$

g)

$$\operatorname{rot} \mathbf{u} \times \mathbf{v} = [\operatorname{grad} \mathbf{u} - (\operatorname{grad} \mathbf{u})^T] \mathbf{v};$$

$$(\mathbf{u} \times \operatorname{grad}) \times \mathbf{v} = (\operatorname{grad} \mathbf{v})^T \mathbf{u} - \mathbf{u}(\operatorname{div} \mathbf{v}) = (\operatorname{grad} \mathbf{v}) \mathbf{u} - \operatorname{rot} \mathbf{v} \times \mathbf{u} - \mathbf{u}(\operatorname{div} \mathbf{v});$$

$$\begin{aligned} \operatorname{grad}(\mathbf{u} \cdot \mathbf{v}) &= (\operatorname{grad} \mathbf{u})^T \mathbf{v} + (\operatorname{grad} \mathbf{v})^T \mathbf{u} \\ &= (\operatorname{grad} \mathbf{u}) \mathbf{v} + (\operatorname{grad} \mathbf{v}) \mathbf{u} + \mathbf{u} \times \operatorname{rot} \mathbf{v} + \mathbf{v} \times \operatorname{rot} \mathbf{u}; \end{aligned}$$

$$\begin{aligned} \operatorname{rot}(\mathbf{u} \times \mathbf{v}) &= \operatorname{div}(\mathbf{u} \otimes \mathbf{v} - \mathbf{v} \otimes \mathbf{u}) \\ &= (\operatorname{grad} \mathbf{u}) \mathbf{v} - (\operatorname{grad} \mathbf{v}) \mathbf{u} + \mathbf{u}(\operatorname{div} \mathbf{v}) - \mathbf{v}(\operatorname{div} \mathbf{u}); \end{aligned}$$

$$\operatorname{div}(\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot \operatorname{rot} \mathbf{u} - \mathbf{u} \cdot \operatorname{rot} \mathbf{v};$$

h) For a vector \mathbf{u} and a tensor $[\mathbf{T}]$ we can write

$$[\mathbf{I}] \cdot \operatorname{grad} \mathbf{u} = \operatorname{div} \mathbf{u};$$

$$\operatorname{div}([\mathbf{T}]\mathbf{u}) = \operatorname{div}[\mathbf{T}^T] \cdot \mathbf{u} + [\mathbf{T}^T] \operatorname{grad} \mathbf{u};$$

i) Let $[\mathbf{W}]$ be a skew-symmetric tensor in E_3 . With this tensor, a vector \mathbf{w} can be associated, called also its dual, such that

$$\operatorname{div}[\mathbf{W}] = -\operatorname{rot} \mathbf{w};$$

$$\operatorname{div}[\mathbf{W}^T] = \operatorname{rot} \mathbf{w}.$$

If $[\mathbf{W}] = [\boldsymbol{\Omega}]$, which is the rotation tensor, then

$$2\mathbf{w} = \boldsymbol{\omega} \quad \text{and} \quad 2\operatorname{div}[\boldsymbol{\Omega}] = -\operatorname{rot} \boldsymbol{\omega}.$$

For an arbitrary vector \mathbf{v} and its dual \mathbf{w} we also have

$$[\mathbf{W}]\mathbf{v} = \mathbf{W} \times \mathbf{v} \quad \text{and} \quad [\mathbf{W}] \cdot \operatorname{grad} \mathbf{v} = \mathbf{w} \cdot \operatorname{rot} \mathbf{v}.$$

References

- [1] Abgrall R., "On Essentially Non-Oscillatory Schemes on Unstructured Meshes: Analysis and Implementation", *J. Comp. Phys.*, 114, pp. 45-58, 1994
- [2] Abraham R., Marsden J. E., "Foundations of Mechanics", 2-nd edition, The Benjamin/Cummings Publishing Company, 1978
- [3] Amon C. H., "Spectral Element-Fourier Method for Transitional Flows in Complex Geometries", *AIAA Journal*, 31,1, pp. 42-48, 1993
- [4] Anderson D. A., Tannehill J. C., Fletcher R. H., "Computational Fluid Mechanics and Heat Transfer", Hemisphere Publishing Corporation, 1984
- [5] Anderson J.D. Jr., "A Time-Dependent Analysis for Vibrational and Chemical Nonequilibrium Nozzle Flows", *AIAA Journal*, 8, 3, pp. 545-550, 1970
- [6] Azaiez M., Fikri A., Labrosse G., "A unique grid spectral solver of the nD Cartesian unsteady Stokes system. Illustrative numerical results", *Finite Elements in Analysis and Design*, 16, pp. 247-260, 1994
- [7] Batoul A., Khallouf H., Labrosse G., "Une methode de resolution directe (pseudospectrale) du probleme de Stokes 2D/3D instationnaire", *C. R. Acad. Sci. Paris, serie II, Mec., Physique, Chimie, Astronomie*, Tome 319, 12, 1994
- [8] Bers L., "Mathematical Aspects of Subsonic and Transonic Gas Dynamics", J. Wiley, N.Y., 1958
- [9] Bers L., John Fr., Schechter M., "Partial Differential Equations" *Am. Mat. Soc.*, 1964
- [10] Birkoff G., "Hydrodynamics", Princeton Univ. Press, 1960
- [11] Brebbia C.A., Telles J.C.F., Wrobel L.C., "Boundary Element Techniques", Springer-Verlag, 1984
- [12] Brown D.L., Cortez R., Minion M.L., "Accurate Projection Methods for the Incompressible Navier–Stokes Equations", *J. Comp. Phys.*, 168, pp. 464-499, 2001

- [13] Canuto C., Hussaini Y. M., Quarteroni A., Zang Th. A., "Spectral Methods in Fluid Dynamics", Springer-Verlag, 1988
- [14] Canuto C., "Spectral Methods for Viscous, Incompressible Flows", in Hussaini Y. M., Kumar A., Salas M. D., eds, "Algorithmic Trends in Computational Fluid Dynamics", Springer-Verlag, pp. 169-193, 1993
- [15] Carabinean A., Bena D., "Conformal Mappings for Neighboring Domains" (in Romanian), Ed. Acad. Rom., Bucuresti, 1993
- [16] Cardos V., "General Integral Method in the Study of Supersonic Flows Past Slender Bodies", Rev. Roum. Math. Pure Appl., 37, p. 57, 1992
- [17] Cesari L., "Functional Analysis and Galerkin's Method", Mich. Math. J., 11, 3, pp. 383-414, 1964
- [18] Chattot J-J., "Computational Aerodynamics and Fluid Dynamics", Springer-Verlag, 2002
- [19] Chorin A.J., Marsden J.E., "A Mathematical Introduction to Fluid Mechanics", Springer-Verlag, 1979
- [20] Chorin A. J., "Vortex Sheets Approximation of Boundary Layers", J. Comp. Phys., 1978
- [21] Chorin A. J., Hughes T. J. R., Marsden J. E., Comm. Pure. Appl. Math. 31, p. 205, 1977
- [22] Chow, C-Y., "An Introduction to Computational Fluid Mechanics", John Wiley & Sons, 1979
- [23] Chu P. C., Fan C., "A Three-Point Combined Compact Difference Scheme", J. Comp. Phys., 140, pp. 370-399, 1998
- [24] Cocarlan P., "Integral Equations and Functional Analysis Elements", in "Classical and Modern Mathematics" (in Romanian), vol. II, (C. Iacob ed.), Ed. Tehnica, Bucuresti, 1979
- [25] Dennis S.C.R., Chang G.-Z., "Numerical Solution for Steady Flow Past a Circular Cylinder at Reynolds Numbers up to 100", J. Fluid Mech. 42, pp. 471-489, 1970
- [26] Dennis S.C.R., Dunwoody J., "The Steady Flow of Viscous Fluid Past a Flat Plate", J. Fluid. Mech. 24, pp. 577-595, 1986
- [27] Dennis S.C.R., Kocabijik S., "The Solution of Two Dimensional Oseen Flow Problem Using Integral Conditions", IMA J. Appl. Math. 45, p.1-3, 1990
- [28] Dennis S.C.R., Walker J.D.A., "Calculation of Steady Flow Past a Sphere at Low and Moderate Reynolds Numbers", J. Fluid Mech. 48, pp. 771-890, 1971
- [29] Dexun F., Yanwen M., "Analysis of Super Compact Finite Difference Method and Application to Simulation of Vortex-Shock Interaction", Int. J. Numer. Meth. Fluids, 36, pp. 773-805, 2001

- [30] Dinca G., "Variational Methods and Applications" (in Romanian), Ed. Tehnica, Bucuresti, 1980
- [31] Dinu L., "Shock Waves Theory in Plasma" (in Romanian), Ed. Stiint. si Encicl., Bucuresti, 1976
- [32] Dragos L., "Tensorial Calculus" in "Classical and Modern Mathematics" (in Romanian), vol II, (C. Iacob ed.), Ed. Tehnica, Bucuresti, 1979
- [33] Dragos L., "The Principles of Mechanics of Continua" (in Romanian), Ed. Tehnica, Bucuresti, 1983
- [34] Dragos L., "Fluid Mechanics I" (in Romanian), Ed Acad. Rom, Bucuresti, 1999
- [35] Dragos L., "Mathematical Methods in Aerohydrodynamics" (in Romanian), Ed. Acad. Rom, Bucuresti, 2000
- [36] Ducaru A., "On the Compressible Subsonic Flows Past Some Profiles Obtained by Conformal Mappings of a Particular Type" (in Romanian), Studii si Cercetari Matematice, 18, 1966
- [37] Dubois T., Jauberteau F., Temam R., "Solution of the Incompressible Navier–Stokes Equations by the Nonlinear Galerkin Method", J. Sci. Comp., 8, 2, pp. 167-194, 1993
- [38] Dushane T. E., Math. Comp., 27, p. 719, 1973
- [39] Ericksen J.L., "Tensor Fields", in Handbuch der Phys. III/1, Springer-Verlag, 1960
- [40] Ericksen I.L., Rivlin, R. S., J. Rational Mech. Anal., 4, p. 323, 1955
- [41] Eringen A.C., "Mechanics of Continua", J. Wiley & Sons, 1967
- [42] Euvrard D., "Résolution numérique des équations aux dérivées partielles", Ed. Masson, 1994
- [43] Feistauer M., "Mathematical Methods in Fluid Dynamics", Longman, 1992
- [44] Fife P. C., Arch. Rat. Mech. 28, p. 184, 1968
- [45] Finn R., "On Steady-State Solutions of Navier–Stokes Partial Differential Equations", Arch. Rat. Mech. Anal., 3, pp.381-396, 1959
- [46] Finn R., Gilbary D., Comm Pure Appl. Math., 10, p. 23, 1957
- [47] Foa E., "Sull'impiega dell' analisi dimensionale nella studio del moto turbolente", L'Industria (Milan), 43, p.426, 1929
- [48] Galdi G.P., Nečas J., "Recent Developments in Theoretical Fluid Mechanics", Longman, 1993
- [49] Galdi G.P., "Existence and Uniqueness at Low Reynolds Number of Stationary Plane Flow of a Viscous Fluid in Exterior Domains", Pitman Research Notes in Mathematics Series, vol. 291, p. 1-33, 1993

- [50] Galdi G.P., Simander C.G., “Existence, Uniqueness and L_q -Estimates for the Stokes Problem in an Exterior Domain”, Arch. Rat. Mech. Anal. 112, p.291-318, 1990
- [51] Georgescu A., “Stability Theory in Hydrodynamics” (in Romanian), Ed. St. si Encicl., Bucuresti, 1976
- [52] Germain P., “Mécanique des milieux continus”, Ed. Masson, 1962
- [53] Gheorghita St. I., “Mathematical Methods in Underground Hydro-Gasdynamics” (in Romanian), Ed. Acad. RSR, Bucuresti, 1966
- [54] Gheorghita St.I., Homentcovschi, “Integral Transforms”, in “Classical and Modern Mathematics” (in Romanian), vol. III, (C. Iacob ed.), Ed. Tehnica, Bucuresti, 1981
- [55] Gheorghiu C. I., “A Constructive Introduction to Finite Element Method”, Quo Vadis, Cluj– Napoca, Romania, 1999
- [56] Goldstein S. L., “Lectures in fluid mechanics”, Cambridge Univ. Press, 1959
- [57] Graffi D., “Il teorema di unicità nella dinamica dei fluidi compressibili”, J. Rat. Mech. Anal., 2, 1953
- [58] Graffi D., “Sur teorema di unicità nella dinamica dei fluidi”, Annali di Matematica, 50, p.379, 1960
- [59] Graffi D., “Sul teorema di unicità per equazioni del moto dei fluidi compressibili in un dominio illimitato”, Atti della Accad. delle Sci. dell’Institut. di Bologna 7, pp. 1-8, 1969
- [60] Griebel M., Dornseifer Th., “Numerical Simulation in Fluid Dynamics”, SIAM, Philadelphia, 1997
- [61] Hald O. H., Manceri del Prete V., Math. Comp., 32, 1978
- [62] Harten A., Enquist B., Osher S., Chakravarthy S., “Uniformly high order essentially non-oscillatory schemes, III”, J. Comp. Phys., 71, pp. 231-303, 1987, republished in J. Comp. Phys., 131, 1997
- [63] Heinrichs W., “A Spectral Multigrid Method for the Stokes Problem in Stream Function Formulation”, J. Comp. Phys. 102, pp.310-318, 1992
- [64] Holt M., “Numerical Methods in Fluid Dynamics”, Springer-Verlag, 1977
- [65] Homentcovski D., Cocora D., “Some Developments of the CVBEM. Application of the Mixed Boundary Value Problem for the Laplace Equation”, Incest – Bucharest, Preprint ser. Math., 13, 1981
- [66] Hopf E., Math. Nachr., 4, pp. 213-231, 1950
- [67] Hopf E., J. Rat. Mech. Anal., 1, p.107, 1952
- [68] Hromadka T.V. II, “The Complex Variable Boundary Element Method”, Springer-Verlag, 1984

- [69] Iacob C., “Introduction mathématique à la mécanique des fluides”, Ed. Acad. RPR – Gauthier Villars, 1959
- [70] Iacob C., “Détermination de la seconde approximation de l’écoulement compressible subsonique autour d’un profil donné”, *Archivum Mechaniki Stosowanej*, 2, 16, 1964
- [71] Iacob C., “Complex Functions”, in “Classical and Modern Mathematics” (in Romanian), vol.II, (C. Iacob ed.), Ed. Tehnica Bucuresti, 1979
- [72] Jiang, B.N., “The Least-Squares Finite Element Method”, Springer-Verlag, 1998
- [73] Kato T., “On Classical Solutions of the Two-Dimensional Non-Stationary Euler Equations”, *Arch. Rat. Mech. Anal*, 25, pp.188-200, 1967
- [74] Kocin N.E., Kibel I.A., Rose N.V., “Theoretical Hydrodynamics” (in Romanian, translated from Russian), Ed. Tehnica, Bucuresti, 1951
- [75] Komatu Y., “Existence Theorem of Conformal Mapping of Double Connected Domains”, *Kodai Math. Sem. Rep.*, 5-6, p. 83, 1979
- [76] Ladyzhenskaya O. A., “The Mathematical Theory of Viscous Incompressible Flows”, Gordon and Breach, 1963
- [77] Ladyzhenskaya O. A., “Annual Review of Fluid Mechanics”, 7, pp. 249-272, 1975
- [78] Lamb H., “Hydrodynamics”, Cambridge University Press, 1932
- [79] Lander, J., “MT405-Numerical Methods II”, Lecture Notes, INTERNET Course, 1997
- [80] Lavrentiev M. A., Sabat B. V., “Methods of the Complex Variable Functions Theory” (in Russian), Izd. Nauka, Moskva, 1973
- [81] Lele S. K., “Compact Finite Difference Schemes with Spectral-like Resolution”, *J. Comp. Phys*, 103, pp. 16-42, 1992
- [82] Leray J., *J. de Math.*, 12, pp. 1-82, 1933
- [83] Li R., Chen Z., Wu, W., “Generalized Difference Methods for Differential Equations”, Marcel Dekker Inc., 2000
- [84] Lichtenstein L., “Gründlagen der Hydrodynamik”, Berlin, 1929
- [85] Lions P.-L., “Mathematical Topics in Fluid Mechanics”, vol.I (Incompressible Models), Oxford Scientific Publications, 1996
- [86] Liubimov A.N., Rusanov V.V., “Gas Flow Past Blunt Bodies”, NASA TT F-714, 1973.
- [87] Loitsyanskij L.G., “Mechanics of Liquids and Gases”, Oxford Univ. Press, 1996
- [88] Marcov N., “Finite Element Method” in “Classical and Modern Mathematics” (in Romanian), vol. IV, (Caius Iacob ed.), Ed. Tehnica, Bucuresti, 1979

- [89] Marcov N. "The Electroconductor Viscous Fluid Flow Past Slender Bodies" (in Romanian), *St. Cerc. Mat.*, 20, p. 199; 21, p.1063, 1968
- [90] Micula G., "Spline Functions" (in Romanian), Ed. Tehnica, Bucuresti, 1978
- [91] Milne-Thomson L.M., "Theoretical hydrodynamics", Mac Millan, 1949
- [92] Murman E. M., Cole J. D., "Calculation of Plane Steady Transonic Flows", *AIAA Journal*, 9,1, pp. 115-121, 1971
- [93] Muskelisvili I.N., "Singular Integral Equations" (in Russian), *Izd. Fizmatgiz, Moskva*, 1962
- [94] Napolitano M., "Efficient Solution of Two Dimensional Steady Separated Flows", *Computers and Fluids*, 20, pp. 213-222, 1991
- [95] Noll W., "A Mathematical Theory of the Mechanical Behavior of Continuous Media", *Arch. Rat. Mech. Anal.*, 4, pp. 323-333, 1955
- [96] Olejnik O. A., *Am. Math. Soc. Translations, Ser. 2*, 33, p. 285, 1965
- [97] Oroveanu T., "Viscous Fluids Mechanics" (in Romanian), Ed. Acad. RSR, Bucuresti, 1967
- [98] Oseen C.W., "Neue Methoden und Ergebnisse in der Hydrodynamik", Leipzig, 1927
- [99] Patankar S.V., "Numerical Heat Transfer and Fluid Flow", McGraw-Hill, 1980
- [100] Petrila, T., Gheorghiu, C. I., "Finite Element Methods and Applications" (in Romanian), Ed. Acad. Rom., Bucuresti, 1987
- [101] Petrila T., "Une nouvelle méthode pour l'étude de l'influence des parois sur l'écoulement fluide plan", *Riv. Mat. Univ. Di Parma*, 3, 2, pp. 47-51, 1973
- [102] Petrila T., "Méthode pratique pour la détermination de l'écoulement produit par un déplacement d'un profile d'aile au point anguleux dans un fluide idéal", *Mathematica*, t. 19 (42), 2, pp. 195 – 201, 1977
- [103] Petrila T., "Une nouvelle méthode pour l'étude de l'écoulement d'un système de n profiles dans un fluide idéal", *Rév. Roum. Math. Pures et Appl.* 10, 1979
- [104] Petrila T., "Une nouvelle méthode pour le calcul de l'influence des parois sur l'écoulement fluide plan", *Rév. Roum. Math. Pures et Appl.* 25, 1, pp. 99–110, 1980
- [105] Petrila T., "An Improved CVBEM for Plane Hydrodynamics", *Rév. de l'Analyse numérique et de la Théorie de l'approximation*, vol. XVI, fasc. 2, pp. 149-157, 1987
- [106] Petrila T., "A Complex Variable Boundary Element Method and its Use to the Theory of Profiles", *Proceedings of the 4-th International Conference on Computational Fluid Dynamics*, Univ. of California, Davis, pp. 997-1003, (&K. Roesner), 1991

- [107] Petrila T., "CVBEM for the Fluid Flow Determined by the Motion of a Dirigible Balloon in a Wind Stream", *Rév. de l'Analyse numérique et de la Théorie de l'Approximation*, t.27, 1, pp. 155-165, 1998
- [108] Petrila T., "The Uniqueness of the Classical Solutions of the Navier–Stokes System for an Incompressible Nonstationary Flow", *Studia Universitatis Babeş-Bolyai, ser. Mathematica*, XXVI, 3, pp. 3-5, 1981
- [109] Petrila T., "Mouvement général d'un profil dans un fluide idéal en présence d'une paroi perméable illimitée. Cadre variationnel et approximation par une méthode d'éléments finis", *Rev. de l'Analyse Numérique et de la Théorie de l'Approximation*, t.8, 1, pp. 67-77, 1979
- [110] Petrila T., "Lectures on Mechanics of Continua" (in Romanian), lito, Universitatea Babeş-Bolyai, Cluj-Napoca, 1980
- [111] Petrila T., "Mathematical Models in Plane Hydrodynamics" (in Romanian), Ed. Acad. RSR, Bucuresti, 1980
- [112] Petrila T., "On Certain Mathematical Problems Connected with the Use of the Complex Variable Boundary Element Method to the Problems of the Plane Hydrodynamics. Gauss Variant of the Procedure", *The Math. Heritage of C.F. Gauss* (G. Rassias ed.), World Scientific Publ. Co. Singapore, pp.585-604, 1991
- [113] Petrila T., "Sur l'influence du sol sur l'écoulement plan autour d'une aile mince à jet", *Mathematica*, 20(43), 1, pp. 195-301, 1978
- [114] Petrila T., Maksay S.I., "Dynamic Boundary Layer with Sliding on a Flat Plate", *Proc. Int. Conf. on Appl. Math.*, Pitesti, Romania (A. Georgescu ed.), 2002
- [115] Petrila T., "A New Attempt to Solve the Equations of Ideal Compressible Flows", Preprint nr. 6 of the University of Cluj, Intinerant Seminar on Functional Equations, pp. 127-136, 1981
- [116] Petrila T., Trif D., "Numerical Alternative Method Scheme for Burgers Equation", *Rev. de l'Anal. Numérique et de la Theorie de l'Approx.*, 22, 1, pp. 87-96, 1993
- [117] Petrila T., Trif D., Bosilca P., Labrosse G., "An Implicit Time Solver for the Spatial High Frequencies of the Pseudospectral Advection-Diffusion Systems", *Proc. 5th Int. Symp. on CFD*, Sendai, vol. II, pp. 429-433, 1993
- [118] Petrila T., Trif D., "A Spectral Matrix-Free Method for the Nonstationary Viscous Incompressible Flows", *Proc. First Asian CFD Conf.*, Hong-Kong, vol.2, pp. 435-440, 1995
- [119] Petrila T., Trif D., "An Improved Viscous Step for a Navier-Stokes Algorithm in Complex Geometries", *Rev. Roum. Math. Pures et Appl.*, 42, 3-4, pp. 311-318, 1997
- [120] Peyret R., Taylor T. D., "Computational Methods for Fluid Flow", Springer-Verlag, 1983

- [121] Peyret R. (ed.), "Handbook of Computational Fluid Mechanics", Academic Press, 1996
- [122] Polotca O., Petrila T., Cardos V., "On the Flow Induced by an Arbitrary Motion of an Airfoil in an Uniform Stream" (unpublished)
- [123] Popp S., "Lectures on Gas Dynamics" (in Romanian), lito, Univ. Bucuresti, 1979
- [124] Pozrikidis, C., "Numerical Computation in Science and Engineering", Oxford Univ. Press, 1998
- [125] Pozrikidis, C., "Introduction to Theoretical and Computational Fluid Dynamics", Oxford Univ. Press, 1997
- [126] Quartapelle L., "Numerical Solution of the Incompressible Navier-Stokes Equations", Birkhäuser, 1996
- [127] Quartapelle L., Valz-Gris F., "Projection Conditions on the Vorticity in Viscous Incompressible Flows", *Int. J. Numer. Meth. Fluids* 1, pp.129-144, 1981
- [128] Quarteroni, A., "Numerical Approximation of Partial Differential Equations", Springer-Verlag, 1994
- [129] Rionero S., Galdi P. "On the Uniqueness of Viscous Fluid Motion", *Arch. Rat. Mech. Anal.* 63, p. 295, 1976
- [130] Rionero S., Maiellaro M., *Rend. Circ. Matem. Di Palermo, S II*, 27, 2, p. 305, 1978
- [131] Roache P. J., "Computational Fluid Dynamics", Hermosa Publishers, 1972
- [132] Savulescu, St. N., Dumitrescu H., Georgescu A., Bucur M., "Mathematical Researches in Modern Boundary Layer Theory" (in Romanian), Ed. Acad. R.S.R., Bucuresti, 1981
- [133] Scheiber E., Lupu M., "Special Mathematics (Problems Solving Assisted by Computer)" (in Romanian), Ed. Tehnica, Bucuresti, 1998
- [134] Sedov, L. I., "Ploskaia zadatcha gydrodynamiki e aerodynamiki (Plane problems of hydrodynamics and aerodynamics)", *Izd. Nauka, Moskva*, 1950
- [135] Serrin, J., "On the Uniqueness of Compressible Fluid Motions", *Arch. Rat. Mech. Anal.* 32, p. 271, 1959
- [136] Shinbrat M., "Lectures on Fluid Mecanics pIII", Gordon and Brench, 1973
- [137] Shu C.W., "High Order ENO and WENO Schemes for Computational Fluid Dynamics", in T.J. Barth, H. Deconinck eds., "High Order Methods for Computational Physics", Springer-Verlag, 1999.
- [138] Sonar T., "On the Construction of Essentially Non-Oscillatory Finite Volume Approximations to Hyperbolic Conservation Laws on General Triangulations: Polynomial Recovery, Accuracy and Stencil Selection", *Computer Methods in Appl. Mech. and Eng.*, 140, pp. 157-181, 1997.

- [139] Taposu I., “Dolphin Profiles” (in Romanian), Ed. Tehnica, Bucuresti, 2002
- [140] Teman R., “Theory and Numerical Analysis of Navier–Stokes Equations”, North Holland Publ., 1977
- [141] Teman R., “Navier–Stokes Equations” North Holland Publ., 1979
- [142] Thomassaut F., “Implementation of Finite Element Methods for Navier–Stokes Equations”, Springer-Verlag, 1981
- [143] Thompson J.F., Thames F.C., Mastin C.W., “Automatic Numerical Generation of Body-Fitted Curvilinear Coordinate System for Field Containing Any Number of Arbitrary Two-Dimensional Bodies”, *J. Comp. Phys.*, 15, pp. 299-319, 1974
- [144] Trefethen L. N., “Spectral Methods in MATLAB”, SIAM, 2000.
- [145] Trif, D., “Numerical Methods for Differential Equations and Dynamical Systems” (in Romanian), Transilvania Press, Cluj, Romania, 1997
- [146] Trif D., “An Almost Explicit Scheme for a Certain Class of Nonlinear Evolution Equations”, *Studia Univ. Babeş-Bolyai, ser. Mathematica*, 38, 2, pp. 103-112, 1993
- [147] Trif D., Petrila T., “An Almost Explicit Algorithm for the Incompressible Navier-Stokes Equations”, *Pure Mathematics Appl.*, 6, 2, pp. 279-285, 1995
- [148] Trif D., Petrila T., “An Analytical-Numerical Algorithm for the Incompressible Navier-Stokes Equations in Complex Domains”, in “Integral Methods in Science and Engineering”, Volume two, 206-209, C. Constanda, J. Saranen, S. Seikkala (eds.), Longman, 1997
- [149] Truesdell C., Noll W., “The non linear field theory of mechanics”, *Handbuch der Phys.*, III/3, Springer-Verlag, 1965
- [150] Truesdell C., “First Course in Rational Continuum Mechanics”, The Johns Hopkins Univ. Press, 1972
- [151] Truesdell C., Toupin R., “The Classical Fields Theories”, in *Handbuch der Phys.* III/I, Springer-Verlag, 1960
- [152] Vladimirov V. S., “Mathematical Physics Equations” (in Russian), Izd. Nauka, Moskva, 1976
- [153] Warsi Z. U. A., “Fluid Dynamics. Theoretical and Computational Approaches”, C.R.C. Press, 1999
- [154] Watson, G. N., “A Treatise on the Theory of Bessel Functions”, Cambridge, 1922.
- [155] Wendt J. F. (ed.), “Computational Fluid Dynamics, An Introduction”, Springer-Verlag, 1992
- [156] Weyl H., “Shock Waves in Arbitrary Fluids”, *Comm. on Pure and Appl. Math.*, 2, p. 103, 1949

[157] Wu J.C., *Int. J. Numer. Methods Fluids*, 4, p. 185, 1984

[158] Wu J.C., *Comput. Fluids*, 1, p. 197, 1973

[159] Zeytounian R.K., “*Mécanique des fluides fondamentales*”, Springer-Verlag, p. 85, 1991

[160] Zeytounian R.K., “*Mécanique des fluides fondamentales*”, vol. I, II, III, lito, Laboratoire de Mécanique de Lille, 1989, 1990

[161] Zierep J., “*Theoretische Gasdynamik*”, Springer-Verlag, 1976

Index

- acceptable grid, 388
- Avogadro, 1
- Busemann A., 128
- Butcher, 212
- characteristics, 47, 123
 - hodograph, 131
 - lines, 227
- circulation, 10
- coefficient
 - drag, 216
 - pressure, 63, 90
 - viscosity
 - first, 39
 - second, 39
- complex
 - potential, 60
 - velocity, 61
- condition
 - adherence, 42
 - boundary
 - essential, 381
 - natural, 381
 - CFL, 259
 - entropy, 48
 - integral type for vorticity, 144
 - Joukowski–Kutta, 77
 - Lipschitz, 208
 - slip, 42, 120
 - smoothness, 6
 - Zorawski, 16
- configuration, 3
- conformal mapping, 65
- consistence, 199
- continuity axiom, 6
- continuum, 1
 - deformable, 4
 - particle, 3
- coordinates
 - Eulerian, 7
 - Lagrangian, 5
 - material, 5
 - spatial, 7
- critical sound speed, 112
- Crocco–Vazsonyi theorem, 54
- curl, 9
- Dario Graffi, 58
- density, 3
- derivative
 - Chebyshev–Galerkin, 454
 - Chebyshev-collocation, 454
 - Fourier–Galerkin, 446
 - Fourier-collocation, 446
 - Legendre–Galerkin, 451
 - Legendre-collocation, 451
 - local, 7
 - material, 7
 - substantive, 7
 - total, 7
- description
 - material, 5
 - spatial, 5, 7
- discontinuity
 - contact, 47
 - strong, 43
- dissipation, 142
- drag, 72
- dual grid, 414
- energy
 - deformation, 25
 - internal, 27
- ENO, 398
- enthalpy, 32
- equation
 - biharmonic, 313
 - Burgers, 288
 - Cauchy–Riemann, 60

- continuity, 18
- Crocco–Vazsonyi, 54
- elliptic, 227
- energy, 24
- Gibbs, 30, 32, 35
- hyperbolic, 227
- integral on the boundary, 380
- Molenbroek–Chaplygin, 131
- parabolic, 227
- Prandtl, 131
- Steichen, 114
- Stokes, 312
- equations
 - balance, 20
 - Cauchy’s motion, 23
 - constitutive, 27
 - Euler, 34
 - Gibbs, 31
 - Navier–Stokes, 134
 - state, 35
- Eucken, 37
- Euler–Lagrange criterion, 13
- Eulerian coordinates, 5
- Fehlberg, 213
- finite differences
 - backward, 248, 254
 - centered, 248, 254
 - forward, 248, 254
 - one-sided, 256, 324
- flow
 - almost (slightly) potential, 92
 - homotropic, 37
 - isentropic, 37
 - simple waves, 123
- fluid
 - barotropic, 38
 - inviscid, 34
 - Newtonian, 39
 - non-Newtonian, 40
 - real, 38
 - Reiner–Rivlin, 39
 - Stokes, 38
 - viscous, 39
- form
 - differential, 44
 - integral, 44
 - variational, 357
 - weak, 44
- formulation
 - inverse, 383, 384
 - original, 384
 - weak, 383, 384
- Gibbs phenomenon, 444
- heat
 - conduction, 28
 - radiation, 28
 - specific, 31
 - total, 32
- hypotheses
 - Weyl, 48
- hypothesis
 - Joukovski, 74
 - Stokes, 39
- identity, Somigliana, 381
- inequality, Clausius–Duhem, 30
- integral representation, 380
- Lagrangian coordinates, 5
- law
 - Cauchy
 - first, 23
 - second, 22
 - conservation, 44
- laws
 - behaviour, 33
 - constitutive, 26, 33
- lemma, Cauchy, 21
- lift, 72
- lowerside, 121
- Mach angle, 116
- mass, 1
 - specific, 3
- material volume, 4
- materialize, 63
- medium
 - homogeneous, 20, 33
 - incompressible, 19
 - izotropic, 34
- method
 - Adams–Bashforth, 216
 - Adams–Moulton, 216
 - ADI, 302
 - characteristics, 293
 - collocation on points, 383
 - collocation on subdomains, 383
 - Crank–Nicolson, 216, 277
 - finite volume, 397
 - fractional step, 239
 - generalized difference, 398
 - Liapunov–Schmidt, 462
 - Liebmann, 297
 - MAC, 316
 - MacCormack, 269
 - multi-step, 216
 - Runge–Kutta, 212
 - S.O.R., 298
 - single-step, 215
 - Uzawa, 474
 - von Neumann, 258
- Morawetz C., 128

- motion
 - permanent, 8
 - plane, 12
 - spectrum, 9
 - steady, 8
- Noll, 33
- number
 - Courant, 259, 280
 - Euler, 160
 - Froude, 160
 - Grashof, 162
 - Knudsen, 42
 - Mach, 111
 - Pecllet, 161, 200
 - Prandtl, 162, 223
 - Reynolds, 139, 160, 312
 - Schmidt, 162
 - similarity, 160
 - Strouhal, 160
- numerical
 - diffusion, 266
 - dispersion, 265
- orthogonal projection operator, 138
- paradox
 - D'Alembert, 72
 - Stokes, 165
- parameters
 - state, 26
 - thermodynamic, 26
- perfect gas, 36, 43
- point
 - cuspidal, 74
 - stagnation, 60
- Prandtl–Meyer flow, 117
- principle
 - Cauchy, 17
 - dependence on the history, 33
 - Fourier–Stokes, 29
 - heat flux, 29
 - indestructibility of matter, 6
 - mass conservation, 18
 - material frame indifference, 33
 - objectivity, 33
 - spatial localization, 33
 - thermodynamics, second, 29
 - variation
 - energy, 24
 - momentum torsor, 20
- problem
 - boundary values, 247
 - Cauchy, 208
 - well-posed, 231, 365
- process
 - adiabatic, 28, 37, 43
 - irreversible, 27
 - reversible, 27
 - thermodynamic, 27
- profile
 - dolphin, 76
 - Joukovski, 75
 - Karman–Trefftz, 76
 - von Mises, 76
- relations
 - Rankine–Hugoniot, 46, 50
 - stresses-deformation, 33
- Reynolds (transport) theorem, 13
- residue, 199
- rotation, 9
- scheme
 - Lax, 277
 - Lax–Wendroff, 279
- sharp trailing edge, 73
- shock
 - back, 47
 - compressive, 48
 - condition, 234
 - front, 47
 - rarefaction, 48
 - wave, 43, 47
- similarity, 159
- simple wave, 117
- solidify, 63
- solution
 - weak, 43
- sonic lines, 114
- spatial instability, 200
- spectral accuracy, 443
- stability, 199
- stream
 - filament, 9
 - function, 11, 60
 - lines, 8
 - surfaces, 8
- stress vector, 21
- stresses, 17
- successive iterations, 209
- Sutherland, 40
- system
 - Chaplygin, 131
 - elliptic, 228
 - hyperbolic, 228
 - Oseen, 166
 - parabolic, 228
 - reduced, 163
 - thermodynamic, 27
- tensor
 - rate-of-strain, 38, 141
 - rotation, 141
 - stress, 21

- theorem
 - Betti, 381
 - Cauchy, 22
 - Cauchy–Eriksen–Rivlin, 39
 - Euler, 13, 15
 - Green, 381
 - Helmholtz, 10
 - Kutta–Jukovski, 72
 - Lagrange, 53
 - Radon–Nycodim, 2
 - Reynolds, 13, 15
 - Riemann–Caratheodory, 65
 - the first Bernoulli, 53
 - the second Bernoulli, 53
 - Thompson (Lord Kelvin), 52
 - transport, 13, 15
- thermodynamic
 - equilibrium, 27
 - state, 27
- trace, 382
- trajectories, 8
- transform
 - discrete, 442
 - discrete polynomial, 449
 - finite, 441
- transformation
 - Joukovski, 75, 77
- triangulation, 366
- truncation, 199
- tube
 - rotation, 10
 - stream, 9
- upperside, 121
- variables
 - state, 26
 - thermodynamic, 26
- velocity potential, 59
- volume support, 3
- vortex
 - lines, 9
 - surfaces, 9
- vorticity, 9
- wing profile, 70