SëMA

BOLETÍN NÚMERO 36

Septiembre 2006

sumario

Boletín de la Sociedad Española de Matemática Aplicada SëMA

Página web de SëMA
`http://www.sema.org.es/`

Dirección Editorial: Boletín de SëMA. Dpto. de Matemática Aplicada. Universidad de Salamanca. Plaza de la Merced, s/n. 37008. Salamanca. `boletin_sema@usal.es`.

ISSN 1575-9822.
Depósito Legal: AS-1442-2002.

Imprime: Gráficas Lope. C/ Laguna Grande, parc. 79, Políg. El Montalvo II 37008. Salamanca.

Diseño de portada: Luis Ferragut Alonso.

Estimados compañeros:

Regresamos tras el descanso estival con un nueva edición de nuestro boletín trimestral. En este número encontraréis el primer capítulo de una serie de tres en que hemos dividido para su publicación un interesante trabajo de Roland Glowinski y sus colaboradores sobre el Método de Elementos Finitos para las ecuaciones de Navier-Stokes. El siguiente artículo, también sobre las ecuaciones de Navier-Stokes, nos lo envia Didier Bresch desde la Universidad Joseph Fourier en Grenoble. Incluimos otros dos artículos de temática variada, el primero sobre autómatas celulares de Ángel Martín del Rey, de la Universidad de Salamanca, y el segundo sobre dinámica de poblaciones de Ángel Calsina y Silvia Cuadrado, de la Universidad Autónoma de Barcelona.

En la sección **Historia de las Matemáticas**, Antonio Cañada, de la Universidad de Granada, nos acerca a Fourier y sus coeficientes.

Nuestro agradecimiento a todos los que han colaborado en esta nueva edición del boletín. Os animamos a todos a participar con artículos de carácter científico o de opinión, o en cualquiera de las otras secciones que lo componen.

Grupo Editor
`boletin_sema@usal.es`

Bol. Soc. Esp. Mat. Apl. nº 36(2006), 7–62

Finite Element Methods for the Numerical Simulation of Incompressible Viscous Fluid Flow Modeled by the Navier-Stokes Equations. Part I

Roland Glowinski*, Tsorng-Whay Pan*,
L. Héctor Juárez V.+ and Edward Dean*

*Department of Mathematics, University of Houston, Houston, Texas 77204-3008, USA
+Departamento de Matematicas, Universidad Autonoma Metropolitana-Iztapalapa, Iztaplapa, D. F. 09340, MEXICO

Introduction.

The Navier–Stokes equations have been known for more than a century and they still provide the most commonly used mathematical model to describe and study the motion of viscous fluids, including phenomena as complicated as turbulent flow. One can only marvel at the fact that these equations accurately describe phenomena whose length scales (resp., time scale) range from fractions of a millimeter (resp., of a second) to thousands of kilometers (resp., several years). Indeed, the Navier–Stokes equations have been validated by numerous comparisons between analytical or computational results and experimental measurements; some of these comparisons are reported in Canuto et al. 1988 [1], Lesieur 1990 [2], Guyon et al. 1991 [3], and Glowinski 2003 [4].

This note does not have the pretension to cover the full field of finite element methods for the Navier–Stokes equations and is organized in sections as follows:

1. The Navier–Stokes equations for incompressible viscous flow
2. Some operator splitting methods for initial value problems and applications to the Navier–Stokes equations
3. Iterative solution of the advection–diffusion sub–problems and the wave-like equation method for the advection sub–problems
4. Iterative solution of the Stokes type sub–problem
5. Finite element approximation of the Navier–Stokes equations
6. Numerical results

Fecha de recepción: 21/10/2005

1 The Navier–Stokes equations for incompressible viscous flow

1.1 Model

Let Ω be an *open* and *connected* region (i.e. a domain) of $\mathbb{R}^d$ ($d = 2$ or 3) filled with a fluid. The generic point of $\mathbb{R}^d$ will be denoted by $x = \{x_i\}_{i=1}^d$ while dx will denote the elementary volume $dx_1dx_2$and $dx_1dx_2dx_3$ for $d = 2$ and $d = 3$, respectively.

Derivations of the Navier–Stokes equations may be found in, e.g., Prager 1961 [5], Batchelor 1967 [6], Guyon et al. 1991 [3], and Chorin et al. 1990 [7], and Glowinski 2003 [4]. Here skipping the details of derivation, we have the following so-called *momentum equation*

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \boldsymbol{\nabla})\mathbf{u} - \nu\Delta\mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \quad in \quad \Omega \times (0, T), \tag{1}$$

and the *continuity equation*

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0 \quad in \quad \Omega \times (0, T) \tag{2}$$

for unsteady, isothermal flows of incompressible, viscous, Newtonian fluids. In (1), (2) (and in the following),

1. $\mathbf{u} = \{u_i\}_{i=1}^d$ is the *velocity* and p is the *pressure*;
2. $\nu(> 0)$ is the *(kinematic) viscosity* coefficient;
3. $\boldsymbol{\nabla} = \{\dfrac{\partial}{\partial x_i}\}_{i=1}^d$, $\Delta = \displaystyle\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$, $\quad \mathbf{u} \cdot \mathbf{v} = \displaystyle\sum_{i=1}^d u_i v_i$, $\quad \forall \quad \mathbf{u} = \{u_i\}_{i=1}^d$, $\mathbf{v} = \{v_i\}_{i=1}^d$,
$$\boldsymbol{\nabla}\mathbf{u} \cdot \boldsymbol{\nabla}\mathbf{v} = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}, \ \forall \ \mathbf{u}, \mathbf{v}, \quad |\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v}, \ |\boldsymbol{\nabla}\mathbf{v}|^2 = \boldsymbol{\nabla}\mathbf{v} \cdot \boldsymbol{\nabla}\mathbf{v};$$
4. $\boldsymbol{\nabla} \cdot \mathbf{v} = \displaystyle\sum_{i=1}^d \frac{\partial v_i}{\partial x_i}$, $\forall \ \mathbf{v}$, $\quad (\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{w} = \{\displaystyle\sum_{j=1}^d v_j \frac{\partial w_i}{\partial x_j}\}_{i=1}^d \ \forall \ \mathbf{v}, \mathbf{w}$;
5. $\mathbf{f} = \{f_i\}_{i=1}^d$ is a density of external forces.

Let Γ be the boundary of Ω (here we suppose that Ω is bounded) and let $\mathbf{n}$ be the unit outward normal vector at Γ. Relations (1), (2) are not sufficient to define a flow; we have to consider further conditions, such as the *initial condition*

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x) \quad (with \ \boldsymbol{\nabla} \cdot \mathbf{u}_0 = 0), \tag{3}$$

and the *boundary condition*

$$\mathbf{u} = \mathbf{g} \quad on \quad \Gamma \times (0, T) \quad (with \int_\Gamma \mathbf{g} \cdot \mathbf{n} \, d\Gamma = 0). \tag{4}$$

The boundary condition (4) is of Dirichlet type; more complicated boundary conditions are described in, e.g., Glowinski 1984 [8], Bristeau et al. 1985 [9], and Pironneau 1989 [10], among them, the following mixed boundary condition which occurs often in applications

$$\mathbf{u} = \mathbf{g}_0 \quad on \quad \Gamma_0 \times (0,T), \quad \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - \mathbf{n}p = \mathbf{g}_1 \quad on \quad \Gamma_1 \times (0,T), \tag{5}$$

with $\dfrac{\partial \mathbf{u}}{\partial \mathbf{n}} = \{\dfrac{\partial u_i}{\partial \mathbf{n}}\}_{i=1}^{d} \left(= \{\boldsymbol{\nabla} u_i \cdot \mathbf{n}\}_{i=1}^{d}\right)$; where, in (5), Γ_0 and Γ_1 are two subsets of Γ satisfying $\Gamma_0 \cap \Gamma_1 = \emptyset$, closure of $\Gamma_0 \cup \Gamma_1 = \Gamma$. Another mixed boundary condition is

$$\mathbf{u} = \mathbf{g}_0 \quad on \quad \Gamma_0 \times (0,T), \quad \boldsymbol{\sigma}\mathbf{n} = \mathbf{g}_1 \quad on \quad \Gamma_1 \times (0,T), \tag{6}$$

with the (stress) tensor $\boldsymbol{\sigma} = 2\nu\mathbf{D}(\mathbf{u}) - p\mathbf{I}$ and $2\mathbf{D}(\mathbf{u}) = \boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^t$. The mixed boundary condition (5) is less physical than (6), but like (6), it is quite useful to implement *downstream boundary conditions* for flow in unbounded regions.

Remark 1 *The Dirichlet conditions in (4), (5) and (6) are called no-slip conditions if* $\mathbf{g} = \mathbf{0}$ *(resp.,* $\mathbf{g}_0 = \mathbf{0}$*) on* Γ *(resp.,* Γ_0*) if* Γ *(resp.,* Γ_0*) is not moving.* □

Remark 2 *The decrease in the popularity of the solution methods for the Navier-Stokes equations based on the stream function–vorticity formulation has been seen in these last years. We see two main reasons for this trend:*

(i) These methods are really convenient for two–dimensional flow. The generalization to three–dimensional flow, although possible, leads to complicated formulations.

(ii) The treatment of the boundary conditions is more delicate than with the velocity–pressure formulation, particularly for flow in multi–connected regions.

In this note, we will not discuss the stream function–vorticity formulation for the Navier-Stokes equations. □

Remark 3 *As today, it is not known if the time dependent Navier-Stokes equations modeling the unsteady flow of three-dimensional incompressible viscous fluids have a unique solution. For those readers who may be surprised that some decisive indications - in one direction or the other - have not been obtained via laboratory or computational experiments we would like to make the following comments:*

(i) The Navier-Stokes equations are just mathematical models (obtained after idealization) for some real life phenomena. Mathematical modeling cannot reflect the full complexity of a laboratory experimentation; indeed, it is

practically impossible to reproduce exactly a given experiment in order to validate its results by those of another one. We also have to remember that at large Reynolds numbers (the interesting case) small perturbations in the data can imply very large differences in the ensuing results.

(ii) Unlike their two-dimensional counterparts, three-dimensional viscous flows at large Reynolds numbers are not routine yet when it comes to numerical simulation. They require a lot of computer resources in time and memory. In order to explore the uniqueness issue it will be necessary to define significant test problems and store in a large data base the results obtained by solution methods using different type of space and time discretizations. We anticipate that this program will take place in the near future and that parallel computing will play an important role in this endeavor. □

Remark 4 *The mathematical theory of the Navier-Stokes equations for incompressible viscous fluids has inspired many investigators. The first rigorous mathematical results were obtained by J. Leray who proved (in Leray 1934 [11]) the existence of solutions when the flow region* Ω *is the full space* $\mathbb{R}^d$ *with* $d = 2$ *or* 3. *The Leray's results were extended to flow regions with boundaries by Leray himself [12] in 1934 and by E. Hopf [13] in 1951. The methods and tools developed by the above two authors have proved to be very useful to the solution of many problems in mechanics, physics, etc., modeled by linear or nonlinear partial differential equations, many of these problems being outside the field of fluid mechanics. The results of J. Leray and E. Hopf have been improved and generalized by several authors (see Leray 1994 [14] for an historical account), one of the most remarkable milestones in that direction being the proof by J.L. Lions and G. Prodi [15] in 1959 that if the flow region is two-dimensional (i.e.* $\Omega \subset \mathbb{R}^2$*), then the time-dependent Navier-Stokes equations have a unique solution. The proof of these existence and uniqueness results and of many others (on the regularity of the solutions, for example) can be found in the books by, e.g., J.L. Lions 1961 (Chapter 10 in [16]), J.L. Lions 1969 (Chapter 1 in [17]), Ladysenskaya 1969 [18], Temam 1977 (Chapters 2 and 3 in [19]), Tartar 1978 [20], Kreiss and Lorenz 1989 (Chapters 9 and 10 in [21]), and P.L. Lions 1996 (Chapters 2 and 3 in [22]). Regarding the Handbook of Numerical Analysis an important source of results (and of methods to obtain them) is the Chapter 1 of Marion and Temam [23]. The above list is far from complete, and the books and articles mentioned contain bibliographical references worth consulting.* □

1.2 Variational Formulations of the Navier-Stokes Equations.

We return now to the Navier-Stokes equations for incompressible Newtonian viscous flow. When using $\boldsymbol{\sigma} = 2\nu\mathbf{D}(\mathbf{u}) - p\mathbf{I}$, we have then

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} = \mathbf{f} \ in\ \Omega\times(0,T), \tag{7}$$

$$\boldsymbol{\nabla}\cdot\mathbf{u} = 0 \ in\ \Omega\times(0,T), \tag{8}$$

$$\mathbf{u}(0) = \mathbf{u}_0 \ (with\ \boldsymbol{\nabla}\cdot\mathbf{u}_0 = 0), \tag{9}$$

that we complete by the following boundary conditions

$$\mathbf{u} = \mathbf{g}_0 \ on\ \Gamma_0\times(0,T), \quad \boldsymbol{\sigma}\mathbf{n} = \mathbf{g}_1 \ on\ \Gamma_1\times(0,T); \tag{10}$$

with Γ_0, Γ_1 as in Section 1.1.

We define now the functional space V_0 by

$$V_0 = \{\mathbf{v} \mid \mathbf{v}\in(H^1(\Omega))^d, \mathbf{v} = 0 \ on\ \Gamma_0\}. \tag{11}$$

Space V_0 is a Hilbert space for the scalar product and norm defined by

$$(\mathbf{v},\mathbf{w})_{V_0} = \textstyle\sum_{i=1}^d (v_i,w_i)_{H^1(\Omega)}, \quad \forall\mathbf{v} = \{v_i\}_{i=1}^d, \mathbf{w} = \{w_i\}_{i=1}^d \in V_0,$$
$$||\mathbf{v}||_{V_0} = (\textstyle\sum_{i=1}^d ||v_i||^2_{H^1(\Omega)})^{1/2}, \quad \forall\mathbf{v} = \{v_i\}_{i=1}^d \in V_0,$$

respectively. In the particular case where $\Gamma_0 \neq \emptyset$ (with $\displaystyle\int_{\Gamma_0} d\Gamma > 0$) and Ω is bounded, we can use over V_0 the scalar product and norm defined by

$$\{\mathbf{v},\mathbf{w}\} \longrightarrow \sum_{i=1}^d\sum_{j=1}^d \int_\Omega \frac{\partial v_i}{\partial x_j}\frac{\partial w_i}{\partial x_j}dx = \sum_{i=1}^d \int_\Omega \boldsymbol{\nabla} v_i\cdot\boldsymbol{\nabla} w_i \ dx,$$

$$\mathbf{v} \longrightarrow (\sum_{i=1}^d \int_\Omega |\boldsymbol{\nabla} v_i|^2 \ dx)^{1/2},$$

respectively. Suppose that $\mathbf{R}$ and $\mathbf{S}$ are two $d\times d$ tensors so that $\mathbf{R} = \{r_{ij}\}$, $\mathbf{S} = \{s_{ij}\}$; from now on we shall use the notation $\mathbf{R} : \mathbf{S}$ for $\displaystyle\sum_{i=1}^d\sum_{j=1}^d r_{ij}s_{ij}$. With this notation the above V_0-scalar product and norm can be written as

$$\int_\Omega \boldsymbol{\nabla}\mathbf{v}{:}\boldsymbol{\nabla}\mathbf{w} \ dx \quad \text{and} \quad (\int_\Omega \boldsymbol{\nabla}\mathbf{v}{:}\boldsymbol{\nabla}\mathbf{v} \ dx)^{1/2},$$

respectively. For simplicity, we shall use in the sequel the notation $|\boldsymbol{\nabla}\mathbf{v}|^2$ for $\boldsymbol{\nabla}\mathbf{v}{:}\boldsymbol{\nabla}\mathbf{v}$. We suppose now that the functions occurring in the system (7)-(9) are sufficiently smooth; taking the $\mathbb{R}^d$-dot product of both sides of (7) with $\mathbf{v}$,

an arbitrary element of V_0, and then integrating over Ω we obtain from Green's formula that for almost any t on $(0,T)$ we have

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}(t)}{\partial t}\cdot \mathbf{v}\ dx + \int_\Omega (\mathbf{u}(t)\cdot\boldsymbol{\nabla})\mathbf{u}(t)\cdot\mathbf{v}\ dx + 2\nu\int_\Omega \mathbf{D}(\mathbf{u}(t)) : \mathbf{D}(\mathbf{v})\ dx \\ \displaystyle -\int_\Omega p(t)\boldsymbol{\nabla}\cdot\mathbf{v}\ dx = \int_\Omega \mathbf{f}(t)\cdot\mathbf{v}\ dx + \int_{\Gamma_1}\mathbf{g}_1(t)\cdot\mathbf{v}\ d\Gamma, \forall \mathbf{v}\in V_0, \end{cases} \tag{12}$$

to be completed by (8), (9) and

$$\mathbf{u} = \mathbf{g}_0 \ on\ \Gamma_0\times(0,T). \tag{13}$$

The "Neumann" condition $\boldsymbol{\sigma}\mathbf{n} = \mathbf{g}_1$ on $\Gamma_1\times(0,T)$ is *automatically* enforced by the formulation (12), which is known as a *variational formulation* of the momentum equation (7). Actually, it can be shown that relation (12) *implies* the momentum equation (7) *and* the "Neumann" condition $\boldsymbol{\sigma}\mathbf{n} = \mathbf{g}_1$ on $\Gamma_1\times(0,T)$.

Suppose now that instead of (10) the boundary conditions are given by

$$\mathbf{u} = \mathbf{g}_0 \ on\ \Gamma_0\times(0,T),\quad \nu\frac{\partial \mathbf{u}}{\partial n} - p\mathbf{n} = \mathbf{g}_1 \ on\ \Gamma_1\times(0,T). \tag{14}$$

Multiplying both sides of (1) by $\mathbf{v}\in V_0$, integrating over Ω and using Green's formula we obtain this time that for almost any $t\in(0,T)$ we have

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t}(t)\cdot\mathbf{v}\,dx + \int_\Omega (\mathbf{u}(t)\cdot\boldsymbol{\nabla})\mathbf{u}(t)\cdot\mathbf{v}\,dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{u}(t) : \boldsymbol{\nabla}\mathbf{v}\,dx \\ \displaystyle -\int_\Omega p(t)\boldsymbol{\nabla}\cdot\mathbf{v}\,dx = \int_\Omega \mathbf{f}(t)\cdot\mathbf{v}\,dx + \int_{\Gamma_1}\mathbf{g}_1(t)\cdot\mathbf{v}d\Gamma, \forall\mathbf{v}\in V_0. \end{cases} \tag{15}$$

Conversely, the variational formulation (15) implies the momentum equation (1) and the generalized Neumann condition $\nu\frac{\partial \mathbf{u}}{\partial n} - p\mathbf{n} = \mathbf{g}_1$ on $\Gamma_1\times(0,T)$.

The variational formulations (12) and (15) of the momentum equation will play a fundamental role in the *finite element* approximation of the Navier-Stokes problems (7)-(10) and (1)-(3), (5), respectively. We shall return to this issue. Actually, for the finite element approximations of the above problems, we shall take advantage of the fact that the *incompressibility condition* $\boldsymbol{\nabla}\cdot\mathbf{u} = 0$ is *equivalent* to

$$\int_\Omega q\boldsymbol{\nabla}\cdot\mathbf{u}\,dx \ = \ 0, \forall q\in L^2(\Omega). \tag{16}$$

2 Operator Splitting Methods for Initial Value Problems: Application to the Navier–Stokes Equations

Solving the above Navier-Stokes equations is a non-trivial task for the following reasons:

(i) the momentum equation is *nonlinear*;

(ii) the incompressibility condition $\boldsymbol{\nabla}\cdot\mathbf{u} = 0$;

(iii) solving the Navier-Stokes equations amounts to solve a *system* of partial differential equations ($d+1$ if $\Omega \subset \mathbb{R}^d$) coupled through the nonlinear term $(\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u}$, the incompressibility condition $\boldsymbol{\nabla}\cdot\mathbf{u}=0$, and sometimes through the viscous term and the boundary conditions (as it is the case in (5) and (6)).

In the following subsections we will only focus on *time discretization* by *operator-splitting* schemes, *theta*–scheme and Marchuk–Yanenko scheme (the details of other well-known operator-splitting schemes, such as Peaceman–Rachford method, Douglas–Rachford method, Crank–Nicolson method, alternating direction method, and etc., can be found in, e.g., Glowinski 2003 [4]), which will partly overcome the above difficulties; in particular, we will be able to decouple the difficulties associated to the *non-linearity* with those associated to the *incompressibility condition*.

2.1 A Family of Initial Value Problems.

We consider the following *initial value problem*:

$$\frac{d\varphi}{dt} + A(\varphi, t) = 0, \quad \varphi(0) = \varphi_0, \tag{17}$$

where, for a given t, A is an operator (possibly nonlinear, and even multivalued) from a Hilbert space H into itself and where $\varphi_0 \in H$.

Suppose now that operator A has the following *nontrivial decomposition* $A = A_1 + A_2$ (by *nontrivial* we mean that A_1 and A_2 are individually simpler than A). It is then quite natural to integrate the initial value problem (17) by numerical methods taking advantage of the decomposition property, $A = A_1 + A_2$; such a goal can be achieved by the *operator splitting schemes*. discussed in the following subsections.

2.2 A θ–scheme.

This scheme, introduced in Glowinski 1985 and 1986 [32, 33], is a variation of schemes discussed in Strang 1968 [34], Beale and Majda 1981 [35], Leveque and Oliger 1983 [36]; it is discussed with further details in Glowinski and Le Tallec 1989 [31]. The θ–scheme to be described below is in fact a variant of the Peaceman–Rachford scheme.

Let θ be a number of the open interval $(0, \frac{1}{2})$ (in practice $\theta \in (0, \frac{1}{3})$); the θ-scheme applied to the solution of the initial value problem (17), when $A = A_1 + A_2$, is described as follows:

$$\varphi^0 = \varphi_0; \tag{18}$$

then for $n \geq 0$, φ^n *being known, we compute* $\varphi^{n+\theta}, \varphi^{n+1-\theta}$ and φ^{n+1} *as follows*:

$$\frac{\varphi^{n+\theta} - \varphi^n}{\theta \Delta t} + A_1\left(\varphi^{n+\theta}, (n+\theta)\Delta t\right) + A_2(\varphi^n, n\Delta t) = 0, \tag{19}$$

$$\frac{\varphi^{n+1-\theta} - \varphi^{n+\theta}}{(1-2\theta)\Delta t} + A_1\left(\varphi^{n+\theta}, (n+\theta)\Delta t\right) + A_2\left(\varphi^{n+1-\theta}, (n+1-\theta)\Delta t\right) = 0 \tag{20}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1-\theta}}{\theta \Delta t} + A_1\left(\varphi^{n+1}, (n+1)\Delta t\right) + A_2\left(\varphi^{n+1-\theta}, (n+1-\theta)\Delta t\right) = 0 \tag{21}$$

We consider now the simple situation where $H = \mathbb{R}^N$, $\varphi_0 \in \mathbb{R}^N$, where A is an $N \times N$ matrix, *symmetric, positive definite* and *independent* of t. The solution of the corresponding *autonomous* system (17) is then

$$\varphi(t) = e^{-At}\varphi_0. \tag{22}$$

If one projects (22) over a vector basis of $\mathbb{R}^N$, consisting of *eigenvectors* of A, we obtain - with obvious notation -

$$\varphi_i(t) = e^{-\lambda_i t}\varphi_{0i}, i = 1, ...N, \tag{23}$$

where $0 < \lambda_1 \leq \lambda_2 ... \leq \lambda_N$ denote the *eigenvalues* of A.

In order to apply scheme (18)–(21), we consider the following decomposition of matrix A

$$A = \alpha A + \beta A, \tag{24}$$

with $\alpha + \beta = 1, 0 < \alpha, \beta < 1$. Applying (18)–(21) with $A_1 = \alpha A, A_2 = \beta A$ yields

$$\varphi^{n+1} = (I + \alpha\theta\Delta t A)^{-2}(I - \beta\theta\Delta t A)^2(I + \beta\theta'\Delta t A)^{-1}(I - \alpha\theta'\Delta t A)\varphi^n, \tag{25}$$

where $\theta' = 1 - 2\theta$, which implies

$$\varphi_i^n = \frac{(1 - \beta\theta\Delta t\lambda_i)^{2n}(1 - \alpha\theta'\Delta t\lambda_i)^n}{(1 + \alpha\theta\Delta t\lambda_i)^{2n}(1 + \beta\theta'\Delta t\lambda_i)^n}\ \varphi_{0i}, \forall i = 1, ...N. \tag{26}$$

Consider now the rational function R_1 defined by

$$R_1(\xi) = \frac{(1 - \beta\theta\xi)^2(1 - \alpha\theta'\xi)}{(1 + \alpha\theta\xi)^2(1 + \beta\theta'\xi)}. \tag{27}$$

Since

$$\lim_{\xi \to +\infty} |R_1(\xi)| = \beta/\alpha, \tag{28}$$

we should prescribe the condition

$$\alpha > \beta \tag{29}$$

which is a *necessary* one for the *stiff A-stability* of the θ–scheme (18)–(21). To obtain the *unconditional stability* we need to have

$$|R_1(\xi)| \leq 1, \forall \xi \in \mathbb{R}_+;$$

actually, a closer inspection of the function R_1 would show that

$$|R_1(\xi)| < 1, \forall \xi > 0, \forall \theta \in [\frac{1}{4}, \frac{1}{2}), \forall \alpha, \beta \; so \; that \; 0 < \beta < \alpha < 1, \alpha + \beta = 1, \quad (30)$$

which implies the unconditional stability of scheme (18)–(21) with respect to Δt (the *lower bound* $\frac{1}{4}$ in (30) is *not optimal* for θ, but we shall be satisfied with it since, as we shall see below, the "optimal" value of θ is $1 - \frac{1}{\sqrt{2}} = 0.292893219... > \frac{1}{4}$).

Concerning now the accuracy of scheme (18)–(21), we can show that in the neighborhood of $\xi = 0, R_1$ satisfies:

$$R_1(\xi) = 1 - \xi + \frac{\xi^2}{2}[1 + (\beta - \alpha)(2\theta^2 - 4\theta + 1)] + O(1)\xi^3. \quad (31)$$

Comparing (31) to the expansion of $e^{-\xi}$

$$e^{-\xi} \; = \; 1 - \xi + \frac{\xi^2}{2} - \frac{\xi^3}{6} + O(1)\xi^4, \quad (32)$$

we obtain that scheme (18)–(21) is *second order accurate if and only if*

$$\alpha = \beta (= \frac{1}{2} \; from \; \alpha + \beta = 1), \quad (33)$$

and/or

$$\theta = 1 - 1/\sqrt{2} = .292893219...; \quad (34)$$

scheme (18)–(21) is *first order accurate* if neither (33) nor (34) holds. If one takes $\alpha = \beta = \frac{1}{2}$ it follows from (26) and (27) that scheme (18)–(21) is *unconditionally stable*, $\forall \theta \in (0, \frac{1}{2})$; however, we have (from (28))

$$\lim_{\xi \to +\infty} |R_1(\xi)| \; = \; 1, \quad (35)$$

implying that in that particular case scheme (18)–(21) *is not stiff A-stable*. Relations (23) show that the larger λ_i, the faster $\varphi_i(t)$ converges to zero as $t \to +\infty$; considering now the discrete analogue of (23), namely (26) we observe that for large values of $\lambda_i \Delta t$ we have $R_1(\lambda_i \Delta t) \sim 1$, implying that, in (26), φ_i^n converges slowly to zero as $n \to +\infty$; from this property (which is also shared by the *Peaceman–Rachford scheme scheme*, the *Douglas–Rachford scheme scheme*, and the *Crank–Nicolson scheme*) we can expect scheme (18)–(21) with $\alpha = \beta = \frac{1}{2}$ and $\theta \in (0, \frac{1}{2})$ to be not well suited (unless Δt is very small) to simulate fast transient phenomena and to capture efficiently the possible *steady state* solutions of (17) (i.e. the solutions of $A(\varphi, +\infty) = 0$), if operator

A is *stiff* (the notion of *stiffness* is defined in, e.g., Crouzeix and Mignot 1984 [30] (pages 86 to 88).

Let us consider now the case where α and β have been chosen so that *we have the same matrix for all the partial steps of the $\theta-$scheme*; in that case α, β, θ have to satisfy

$$\alpha\theta = \beta(1-2\theta), \tag{36}$$

which implies

$$\alpha \ = \ (1-2\theta)/(1-\theta), \ \beta \ = \ \theta/(1-\theta). \tag{37}$$

Combining (29) and (37) yields

$$0 < \theta < 1/3; \tag{38}$$

for $\theta = 1/3$, (37) implies $\alpha = \beta = 1/2$, a situation which has been discussed already.

If $0 < \theta < 1/3$ and if α and β are given by (37) we have

$$\lim_{\xi \to +\infty} |R_1(\xi)| = \beta/\alpha = \theta/(1-2\theta) < 1. \tag{39}$$

Indeed, we can prove that if $\theta^* \leq \theta \leq 1/3$ (with $\theta^* = .087385580...$) and if α and β are given by (37), then scheme (18)-(21) is *unconditionally stable*; moreover if $\theta^* < \theta < 1/3$ (with α and β still given by (37)), property (39) implies that scheme (18)-(21) is *stiff A-stable* and has therefore good asymptotic properties as $n \to +\infty$, *making it well suited to compute steady state solutions.*

If $\theta = 1-1/\sqrt{2}$ (resp., $\theta = 1/4$) we have $\alpha = 2-\sqrt{2}, \beta = \sqrt{2}-1, \beta/\alpha = 1/\sqrt{2}$ (resp., $\alpha = 2/3, \beta = 1/3, \beta/\alpha = 1/2$).

Remark 5 *We consider the case where in (17) we have*

$$A(\varphi, t) = B(\varphi) - f(t) \text{ with } B = B_1 + B_2. \tag{40}$$

In order to decide how to decompose f when applying the θ-scheme (18)-(21) to the solution of the initial value problem

$$\begin{cases} \dfrac{d\varphi}{dt} + B(\varphi) = f, \\ \varphi(0) \ = \ \varphi_0, \end{cases} \tag{41}$$

as before, we suppose therefore, that

$$f = f_1 + f_2 \tag{42}$$

with $f_1 = \alpha f$, $f_2 = \beta f$, $0 \leq \alpha, \ \beta \leq 1$, $\alpha + \beta = 1$, and we assume that $B = 0$, for simplicity.

Applying scheme (18)-(21) to the solution of

$$\frac{d\varphi}{dt} = f, \ \varphi(0) = \varphi_0, \tag{43}$$

we obtain (with $\theta' = 1 - 2\theta$*)*

$$\varphi^0 = \varphi_0, \tag{44}$$

and for $n \geq 0$*,*

$$\frac{\varphi^{n+\theta} - \varphi^n}{\theta \Delta t} = \alpha f\left((n+\theta)\Delta t\right) + \beta f(n\Delta t). \tag{45}$$

$$\frac{\varphi^{n+1-\theta} - \varphi^{n+\theta}}{\theta' \Delta t} = \alpha f\left((n+\theta)\Delta t\right) + \beta f\left((n+1-\theta)\Delta t\right), \tag{46}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1-\theta}}{\theta \Delta t} = \alpha f\left((n+1)\Delta t\right) + \beta f\left((n+1-\theta)\Delta t\right), \tag{47}$$

which imply that

$$\begin{cases} \varphi^n = \varphi_0 + \Delta t \sum_{q=0}^{n-1} \{\beta\theta f(q\Delta t) + \alpha(1-\theta) f\left((q+\theta)\Delta t\right) + \\ \beta(1-\theta) f\left((q+1-\theta)\Delta t\right) + \alpha\theta f\left((q+1)\Delta t\right)\}. \end{cases} \tag{48}$$

Since $\beta\theta + \alpha(1-\theta) + \beta(1-\theta) + \alpha\theta = 1$*, the numerical integration rule which, in (48), approximates* $\int_{q\Delta t}^{(q+1)\Delta t} f(t)dt$*, is first-order accurate, at least; actually, it is second-order accurate, if and only if*

$$\alpha(1-\theta)\theta + \beta(1-\theta)^2 + \alpha\theta = \frac{1}{2},$$

or equivalently

$$(\beta - \alpha)(2\theta^2 - 4\theta + 1) = 0. \tag{49}$$

Not surprisingly, we recover from (49) conditions (33) and (34), namely scheme (44)-(47) is second-order accurate if and only if

$$\alpha = \beta = \frac{1}{2} \tag{50}$$

and/or

$$\theta = 1 - 1/\sqrt{2}. \tag{51}$$

Assuming that (51) holds, we can wonder if there are values of α *and* β *for which scheme (44)-(47) is third-order accurate; this will be the case if and only if the numerical integration rule in (48) is exact for second degree polynomials, i.e. if and only if*

$$\alpha(1-\theta)\theta^2 + \beta(1-\theta)^3 + \alpha\theta = \frac{1}{3} \tag{52}$$

with $\theta = 1 - 1/\sqrt{2}$ *in (52). Taking* $\beta = 1 - \alpha$ *into account, it follows from (52) that*

$$\alpha(2\theta^2 - 4\theta + 1) = (1-\theta)^3 - 1/3.$$

which implies in turn, since (51) holds, that

$$0 = \frac{3 - 2\sqrt{2}}{6\sqrt{2}} \tag{53}$$

which makes no sense. Strictly speaking, therefore, if $\theta = 1 - 1/\sqrt{2}$ *scheme (44)-(47) is never third-order accurate,* $\forall \alpha, \beta$, *so that* $0 \leq \alpha, \beta \leq 1, \alpha + \beta = 1$. *However, since* $\frac{3 - 2\sqrt{2}}{6\sqrt{2}} \simeq 2 \times 10^{-2}$ *we can say that (52) is "almost" verified, implying that scheme (44)-(47) is "not far" from being third-order accurate if* $\theta = 1 - 1/\sqrt{2}$. *Similarly, if* $\alpha = \beta = 1/2$, *we can prove that there is no value of* θ *in (0,1/2) so that scheme (44)-(47) is third-order accurate.*

From the above results, we suggest to proceed as follows when applying the θ*-scheme (18)-(21) to the solution of the initial value problem (40):*

1) If $\theta \neq 1 - 1/\sqrt{2}$, *use*

$$\varphi^0 = \varphi_0, \tag{54}$$

and for $n \geq 0$

$$\frac{\varphi^{n+\theta} - \varphi^n}{\theta \Delta t} + B_1(\varphi^{n+\theta}) + B_2(\varphi^n) = \frac{1}{2}(f^{n+\theta} + f^n), \tag{55}$$

$$\frac{\varphi^{n+1-\theta} - \varphi^{n+\theta}}{(1 - 2\theta)\Delta t} + B_1(\varphi^{n+\theta}) + B_2(\varphi^{n+1-\theta}) = \frac{1}{2}(f^{n+\theta} + f^{n+1-\theta}), \tag{56}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1-\theta}}{\theta \Delta t} + B_1(\varphi^{n+1}) + B_2(\varphi^{n+1-\theta}) = \frac{1}{2}(f^{n+1} + f^{n+1-\theta}). \tag{57}$$

2) If $\theta = 1 - 1/\sqrt{2}$ *we can still use scheme (54)-(57), but simpler choices are provided by*

$$\varphi^0 = \varphi_0, \tag{58}$$

and for $n \geq 0$

$$\frac{\varphi^{n+\theta} - \varphi^n}{\theta \Delta t} + B_1(\varphi^{n+\theta}) + B_2(\varphi^n) = f^{n+\theta}, \tag{59}$$

$$\frac{\varphi^{n+1-\theta} - \varphi^{n+\theta}}{(1 - 2\theta)\Delta t} + B_1(\varphi^{n+\theta}) + B_2(\varphi^{n+1-\theta}) = f^{n+\theta}, \tag{60}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1-\theta}}{\theta \Delta t} + B_1(\varphi^{n+1}) + B_2(\varphi^{n+1-\theta}) = f^{n+1} \tag{61}$$

(which corresponds to $\{\alpha, \beta\} = \{1, 0\}$*) and by*

$$\varphi^0 = \varphi_0, \tag{62}$$

and for $n \geq 0$

$$\frac{\varphi^{n+\theta} - \varphi^n}{\theta \Delta t} + B_1(\varphi^{n+\theta}) + B_2(\varphi^n) = f^n, \tag{63}$$

$$\frac{\varphi^{n+1-\theta} - \varphi^{n+\theta}}{(1 - 2\theta)\Delta t} + B_1(\varphi^{n+\theta}) + B_2(\varphi^{n+1-\theta}) = f^{n+1-\theta}, \tag{64}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1-\theta}}{\theta \Delta t} + B_1(\varphi^{n+1}) + B_2(\varphi^{n+1-\theta}) = f^{n+1-\theta} \tag{65}$$

(which corresponds to $\{\alpha, \beta\} = \{0, 1\}$*).* □

2.3 Fractional–step scheme à la Marchuk–Yanenko

Among the many operator–splitting methods which can be employed to solve (17), we also advocate (following, e.g., Marchuk 1990 [29]) the very simple one below; it is only first order accurate, but its low order accuracy is compensated by easy implementation, less cost in computation, good stability, and robustness properties. We consider the initial value problem (17) with $A = A_1 + A_2$ where A_1 and A_2 are linear and independent of t; we have then (at least formally)

$$\varphi(t) = e^{-(A_1+A_2)t}\varphi_0. \tag{66}$$

We consider a time discretization step $\Delta t(> 0)$ and denote $(n + \alpha)\Delta t$ by $t^{n+\alpha}$. Then from (66) we have

$$\varphi(t^{n+1}) = e^{-(A_1+A_2)\Delta t}\varphi(t^n). \tag{67}$$

Now we suppose that A_1 and A_2 do not commute. We have then

$$e^{-(A_1+A_2)\Delta t} = e^{-A_2\Delta t}e^{-A_1\Delta t} + O(\Delta t^2). \tag{68}$$

Relation (68) leads to the following first order scheme for the solution of problem (17):

$$\varphi^0 = \varphi_0, \tag{69}$$

for $n \geq 0, \varphi^n$ being known, we compute $\varphi^{n+1/2}, \varphi^{n+1}$ via the solution of two initial value problems below:

$$d\varphi/dt + A_1\varphi = 0 \text{ on } (t^n, t^{n+1}),\ \varphi(t^n) = \varphi^n;\ \varphi^{n+1/2} = \varphi(t^{n+1}), \tag{70}$$

$$d\varphi/dt + A_2\varphi = 0 \text{ on } (t^n, t^{n+1}), \varphi(t^n) = \varphi^{n+1/2};\ \varphi^{n+1} = \varphi(t^{n+1}), \tag{71}$$

We consider again the simple situation where $H = \mathbb{R}^N$, $\varphi_0 \in \mathbb{R}^N$, where A is an $N \times N$ matrix, *symmetric, positive definite* and *independent* of t. Applying (70), (71) with $A_1 = \alpha A, A_2 = \beta A$ satisfying $\alpha + \beta = 1$, $0 < \alpha, \beta < 1$ and backward Euler method yields

$$\frac{\varphi^{n+1/2} - \varphi^n}{\Delta t} + \alpha A\varphi^{n+1/2} = 0, \tag{72}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1/2}}{\Delta t} + \beta A\varphi^{n+1} = 0, \tag{73}$$

and

$$\varphi^{n+1} = (I + \beta\Delta tA)^{-1}(I + \alpha\Delta tA)^{-1}\varphi^n$$

which implies

$$\varphi_i^n = \frac{\varphi_{0i}}{(1 + \beta\Delta t\lambda_i)^n(1 + \alpha\Delta t\lambda_i)^n}, \forall i = 1, ...N.$$

Hence $\varphi_i^n \to 0$ as $n \to \infty$ for all α, β satisfying $\alpha + \beta = 1$, $0 < \alpha, \beta < 1$. So the scheme is unconditionally stable. Consider now the rational function R_2 defined by

$$R_2(\xi) \ = \ (1 + \beta\xi)^{-1}(1 + \alpha\xi)^{-1}.$$

We have

$$R_2(\xi) \ = 1 - \xi + \frac{1}{2}\xi^2 + O(1)\xi^3.$$

Comparing above expansion of $R_2(\xi)$ to the expansion of $e^{-\xi}$ in (32), we obtain that scheme à la Marchuk–Yanenko is first order accurate (due to the way we approximate the problem (17) by two problems (70) and (71)) and unconditionally stable at least for the above simple case consideration.

Remark 6 *A second order scheme can be obtained by symmetrization (see, e.g., Dean and Glowinski 1997 [100] and Dean, Glowinski, and Pan [101] for the application of symmetrized splitting schemes to the solution of the Navier-Stokes equations).*

Remark 7 *We consider again the case where in (17) we have* $A(\varphi, t) = -f(t)$. *As before, we suppose that, that* $f = f_1 + f_2$ *with* $f_1 = \alpha f$, $f_2 = \beta f$ *and* $0 \leq \alpha, \ \beta \leq 1$, $\alpha + \beta = 1$.

Applying scheme (72)-(73) to the solution of

$$\frac{d\varphi}{dt} = f, \ \varphi(0) = \varphi_0,$$

we obtain

$$\varphi^0 = \varphi_0, \tag{74}$$

and for $n \geq 0$,

$$\frac{\varphi^{n+1/2} - \varphi^n}{\Delta t} = \alpha f(t^{n+1}), \tag{75}$$

$$\frac{\varphi^{n+1} - \varphi^{n+1/2}}{\Delta t} = \beta f(t^{n+1}), \tag{76}$$

which imply that

$$\varphi^n = \varphi_0 + \Delta t \sum_{q=1}^{n} f(t^q).$$

Hence it is first order accurate if f' *is continuous on* $[0, t^n]$.

2.4 Application to the Navier-Stokes equations.

We discuss now the application of the time discretization schemes described in the above sections to the solution of the *time-dependent Navier-Stokes equations* (1)-(3), (5).

Actually, we shall consider application of the θ-scheme, since it is the one which gives the best results regarding accuracy and convergence to steady-state solutions. We obtain then the following time discretization scheme (with $0 < \alpha < 1, 0 < \beta < 1$ and $\alpha + \beta = 1$) :

$$\mathbf{u}^0 = \mathbf{u}_0; \tag{77}$$

then for $n \geq 0, \mathbf{u}^n$ *being known, we compute* $\mathbf{u}^{n+\theta}, \mathbf{u}^{n+1-\theta}$ *and* $\mathbf{u}^{n+1}$ *via the solution of*

$$\frac{\mathbf{u}^{n+\theta} - \mathbf{u}^n}{\theta \Delta t} - \alpha\nu\Delta\mathbf{u}^{n+\theta} + \boldsymbol{\nabla} p^{n+\theta} = \mathbf{f}^{n+\theta} + \beta\nu\Delta\mathbf{u}^n - (\mathbf{u}^n \cdot \boldsymbol{\nabla})\mathbf{u}^n \text{ in } \Omega, \tag{78}$$

$$\boldsymbol{\nabla} \cdot \mathbf{u}^{n+\theta} = 0 \text{ in } \Omega, \tag{79}$$

$$\mathbf{u}^{n+\theta} = \mathbf{g}_0^{n+\theta} \text{ on } \Gamma_0,$$

$$\alpha\nu\frac{\partial \mathbf{u}^{n+\theta}}{\partial n} - \mathbf{n}p^{n+\theta} = \mathbf{g}_1^{n+\theta} - \beta\nu\frac{\partial \mathbf{u}^n}{\partial n} \text{ on } \Gamma_1, \tag{80}$$

and then, of

$$\frac{\mathbf{u}^{n+1-\theta} - \mathbf{u}^{n+\theta}}{(1-2\theta)\Delta t} - \beta\nu\Delta\mathbf{u}^{n+1-\theta} + (\mathbf{u}^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}^{n+1-\theta} = \mathbf{f}^{n+\theta} + \alpha\nu\Delta\mathbf{u}^{n+\theta} - \boldsymbol{\nabla} p^{n+\theta} \text{ in } \Omega, \tag{81}$$

$$\mathbf{u}^{n+1-\theta} = \mathbf{g}_0^{n+1-\theta} \text{ on } \Gamma_0,$$

$$\beta\nu\frac{\partial \mathbf{u}^{n+1-\theta}}{\partial n} = \mathbf{g}_1^{n+\theta} + \mathbf{n}p^{n+\theta} - \alpha\nu\frac{\partial \mathbf{u}^{n+\theta}}{\partial n} \text{ on } \Gamma_1, \tag{82}$$

and finally, of

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^{n+1-\theta}}{\theta \Delta t} - \alpha\nu\Delta\mathbf{u}^{n+1} + \boldsymbol{\nabla} p^{n+1} = \mathbf{f}^{n+1} + \beta\nu\Delta\mathbf{u}^{n+1-\theta} - (\mathbf{u}^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}^{n+1-\theta} \text{ in } \Omega, \tag{83}$$

$$\boldsymbol{\nabla} \cdot \mathbf{u}^{n+1} = 0 \text{ in } \Omega, \tag{84}$$

$$\mathbf{u}^{n+1} = \mathbf{g}_0^{n+1} \text{ on } \Gamma_0,$$

$$\alpha\nu\frac{\partial \mathbf{u}^{n+1}}{\partial n} - \mathbf{n}p^{n+1} = \mathbf{g}_1^{n+1} - \beta\nu\frac{\partial \mathbf{u}^{n+1-\theta}}{\partial n} \text{ on } \Gamma_1; \tag{85}$$

the choice of α and β will be discussed below. We observe that using the θ-scheme we have been able to *decouple* the nonlinearity and the incompressibility in the Navier-Stokes equations (1)-(3), (5). We observe also that $\mathbf{u}^{n+\theta}$ and $\mathbf{u}^{n+1}$ are obtained from the solution of *linear* problems very close to the *Stokes problem*

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \text{ in } \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \text{ in } \Omega, \\ \mathbf{u} = \mathbf{g}_0 \text{ on } \Gamma_0, \quad \nu\dfrac{\partial \mathbf{u}}{\partial n} - \mathbf{n}p = \mathbf{g}_1 \text{ on } \Gamma_1. \end{cases} \tag{86}$$

In Sections 3 and 4, we shall describe the specific treatment of the subproblems encountered at each step of scheme (77)-(85). Concerning now the choice of α and β, we advocate the one given by (37); with such a choice many computer subprograms are *common* to both the linear and nonlinear subproblems, saving therefore quite a substantial amount of core memory. Concerning θ, numerical experiments show that $\theta = 1 - 1/\sqrt{2}$ seems to produce the best results, even in those situations where the Reynolds number is large.

Remark 8 *Numerical experiments show that there is practically no loss in accuracy and stability by replacing* $(\mathbf{u}^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}^{n+1-\theta}$ *by* $(\mathbf{u}^{n+\theta} \cdot \boldsymbol{\nabla})\mathbf{u}^{n+1-\theta}$ *in (81). This observation has important practical consequences since the following problem*

$$\begin{cases} \dfrac{\mathbf{u}^{n+1-\theta} - \mathbf{u}^{n+\theta}}{(1-2\theta)\Delta t} - \beta\nu\Delta\mathbf{u}^{n+1-\theta} + (\mathbf{u}^{n+\theta} \cdot \boldsymbol{\nabla})\mathbf{u}^{n+1-\theta} = \\ \mathbf{f}^{n+\theta} + \alpha\nu\Delta\mathbf{u}^{n+\theta} - \boldsymbol{\nabla}p^{n+\theta} \text{ in } \Omega, \\ \mathbf{u}^{n+1-\theta} = \mathbf{g}_0^{n+1-\theta} \text{ on } \Gamma_0, \beta\nu\dfrac{\partial\mathbf{u}^{n+1-\theta}}{\partial n} = \mathbf{g}_1^{n+\theta} + \mathbf{n}p^{n+\theta} - \alpha\nu\dfrac{\partial\mathbf{u}^{n+\theta}}{\partial n} \text{ on } \Gamma_1, \end{cases} \tag{87}$$

being linear, is easier to solve than the nonlinear problem (81). □

Remark 9 *Operator splitting methods have always been popular tools for the numerical simulation of incompressible viscous flow. To be more precise, the so-called projection methods, which have been used for more than thirty years now, for solving the Navier-Stokes equations can be viewed as operator splitting methods. The projection methods can also be viewed as predictor-corrector schemes, where a predicted value (not necessarily divergence-free) of the approximate solution at time* $(n+1)\Delta t$ *is projected in the* $L^2(\Omega)$*-sense over an appropriate space of divergence-free functions. We will discuss a projection method obtained by the scheme à la Marchuk-Yanenko in Section 5.3 (to appear in Part II). To our knowledge, projection methods for solving the Navier-Stokes equations have been introduced by Chorin 1967 and 1968 [38, 39] and Temam 1969 [40, 41]; the original projection methods contained several drawbacks, concerning particularly the quality of the approximate pressure at low Reynolds numbers, but, fortunately, these flaws have been essentially eliminated in the modern projection methods. A concise, but fairly complete introduction to projection schemes can be found in Quarteroni and Valli 1994 [42] (Section 13.5), a more detailed one being Marion and Temam 1998 [23] (Chapter 3).*

3 Classical and Variational Formulations of the Advection-Diffusion Subproblems Associated with the Operator Splitting Schemes.

At each full step of scheme (77)-(85) we have to solve a *nonlinear elliptic system* of the following type (with Ω, Γ, Γ_0 and Γ_1 as in Section 1):

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u} = \mathbf{f} \ in\ \Omega, \\ \mathbf{u} = \mathbf{g}_0 \ on\ \Gamma_0,\ \nu\dfrac{\partial\mathbf{u}}{\partial n} = \mathbf{g}_1 \ on\ \Gamma_1, \end{cases} \tag{88}$$

where α and ν are two positive constants, and $\mathbf{f}, \mathbf{g}_0$ and $\mathbf{g}_1$ are three given functions, defined on Ω, Γ_0 and Γ_1, respectively. We shall not discuss here the existence and uniqueness of solution for problem (88), which can be found in, e.g., Glowinski 2003 [4] (Section 15). We consider now the following functional spaces of *Sobolev* type:

$$H^1(\Omega) = \{\varphi | \varphi \in L^2(\Omega),\ \frac{\partial\varphi}{\partial x_i} \in L^2(\Omega),\ \forall i = 1, ...d\}, \tag{89}$$

$$H_0^1(\Omega) = \{\varphi | \varphi \in H^1(\Omega),\ \varphi = 0 \ on\ \Gamma\}, \tag{90}$$

$$V_0 = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d,\ \mathbf{v} = \mathbf{0} \ on\ \Gamma_0\}, \tag{91}$$

$$V_g = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d,\ \mathbf{v} = \mathbf{g}_0 \ on\ \Gamma_0\}; \tag{92}$$

if $\mathbf{g}_0$ is sufficiently smooth, then space V_g is nonempty.

Using *Green's formula* we can prove that for sufficiently smooth functions $\mathbf{u}$ and $\mathbf{v}$ belonging to $(H^1(\Omega))^d$ and V_0, respectively, we have

$$\int_{\Gamma_1} \frac{\partial\mathbf{u}}{\partial n}\cdot\mathbf{v}d\Gamma = \int_\Omega \boldsymbol{\nabla}\mathbf{u} : \boldsymbol{\nabla}\mathbf{v}\, dx + \int_\Omega \Delta\mathbf{u}\cdot\mathbf{v}\, dx. \tag{93}$$

Taking now the dot-product with $\mathbf{v}$ of both sides of the first equation (88), using (93) and taking the boundary conditions in (88) into account we obtain that if $\mathbf{u}$ is a solution of problem (88) belonging to V_g, it is also a solution of the following *nonlinear variational problem*:

$$\begin{cases} \mathbf{u} \in V_g;\ \forall \mathbf{v} \in V_0 \ we\ have \\ \alpha\displaystyle\int_\Omega \mathbf{u}\cdot\mathbf{v}\, dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{u} : \boldsymbol{\nabla}\mathbf{v}\, dx + \int_\Omega (\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u}\cdot\mathbf{v}\, dx = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, dx + \int_{\Gamma_1} \mathbf{g}_1\cdot\mathbf{v}d\Gamma. \end{cases} \tag{94}$$

Actually, the reciprocal property is true and (94) implies (88). Problem (88), (94) is equivalent to a problem of the *Calculus of Variations*, since neither $(\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v}$ nor $(\mathbf{V}\cdot\boldsymbol{\nabla})\mathbf{v}$ are the differentials of a functional of $\mathbf{v}$; using, however, a convenient *least squares formulation* we shall be able to solve the above advection-diffusion problem by iterative methods from *Nonlinear Programming*, such as *conjugate gradient algorithms*.

3.1 Leastr–Squares Formulation of (88), (94)

Let $\mathbf{v} \in V_g$; to $\mathbf{v}$ we associate the solution $\mathbf{y} = \mathbf{y}(\mathbf{v}) \in V_0$ of

$$\begin{cases} \alpha\mathbf{y} - \nu\Delta\mathbf{y} = \alpha\mathbf{v} - \nu\Delta\mathbf{v} + (\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v} - \mathbf{f} \ in\ \Omega, \\ \mathbf{y} = \mathbf{0}\ on\ \Gamma_0,\ \nu\dfrac{\partial\mathbf{y}}{\partial n} = \nu\dfrac{\partial\mathbf{v}}{\partial n} - \mathbf{g}_1\ on\ \Gamma_1. \end{cases} \tag{95}$$

We observe that $\mathbf{y}$ is obtained from $\mathbf{v}$ via the solution of d uncoupled linear elliptic problems (one for each component of $\mathbf{y}$); using (93), it is easily shown that (95) is equivalent to the linear variational problem

$$\begin{cases} \mathbf{y} \in V_0;\ \forall \mathbf{z} \in V_0\ we\ have \\ \alpha\displaystyle\int_\Omega \mathbf{y}\cdot\mathbf{z}\,dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{y} : \boldsymbol{\nabla}\mathbf{z}\,dx = \alpha\int_\Omega \mathbf{v}\cdot\mathbf{z}\,dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{v} : \boldsymbol{\nabla}\mathbf{z}\,dx \\ \quad + \displaystyle\int_\Omega (\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v}\cdot\mathbf{z}\,dx - \int_\Omega \mathbf{f}\cdot\mathbf{z}\,dx - \int_{\Gamma_1} \mathbf{g}_1\cdot\mathbf{z}d\Gamma, \end{cases} \tag{96}$$

which has a unique solution.

Suppose now that $\mathbf{v}$ is a solution of the nonlinear problem (88), (94); the corresponding $\mathbf{y}$ (obtained from the solution of (95), (96)) is clearly $\mathbf{y} = \mathbf{0}$; from this observation, it is quite natural to introduce the following (nonlinear) least–squares formulation of (88), (94):

$$\begin{cases} find\ \mathbf{u} \in V_g\ such\ that \\ J(\mathbf{u}) \le J(\mathbf{v}),\ \forall \mathbf{v} \in V_g, \end{cases} \tag{97}$$

where the functional $J : (H^1(\Omega))^d \to \mathbb{R}$ is defined by

$$J(\mathbf{v}) = \frac{1}{2}\int_\Omega \{\alpha|\mathbf{y}|^2 + \nu|\boldsymbol{\nabla}\mathbf{y}|^2\}\,dx \tag{98}$$

with $\mathbf{y}$ defined from $\mathbf{v}$ by (95), (96). Observe that if $\mathbf{u}$ is a solution of (97), such that $J(\mathbf{u}) = 0$, then it is also a solution of (88), (94).

3.2 Conjugate Gradient Methods for the Solution of Minimization Problems in Hilbert Spaces.

The main goal of this subsection is to discuss the *iterative solution of minimization problems in Hilbert spaces by conjugate gradient algorithms.* For years, our main sources of information concerning conjugate gradient algorithms have been Daniel 1970 [43] and Polak 1971 [44], the first reference in particular since it is also concerned with infinite dimensional problems.

Conjugate gradient algorithms have been introduced by M. Hestenes and E. Stiefel in the early fifties for the solution of finite dimensional linear systems associated with *symmetric and positive definite* matrices (see Hestenes and Stiefel 1952 [45] for details). Since then, these methods have enjoyed considerable generalizations and have motivated a very large number of

publications. The interested reader may find abundant information on these methods and their implementation in, e.g., the review articles Freund, Golub and Nachtigal 1992 [46], Nocedal 1992 [47] and in the monographs Kelley 1995 [48] (Chapter 2), Saad 1995 [49] (see also the references therein, and Golub and O'Leary 1989 [50] for an historical account).

3.2.1 Conjugate Gradient Solution of Linear Variational Problems in Hilbert Spaces.

We shall discuss first the *conjugate gradient solution* of the *linear variational problems in Hilbert spaces*. We consider:

(i) V is a *real Hilbert space* for the scalar product $(\cdot,\cdot)$ and the associated norm $||\cdot||$;

(ii) $a(\cdot,\cdot)$ is a *bilinear* functional from $V \times V \to \mathbb{R}$, *continuous* and V-*elliptic* (i.e., $\exists\ \alpha > 0$ such that $a(v,v) \geq \alpha ||v||^2,\ \forall v \in V$);

(iii) L is *linear* and *continuous* over V.

In this section we make the following additional assumption on the bilinear functional $a(\cdot,\cdot)$:

$$\begin{cases} \textit{the bilinear functional } a(\cdot,\cdot) \textit{ is symmetric}, \\ \textit{i.e., } a(v,w) = a(w,v),\ \forall v,w \in V. \end{cases} \tag{99}$$

If the symmetry property (99) holds, then the linear variational problem

$$\begin{cases} u \in V, \\ a(u,v) = L(v),\ \forall v \in V, \end{cases} \tag{100}$$

has a *unique* solution by the Lax-Milgram theorem, which is also the solution of the *minimization* problem

$$\begin{cases} u \in V, \\ J(u) \leq J(v),\ \forall v \in V, \end{cases} \tag{101}$$

with

$$J(v) = \frac{1}{2} a(v,v) - L(v),\ \forall v \in V. \tag{102}$$

Here are a typical examples of above linear variational problem:
Example 3.1: Here we consider the one associated with the homogeneous Dirichlet problem:

$$\begin{cases} u \in H_0^1(\Omega), \\ \displaystyle\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, dx = \int_\Omega f v \, dx,\ \forall v \in H_0^1(\Omega), \end{cases}$$

with $f \in L^2(\Omega)$. In this example, we have $V = H_0^1(\Omega)$,

$$a(u,v) = \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, dx, \quad L(v) = \int_\Omega f v \, dx$$

and

$$J(v) = \frac{1}{2}\int_\Omega |\boldsymbol{\nabla} v|^2 \, dx - \int_\Omega f v \, dx.$$

Description of the conjugate gradient algorithm.

In order to solve problem (100), (101) we propose the following conjugate gradient algorithm.

Step 0: Initialization

$$u^0 \in V \text{ is given;} \tag{103}$$

solve

$$\begin{cases} g^0 \in V, \\ (g^0, v) = a(u^0, v) - L(v), \ \forall v \in V, \end{cases} \tag{104}$$

and set

$$w^0 = g^0. \square \tag{105}$$

For $n \geq 0$, assuming that u^n, g^n, w^n are known with $g^n \neq 0$ and $w^n \neq 0$, compute u^{n+1}, g^{n+1}, w^{n+1} as follows

Step 1: Steepest descent

Compute

$$\rho_n = ||g^n||^2 / a(w^n, w^n) \tag{106}$$

and set

$$u^{n+1} = u^n - \rho_n w^n. \tag{107}$$

Step 2: Testing the convergence and construction of the new descent direction

Solve

$$\begin{cases} g^{n+1} \in V, \\ (g^{n+1}, v) = (g^n, v) - \rho_n a(w^n, v), \ \forall v \in V. \end{cases} \tag{108}$$

If $||g^{n+1}||/||g^0|| \leq \varepsilon$ take $u = u^{n+1}$; else, compute

$$\gamma_n = ||g^{n+1}||^2 / ||g^n||^2 \tag{109}$$

and update w^n by

$$w^{n+1} = g^{n+1} + \gamma_n w^n. \tag{110}$$

Do $n = n + 1$ and return to (106). $\square$

Despite its apparent simplicity, algorithm (103)-(110) is *one of the most powerful tools* of Scientific Computing; it is currently used to solve very complicated problems from Science and Engineering which may involve many millions of unknowns. Large scale application of the above algorithm will be found in several parts of this article.

Convergence of algorithm (103)-(110).

Before discussing the convergence of algorithm (103)-(110), it can be shown by using the Riesz theorem that problem (100), (101) is equivalent to

$$Au = l, \tag{111}$$

where $l \in V, A \in \mathcal{L}(V, V)$ and verify

$$L(v) = (l, v),\ \forall v \in V \text{ and } a(v, w) = (Av, w),\ \forall v, w \in V;$$

operator A is an *automorphism* of V (*symmetric* since $a(\cdot, \cdot)$ is symmetric). Incidentally, we have

$$\alpha||v||^2 \leq a(v, v) \leq ||A||||v||^2,\ \forall v \in V; \tag{112}$$

in (112), the best constant α (i.e., the largest one) is given by $1/||A^{-1}||$.

Concerning the *convergence* of algorithm (103)-(110), we are going to prove the following:

Theorem 1 *Suppose that $\varepsilon = 0$ in algorithm (103)-(110); we have then*

$$\lim_{n \to +\infty} ||u^n - u|| = 0,\ \forall u^0 \in V, \tag{113}$$

where u is the solution of problem (100), (101).

PROOF: For clarity, the proof has been divided in two parts.
Orthogonality properties: First, we are going to show that the following *orthogonality* properties hold, as long as we can iterate (i.e., as long as g^n and w^n are different from 0 in (103)-(110)):

$$(g^i, g^j) = 0,\ \forall i, j, i \neq j, \tag{114}$$

$$(g^i, w^j) = 0,\ \forall i, j, i > j, \tag{115}$$

$$a(w^i, w^j) = 0,\ \forall i, j, i \neq j. \tag{116}$$

We are going to proceed by *induction*, assuming first that relations (114)-(116) hold up to n; let us show that they also hold up to $n + 1$. We start with (114):

We have, from (108) and from (110) (with n replaced by $n - 1$)

$$\begin{aligned} (g^{n+1}, g^n) &= ||g^n||^2 - \rho_n a(w^n, g^n) \\ &= ||g^n||^2 - \rho_n a(w^n, w^n - \gamma_{n-1} w^{n-1}); \end{aligned}$$

using (116) (true up to n) and (106) we obtain

$$(g^{n+1}, g^n) = ||g^n||^2 - \rho_n a(w^n, w^n) = 0.$$

Similarly, we have for $j < n$

$$\begin{cases} (g^{n+1}, g^j) = (g^n, g^j) - \rho_n a(w^n, g^j) \\ \qquad\qquad = (g^n, g^j) - \rho_n a(w^n, w^j - \gamma_{j-1} w^{j-1}) = 0. \end{cases}$$

We have thus shown that if (114) holds up to n, it also holds up to $n+1$. $\square$

We consider now the relations (115); operating as above we have

$$\begin{aligned}(g^{n+1},w^n) &= (g^n,w^n)-\rho_n a(w^n,w^n)\\ &= (g^n,g^n+\gamma_{n-1}w^{n-1})-\rho_n a(w^n,w^n)\\ &= ||g^n||^2-\rho_n a(w^n,w^n)=0,\end{aligned}$$

and for $j<n$

$$(g^{n+1},w^j)=(g^n,w^j)-\rho_n a(w^n,w^j)=0.$$

We have shown, here also, that if (115) holds up to n, it holds up to $n+1$. $\square$

Proving similar results for (116) is slightly more complicated; however, using the relations in algorithm (103)-(110) and the fact that (114), (115) (resp., (116)) hold up to $n+1$ (resp., n) we have

$$\begin{aligned}a(w^{n+1},w^n) &= a(w^n,w^{n+1})=\rho_n^{-1}[(g^n,w^{n+1})-(g^{n+1},w^{n+1})]\\ &= \rho_n^{-1}[(g^n,g^{n+1}+\gamma_n w^n)-(g^{n+1},g^{n+1}+\gamma_n w^n)]\\ &= \rho_n^{-1}[\gamma_n(g^n,w^n)-||g^{n+1}||^2]\\ &= \rho_n^{-1}[\gamma_n(g^n,g^n+\gamma_{n-1}w^{n-1})-||g^{n+1}||^2]\\ &= \rho_n^{-1}[\gamma_n||g^n||^2-||g^{n+1}||^2]=0,\end{aligned}$$

and then for $j<n$

$$\begin{aligned}a(w^{n+1},w^j) &= a(g^{n+1}+\gamma_n w^n,w^j)=a(g^{n+1},w^j)\\ &= a(w^j,g^{n+1})\\ &= \rho_j^{-1}[(g^j,g^{n+1})-(g^{j+1},g^{n+1})]=0;\end{aligned}$$

the above relations imply that (116) hold up to $n+1$ if it holds up to n. $\square$

To complete the proof of (114)-(116) it suffices to show that these relations also hold for $i=1$ and $j=0$. Using the fact that $w^0=g^0$, we have

$$\begin{aligned}(g^1,g^0) &= ||g^0||^2-\rho_0 a(w^0,g^0)=||g^0||^2-\rho_0 a(w^0,w^0)=0,\\ (g^1,w^0) &= 0.\end{aligned}$$

Concerning now $a(w^1,w^0)$, we have

$$\begin{aligned}a(w^1,w^0) &= a(w^0,w^1)=\rho_0^{-1}[(g^0,w^1)-(g^1,w^1)]\\ &= \rho_0^{-1}[(g^0,g^1+\gamma_0 w^0)-(g^1,g^1+\gamma_0 w^0)]\\ &= \rho_0^{-1}[\gamma_0(g^0,w^0)-||g^1||^2]\\ &= \rho_0^{-1}[\gamma_0||g^0||^2-||g^1||^2]=0,\end{aligned}$$

which completes the proof of relations (114)-(116). $\square$

Convergence: We can easily show (by induction, again) that

$$(g^{n+1},v)=a(u^{n+1},v)-L(v),\ \forall v\in V.$$

If $g^{n+1}=0$ in algorithm (103)-(110), we have therefore $u^{n+1}=u$ (since problem (100) has a unique solution). Suppose now that $w^{n+1}=0$; it follows from (110) that

$$g^{n+1}+\gamma_n w^n=0$$

which implies in turn (from (115)) that

$$||g^{n+1}||^2 + \gamma_n(g^{n+1}, w^n) = ||g^{n+1}||^2 = 0;$$

we have thus $u^{n+1} = u$.

Suppose now that we have $g^n \neq 0$ and $w^n \neq 0,\ \forall n \geq 0$; in order to show that $\lim_{n\to+\infty} u^n = u$ we consider the difference $J(u^n) - J(u^{n+1})$; we clearly have (Taylor's expansion)

$$\begin{aligned} J(u^{n+1}) &= J(u^n - \rho_n w^n) = J(u^n) - \rho_n(J'(u^n), w^n) + \tfrac{1}{2}\rho_n^2 a(w^n, w^n) \\ &= J(u^n) - \rho_n[a(u^n, w^n) - L(w^n)] + \tfrac{1}{2}\rho_n^2 a(w^n, w^n) \\ &= J(u^n) - \rho_n(g^n, w^n) + \tfrac{1}{2}\rho_n^2 a(w^n, w^n) \\ &= J(u^n) - \rho_n(g^n, g^n + \gamma_{n-1}w^{n-1}) + \tfrac{1}{2}\rho_n^2 a(w^n, w^n) \\ &= J(u^n) - \rho_n||g^n||^2 + \tfrac{1}{2}\rho_n^2 a(w^n, w^n) \end{aligned}$$

which implies that

$$J(u^n) - J(u^{n+1}) = \rho_n||g^n||^2 - \frac{1}{2}\rho_n^2 a(w^n, w^n) = \frac{1}{2}||g^n||^4/a(w^n, w^n),\ \forall n \geq 0. \tag{117}$$

It follows from (117) that the sequence $\{J(u^n)\}_{n\geq 0}$ is a decreasing one; since it is bounded from below by $J(u)$, it converges to some limit ($\geq J(u)$) which implies that

$$\lim_{n\to+\infty}[J(u^n) - J(u^{n+1})] = 0.$$

We have thus shown (from (117)) that

$$\lim_{n\to+\infty} ||g^n||^4/a(w^n, w^n) = 0. \tag{118}$$

Since $g^n = w^n - \gamma_{n-1}w^{n-1}$, we have (from (116)) that

$$a(g^n, g^n) = a(w^n, w^n) + \gamma_{n-1}^2 a(w^{n-1}, w^{n-1}) \geq a(w^n, w^n) > 0; \tag{119}$$

we also have, from (112),

$$a(g^n, g^n) \leq ||A||||g^n||^2. \tag{120}$$

Combining (118), (119), (120) yields $\lim_{n\to+\infty} ||g^n|| = 0$, which implies in turn (since $g^n = Au^n - l,\ \forall n \geq 0$)

$$\lim_{n\to+\infty} u^n = A^{-1}l = u,$$

which completes the proof of the theorem.

Remark 10 *Suppose that V is finite dimensional with $dimV = d$; in that case we have convergence in d iterations at most. Suppose that it is not the case, then $\{g^0, g^1, \ldots g^d\}$ will be a system of $d+1$ vectors of V, linearly independent since all different from zero and mutually orthogonal (from (114)). Since this is impossible there exists $N \leq d$ such that $g^N = 0$, which implies in turn that $u^N = u$.*

Remark 11 *The above proof of Theorem 1 is a variant of the classical one used to prove, in finite dimension, the finite termination property discussed in Remark 10; these proofs completely rely on the orthogonality properties (114)-(116). Computer implementations (necessarily finite-dimensional) of algorithm (103)-(110) will suffer from the effects of round-off errors, one of the effects being precisely the loss of the above orthogonality properties; we can wonder, therefore, about the convergence properties of algorithm (103)-(110) in practice. Actually they are quite good, in general, despite the fact that the finite termination is lost, strictly speaking. This good behavior of algorithm (103)-(110) is a direct consequence of the following estimate of its speed of convergence (proved in, e.g., Daniel 1970 [43]):*

$$a(u^n - u, u^n - u) \leq 4a(u^0 - u, u^0 - u)\left(\frac{\sqrt{\nu_a} - 1}{\sqrt{\nu_a} + 1}\right)^{2n}, \ \forall n \geq 1, \tag{121}$$

where, in (121), the condition number ν_a of the bilinear functional $a(\cdot, \cdot)$ is defined by

$$\nu_a = \sup_{v \in S} a(v, v) / \inf_{v \in S} a(v, v), \tag{122}$$

with $S = \{v | v \in V, ||v|| = 1\}$ (we can easily show that $\nu_a = ||A||||A^{-1}||$, operator A being this element of $\mathcal{L}(V, V)$ such that $a(v, w) = (Av, w), \ \forall v, w \in V$). We observe that the closer ν_a is to 1, the faster is the speed of convergence. For problems of large dimension the convergence behavior associated with (121) is much more important than the hypothetical finite termination property mentioned above.

Using the following *equivalence* relations between the norms $||v||$ and $\sqrt{a(v, v)}$

$$||A^{-1}||^{-1}||v||^2 \leq a(v, v) \leq ||A||||v||^2, \ \forall v \in V,$$

we can easily show that (121) implies

$$||u^n - u|| \leq 2\sqrt{\nu_a}\left(\frac{\sqrt{\nu_a} - 1}{\sqrt{\nu_a} + 1}\right)^n ||u^0 - u||, \ \forall n \geq 1, \tag{123}$$

which is less sharp than (121).

3.2.2 Conjugate Gradient Methods for the Solution of Minimization Problems in Hilbert Spaces.

Formulation of the Minimization Problems.

The *minimization problems* to be considered have the following formulation:

$$\begin{cases} u \in V, \\ J(u) \leq J(v), \ \forall v \in V, \end{cases} \tag{124}$$

where:

• V is a Hilbert space for the scalar product $(\cdot,\cdot)$ and the corresponding norm $||\cdot||$; we do not assume, here, that V has been identified to its dual space V'.

• $J : V \to \mathbb{R}$ is a *differentiable* functional whose differential is denoted by J' (some authors use the notation $\boldsymbol{\nabla} J$ for the differential of J).

Since V has not been necessarily identified to V', it is convenient to introduce the *duality isomorphism* $S : V \to V'$, which is the *unique* operator in *Isom* (V, V') such that

$$< Sv, w >=< Sw, v >= (v, w),\ \forall v, w \in V, \tag{125}$$

where $< \cdot,\cdot >$ denotes the *duality pairing* between V' and V; operator S is *self-adjoint* and *strongly-elliptic* over V since (125) implies

$$< Sv, v >= ||v||^2,\ \forall v \in V. \tag{126}$$

Actually, in addition to (126), relation (125) implies

$$||f||_*^2 =< f, S^{-1}f >,\ \forall f \in V' \tag{127}$$

(where the dual norm $||\cdot||_*$ is defined - classically - by

$$||f||_* = \sup_{v\in\Sigma} | < f, v > |\ with\ \Sigma = \{v | v \in V, ||v|| = 1\}),$$

and

$$(f, g)_* =< f, S^{-1}g >,\ \forall f, g \in V', \tag{128}$$

where $(\cdot,\cdot)_*$ denotes the scalar product in V', compatible with the norm $||\cdot||_*$.

Concerning now the *differentiability* of J, we shall assume that J is either *Fréchet-differentiable* or *Gâteaux-differentiable*. We recall (see, e.g., Zeidler 1986 [51] (Chapter 4)) that J is *Fréchet-differentiable* over V if, $\forall v \in V$, there exists $J'(v) \in V'$, the derivative of J at v, such that

$$J(v + w) - J(v) =< J'(v), w > +||w||\varepsilon(v, w), \tag{129}$$

with $\lim_{w\to 0} \varepsilon(v, w) = 0$. Similarly, (see again Zeidler 1986 [51], loc. cit.), J is *Gâteaux-differentiable* over V, if, $\forall v, w \in V$, there exists $J'(v) \in V'$ such that

$$J(v + tw) - J(v) = t < J'(v), w > +t\varepsilon(t, v, w). \tag{130}$$

with $\lim_{t\to 0} \varepsilon(t, v, w,) = 0$. It is quite obvious that the Fréchet-differentiability of J implies its continuity and its Gâteaux-differentiability.

Back to (124), if we suppose that the minimization problem has a solution u, it *necessarily* verifies

$$J'(u) = 0. \tag{131}$$

Proving (131) is fairly obvious, but owing to the importance of this result, we feel obliged to prove it. Observe, therefore, that (124) implies

$$\frac{J(u + tv) - J(u)}{t} \geq 0,\ \forall v \in V\ and\ \ \forall t > 0; \tag{132}$$

taking the limit in (132), as $t \to 0_+$, we obtain, from (130),

$$< J'(u), v >\geq 0, \forall v \in V,$$

which clearly implies (replace v by $-v$)

$$< J'(u), v >= 0, \forall v \in V. \tag{133}$$

Finally, to show (131) take $v = S^{-1}J'(u)$ in (133) and use relation (127).

As already mentioned, the optimal condition (131) is *sufficient* if J is *convex*; furthermore if J is *strictly convex*, i.e.

$$\begin{cases} J(tv + (1-t)w) < tJ(v) + (1-t)J(w), \\ \forall t \in (0,1),\ \forall v, w \in V, v \neq w, \end{cases} \tag{134}$$

then *existence implies uniqueness.*

We shall conclude this section by mentioning typical conditions which imply the existence of a solution to the minimization problem (124); these conditions are

$$\lim_{||v|| \to +\infty} J(v) = +\infty, \tag{135}$$

$$J \text{ is weakly lower semi} - \text{continuous over } V; \tag{136}$$

condition (136) means that:

$$If \lim_{n \to +\infty} v_n = v \text{ weakly in } V, \text{ then } \lim_{n \to +\infty} \inf\ J(v_n) \geq J(v).$$

Showing that (135), (136) implies the existence of a solution to problem (124) is fairly easy.

Remark 12 *If J is convex and differentiable over V, condition (136) is automatically satisfied. To show this result we observe that from the convexity of J we have (by definition)*

$$J((1-t)v + tw) \leq tJ(w) + (1-t)J(v),\ \forall v, w \in V,\ \forall t \in (0,1],$$

which can be rewritten as

$$\frac{J(v + t(w-v)) - J(v)}{t} \leq J(w) - J(v),\ \forall v, w \in V,\ \forall t \in (0,1]. \tag{137}$$

Taking the limit in (137), as $t \to 0_+$ we obtain (from (130))

$$J(w) - J(v) \geq < J'(v), w - v >,\ \forall v, w \in V. \tag{138}$$

Condition (138) is in fact a celebrated characterization of the convexity of differentiable functionals (as shown in, e.g., Ekeland and Temam 1976 [52]). Consider now a sequence $\{v_n\}_{n \geq 0}$ in V such that $\lim_{n \to +\infty} v_n = v$ weakly in V; we have, from (138),

$$J(v_n) - J(v) \geq < J'(v), v_n - v >,\ \forall n \geq 0,$$

which implies at the limit, as $n \to +\infty$,

$$\lim_{n\to+\infty} \inf J(v_n) \geq J(v),$$

which shows the weak lower semi-continuity of J.

Description of Conjugate Gradient Algorithm for the Solution of Problem (124).

In order to solve problem (124) we shall use the following conjugate gradient type algorithms:

Step 0: Initialization

$$u^0 \in V \text{ is given;} \tag{139}$$

solve

$$\begin{cases} g^0 \in V, \\ (g^0, v) =< J'(u^0), v >, \forall v \in V, \end{cases} \tag{140}$$

and set

$$w^0 = g^0. \square \tag{141}$$

Then for $n \geq 0$, *assuming that* u^n, g^n, w^n *are known, compute* $u^{n+1}, g^{n+1}, w^{n+1}$ *as follows*:

Step 1: Steepest Descent

Solve

$$\begin{cases} \rho_n \in \mathbb{R}, \\ J(u^n - \rho_n w^n) \leq J(u^n - \rho w^n), \ \forall \rho \in \mathbb{R} \end{cases} \tag{142}$$

and set

$$u^{n+1} = u^n - \rho_n w^n. \tag{143}$$

Step 2: Testing the convergence and construction of the new descent direction

Solve

$$\begin{cases} g^{n+1} \in V, \\ (g^{n+1}, v) =< J'(u^{n+1}), v >, \forall v \in V; \end{cases} \tag{144}$$

if $||g^{n+1}||/||g^0|| \leq \varepsilon$ *take* $u = u^{n+1}$; *else, compute either*

$$\gamma_n = ||g^{n+1}||^2/||g^n||^2 \ (Fletcher - Reeves \ update) \tag{145}$$

or

$$\gamma_n = (g^{n+1} - g^n, g^{n+1})/||g^n||^2 \ (Polak - Ribière \ update) \tag{146}$$

and then

$$w^{n+1} = g^{n+1} + \gamma_n w^n. \tag{147}$$

Do $n = n + 1$ *and return to (142).* $\square$

Remark 13 *Suppose that the functional J in (124) is given by (102). We can easily show that*

$$< J'(v), w >= a(v, w) - L(w), \ \forall v, w \in V \tag{148}$$

and that algorithm (139)-(147) applied to the minimization of J yields

$$\rho_n = (g^n, w^n)/a(w^n, w^n) \text{ in (142).} \tag{149}$$

Consider now algorithm (103)-(110): the orthogonality conditions (114)-(116) imply that

$$\rho_n = ||g^n||^2/a(w^n, w^n) = (g^n, w^n)/a(w^n, w^n) \text{ in (106),} \tag{150}$$

$$\gamma_n = ||g^{n+1}||^2/||g^n||^2 = (g^{n+1}, g^{n+1} - g^n)/||g^n||^2 \text{ in (109).} \tag{151}$$

It follows from (148)-(151) that algorithms (103)-(110) and (139)-(147) coincide if J is given by (102). □

The convergence properties of the Fletcher-Reeves and Polak-Ribière conjugate gradient algorithms have inspired many investigators; let us mention among others Daniel 1970 [43], Ortega and Rheinboldt 1970 and 1972 [53, 54, 55], Polak 1971 [44] (Chapter 6), Avriel 1976 [56] (Chapter 10), Powell 1976 and 1977 [57, 58], Girault and Raviart 1986 [59] (Chapter 4) and also two recent references, namely Nocedal 1992 [47] and Hiriart-Urruty and Lemarechal 1993 [60] (Chapter 2). We found the last two references particularly interesting, since they contain a large number of further references on conjugate gradient algorithms, and also very detailed advices and recipes on the practical implementation of these algorithms, based on three decades of theoretical investigations and computer experiments.

3.2.3 Application to the advection-diffusion problem (88).

In order to solve problem (88) by the least-squares/conjugate gradient techniques discussed in previous subsections, we need to equip V_0 and V_g with an appropriate Hilbertian structure; we chose as scalar product on V_0 and V_g

$$\{\mathbf{v}, \mathbf{w}\} \rightarrow \int_\Omega (\alpha \mathbf{v} \cdot \mathbf{w} + \nu \boldsymbol{\nabla} \mathbf{v} : \boldsymbol{\nabla} \mathbf{w}) \, dx,$$

the corresponding norm being, obviously,

$$\mathbf{v} \rightarrow (\int_\Omega (\alpha |\mathbf{v}|^2 + \nu |\boldsymbol{\nabla} \mathbf{v}|^2) \, dx)^{\frac{1}{2}}.$$

To apply the Fletcher-Reeves algorithm (139)-(147) to the solution of the problem (88), (97) we need, in principle, to take as unknown $\tilde{\mathbf{u}} = \mathbf{u} - \tilde{\mathbf{g}}_0$ with some $\tilde{\mathbf{g}}_0 \in V_g$ so that $\mathbf{g}_0 = \tilde{\mathbf{g}}_0|_{\Gamma_0}$, in order to transform the problem (88),

(97) into equivalent problems in V_0; actually, this is not necessary and we can proceed directly with algorithm (139)-(147). We obtain then:

$$\mathbf{u}^0 \in V_g \text{ is given;} \tag{152}$$

solve

$$\begin{cases} \mathbf{g}^0 \in V_0, \\ \displaystyle\int_\Omega (\alpha \mathbf{g}^0 \cdot \mathbf{z} + \nu \boldsymbol{\nabla} \mathbf{g}^0 : \boldsymbol{\nabla} \mathbf{z}) dx =< J'(\mathbf{u}^0), \mathbf{z} >, \quad \forall \mathbf{z} \in V_0, \end{cases} \tag{153}$$

and set

$$\mathbf{w}^0 = \mathbf{g}^0. \tag{154}$$

For $n \geq 0$, assuming that $\mathbf{u}^n, \mathbf{g}^n, \mathbf{w}^n$ are known, we obtain $\mathbf{u}^{n+1}, \mathbf{g}^{n+1}, \mathbf{w}^{n+1}$ by:
Step 1: Steepest Descent
Solve

$$\begin{cases} \rho_n \in \mathbb{R}, \\ J(\mathbf{u}^n - \rho_n \mathbf{w}^n) \leq J(\mathbf{u}^n - \rho \mathbf{w}^n), \ \forall \rho \in \mathbb{R} \end{cases} \tag{155}$$

and set

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \rho_n \mathbf{w}^n. \tag{156}$$

Step 2: Testing the convergence and construction of the new descent direction
Solve

$$\begin{cases} \mathbf{g}^{n+1} \in V_0, \\ \displaystyle\int_\Omega (\alpha \mathbf{g}^{n+1} \cdot \mathbf{z} + \nu \boldsymbol{\nabla} \mathbf{g}^{n+1} : \boldsymbol{\nabla} \mathbf{z}) dx =< J'(\mathbf{u}^{n+1}), \mathbf{z} >, \quad \forall \mathbf{z} \in V_0, \end{cases} \tag{157}$$

if $\displaystyle\int_\Omega (\alpha |\mathbf{g}^{n+1}|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^{n+1}|^2)\, dx / \int_\Omega (\alpha |\mathbf{g}^0|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^0|^2)\, dx \leq \varepsilon^2$, *take* $\mathbf{u} = \mathbf{u}^{n+1}$;
else, compute

$$\gamma_n = \int_\Omega (\alpha |\mathbf{g}^{n+1}|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^{n+1}|^2)\, dx / \int_\Omega (\alpha |\mathbf{g}^n|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^n|^2)\, dx \tag{158}$$

and then

$$\mathbf{w}^{n+1} = \mathbf{g}^{n+1} + \gamma_n \mathbf{w}^n. \tag{159}$$

Do $n = n + 1$ and return to (155). □

Calculations of J' and ρ_n when solving the non-linear problem (88), (97).

To compute $J'(\mathbf{u}^n)$ at each iteration in above algorithm (152)-(159) when solving the non-linear problem (88), (97), again let us follow the definition (130).

For $\mathbf{v} \in V_g$ and $\mathbf{w} \in V_0$, we have

$$\begin{cases} \mathbf{y}(t) \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \alpha \displaystyle\int_\Omega \mathbf{y}(t) \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{y}(t) : \boldsymbol{\nabla}\mathbf{z}\, dx \\ \quad = \alpha \displaystyle\int_\Omega (\mathbf{v} + t\mathbf{w}) \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla}(\mathbf{v} + t\mathbf{w}) : \boldsymbol{\nabla}\mathbf{z}\, dx \\ \quad + \displaystyle\int_\Omega ((\mathbf{v} + t\mathbf{w}) \cdot \boldsymbol{\nabla})(\mathbf{v} + t\mathbf{w}) \cdot \mathbf{z}\, dx - \int_\Omega \mathbf{f} \cdot \mathbf{z}\, dx - \int_{\Gamma_1} \mathbf{g}_1 \cdot \mathbf{z}\, d\Gamma, \end{cases} \tag{160}$$

and

$$\begin{cases} \mathbf{y} \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \alpha \displaystyle\int_\Omega \mathbf{y} \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{y} : \boldsymbol{\nabla}\mathbf{z}\, dx = \alpha \int_\Omega \mathbf{v} \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{v} : \boldsymbol{\nabla}\mathbf{z}\, dx \\ \quad + \displaystyle\int_\Omega (\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{v} \cdot \mathbf{z}\, dx - \int_\Omega \mathbf{f} \cdot \mathbf{z}\, dx - \int_{\Gamma_1} \mathbf{g}_1 \cdot \mathbf{z}\, d\Gamma. \end{cases} \tag{161}$$

Clearly we have $\mathbf{y}(t) = \mathbf{y} + t\ \delta\mathbf{y}(t)$ where $\delta\mathbf{y}(t)$ is the solution of

$$\begin{cases} \delta\mathbf{y}(t) \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \alpha \displaystyle\int_\Omega \delta\mathbf{y}(t) \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla}\delta\mathbf{y}(t) : \boldsymbol{\nabla}\mathbf{z}\, dx = \alpha \int_\Omega \mathbf{w} \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{w} : \boldsymbol{\nabla}\mathbf{z}\, dx \\ \quad + \displaystyle\int_\Omega (\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{w} \cdot \mathbf{z}\, dx + \int_\Omega (\mathbf{w} \cdot \boldsymbol{\nabla})\mathbf{v} \cdot \mathbf{z}\, dx + t \int_\Omega (\mathbf{w} \cdot \boldsymbol{\nabla})\mathbf{w} \cdot \mathbf{z}\, dx \end{cases} \tag{162}$$

Hence we have the difference

$$\begin{aligned} J(\mathbf{v} + t\mathbf{w}) - J(\mathbf{v}) &= \frac{1}{2}\int_\Omega (\alpha|\mathbf{y} + t\ \delta\mathbf{y}(t)|^2 + \nu|\boldsymbol{\nabla}(\mathbf{y} + t\ \delta\mathbf{y}(t))|^2)\, dx \\ &\qquad -\frac{1}{2}\int_\Omega (\alpha|\mathbf{y}|^2 + \nu|\boldsymbol{\nabla}\mathbf{y}|^2)\, dx \\ &= t\int_\Omega (\alpha\mathbf{y} \cdot \delta\mathbf{y}(t) + \nu\boldsymbol{\nabla}\mathbf{y} : \boldsymbol{\nabla}\delta\mathbf{y}(t))\, dx \\ &\qquad +\frac{t^2}{2}\int_\Omega (\alpha|\delta\mathbf{y}(t)|^2 + \nu|\boldsymbol{\nabla}\delta\mathbf{y}(t)|^2)\, dx \end{aligned} \tag{163}$$

and

$$< J'(\mathbf{v}), \mathbf{w} >= \lim_{t\to 0} \frac{J(\mathbf{v} + t\mathbf{w}) - J(\mathbf{v})}{t} = \int_\Omega (\alpha\mathbf{y} \cdot \delta\mathbf{y}(0) + \nu\boldsymbol{\nabla}\mathbf{y} : \boldsymbol{\nabla}\delta\mathbf{y}(0))\, dx. \tag{164}$$

Since $\mathbf{y} \in V_0$, we set $\mathbf{z} = \mathbf{y}$ and $t = 0$ in (162) and obtain

$$< J'(\mathbf{v}), \mathbf{w} >= \alpha\int_\Omega \mathbf{y}\cdot\mathbf{w}\, dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{y} : \boldsymbol{\nabla}\mathbf{w}\, dx + \int_\Omega (\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{w}\cdot\mathbf{y}\, dx + \int_\Omega (\mathbf{w}\cdot\boldsymbol{\nabla})\mathbf{v}\cdot\mathbf{y}\, dx. \tag{165}$$

Therefore $< J'(\mathbf{v}), \mathbf{w} >$ has a purely integral representation, which is of major importance in view of finite element implementation of algorithm (152)-(159).

Another problem of practical importance is the calculation of ρ_n in (155) when solving the non-linear problem (88), (97). Let $\mathbf{y}^n(\rho)$ be the solution of (95), (96) with given $v = \mathbf{u}^n - \rho\mathbf{w}^n$, then we have $\mathbf{y}^n(0) = \mathbf{y}^n$ and $\mathbf{y}^n(\rho_n) = \mathbf{y}^{n+1}$ and

$$\mathbf{y}^n(\rho) = \mathbf{y}^n - \rho\mathbf{y}_1^n + \rho^2\mathbf{y}_2^n, \tag{166}$$

where $\mathbf{y}_1^n$ and $\mathbf{y}_2^n$ are the solution of

$$\begin{cases} \mathbf{y}_1^n \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha\mathbf{y}_1^n \cdot \mathbf{z} + \nu\boldsymbol{\nabla}\mathbf{y}_1^n : \boldsymbol{\nabla}\mathbf{z})dx = \alpha\int_\Omega \mathbf{w}^n \cdot \mathbf{z}\,dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{w}^n : \boldsymbol{\nabla}\mathbf{z}\,dx \\ \quad + \displaystyle\int_\Omega (\mathbf{u}^n \cdot \boldsymbol{\nabla})\mathbf{w}^n \cdot \mathbf{z}\,dx + \int_\Omega (\mathbf{w}^n \cdot \boldsymbol{\nabla})\mathbf{u}^n \cdot \mathbf{z}\,dx, \end{cases} \tag{167}$$

$$\begin{cases} \mathbf{y}_2^n \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha\mathbf{y}_2^n \cdot \mathbf{z} + \nu\boldsymbol{\nabla}\mathbf{y}_2^n : \boldsymbol{\nabla}\mathbf{z})dx = \int_\Omega (\mathbf{w}^n \cdot \boldsymbol{\nabla})\mathbf{w}^n \cdot \mathbf{z}\,dx, \end{cases} \tag{168}$$

respectively. Since

$$J(\mathbf{u}^n - \rho\mathbf{w}^n) = \frac{1}{2}\int_\Omega \{\alpha|\mathbf{y}^n(\rho)|^2 + \nu|\boldsymbol{\nabla}\mathbf{y}^n(\rho)|^2\}\,dx, \tag{169}$$

the function $j_n(\rho) = J(\mathbf{u}^n - \rho\mathbf{w}^n)$ is a *quartic polynomial* in ρ; ρ_n is therefore a solution of the *cubic* equation

$$j_n{}'(\rho) = 0. \tag{170}$$

We shall use the standard *Newton's method* to compute ρ_n from (170), starting from $\rho = 0$. The resulting algorithm is as follows:

$$\rho^0 = 0, \tag{171}$$

and for $k \geq 0, \rho^k$ *being known,*

$$\rho^{k+1} = \rho^k - j_n{}'(\rho^k)/j_n{}''(\rho^k). \tag{172}$$

Conjugate gradient algorithm for the non-linear problem (88), (97).

From above results, the algorithm (152)-(159) can be written in details as follows:

$$\mathbf{u}^0 \in V_g \text{ is given}; \tag{173}$$

solve

$$\begin{cases} \mathbf{y}^0 \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \alpha\displaystyle\int_\Omega \mathbf{y}^0 \cdot \mathbf{z}\,dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{y}^0 : \boldsymbol{\nabla}\mathbf{z}\,dx = \alpha\int_\Omega \mathbf{u}^0 \cdot \mathbf{z}\,dx + \nu\int_\Omega \boldsymbol{\nabla}\mathbf{u}^0 : \boldsymbol{\nabla}\mathbf{z}\,dx \\ \quad + \displaystyle\int_\Omega (\mathbf{u}^0 \cdot \boldsymbol{\nabla})\mathbf{u}^0 \cdot \mathbf{z}\,dx - \int_\Omega \mathbf{f} \cdot \mathbf{z}\,dx - \int_{\Gamma_1} \mathbf{g}_1 \cdot \mathbf{z}\,d\Gamma, \end{cases} \tag{174}$$

and

$$\begin{cases} \mathbf{g}^0 \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha \mathbf{g}^0 \cdot \mathbf{z} + \nu \boldsymbol{\nabla} \mathbf{g}^0 : \boldsymbol{\nabla} \mathbf{z}) dx = \alpha \int_\Omega \mathbf{y}^0 \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{y}^0 : \boldsymbol{\nabla} \mathbf{z}\, dx \\ \quad + \displaystyle\int_\Omega (\mathbf{u}^0 \cdot \boldsymbol{\nabla}) \mathbf{z} \cdot \mathbf{y}^0\, dx + \int_\Omega (\mathbf{z} \cdot \boldsymbol{\nabla}) \mathbf{u}^0 \cdot \mathbf{y}^0\, dx; \end{cases} \tag{175}$$

and set

$$\mathbf{w}^0 = \mathbf{g}^0. \tag{176}$$

For $n \geq 0$, *assuming that* $\mathbf{u}^n$, $\mathbf{y}^n$, $\mathbf{g}^n$, $\mathbf{w}^n$ *are known, we obtain* $\mathbf{u}^{n+1}$, $\mathbf{y}^{n+1}$, $\mathbf{g}^{n+1}$, $\mathbf{w}^{n+1}$ *by*:
Solve

$$\begin{cases} \mathbf{y}_1^n \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha \mathbf{y}_1^n \cdot \mathbf{z} + \nu \boldsymbol{\nabla} \mathbf{y}_1^n : \boldsymbol{\nabla} \mathbf{z}) dx = \alpha \int_\Omega \mathbf{w}^n \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{w}^n : \boldsymbol{\nabla} \mathbf{z}\, dx \\ \quad + \displaystyle\int_\Omega (\mathbf{u}^n \cdot \boldsymbol{\nabla}) \mathbf{w}^n \cdot \mathbf{z}\, dx + \int_\Omega (\mathbf{w}^n \cdot \boldsymbol{\nabla}) \mathbf{u}^n \cdot \mathbf{z}\, dx, \end{cases} \tag{177}$$

$$\begin{cases} \mathbf{y}_2^n \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha \mathbf{y}_2^n \cdot \mathbf{z} + \nu \boldsymbol{\nabla} \mathbf{y}_2^n : \boldsymbol{\nabla} \mathbf{z}) dx = \int_\Omega (\mathbf{w}^n \cdot \boldsymbol{\nabla}) \mathbf{w}^n \cdot \mathbf{z}\, dx. \end{cases} \tag{178}$$

Define

$$\mathbf{y}^n(\rho) = \mathbf{y}^n - \rho \mathbf{y}_1^n + \rho^2 \mathbf{y}_2^n, \tag{179}$$

$$j_n(\rho) = \frac{1}{2} \int_\Omega \{\alpha |\mathbf{y}^n(\rho)|^2 + \nu |\boldsymbol{\nabla} \mathbf{y}^n(\rho)|^2\}\, dx, \tag{180}$$

and solve the cubic equation

$$j_n^{'}(\rho_n) = 0; \tag{181}$$

we have then

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \rho_n \mathbf{w}^n, \tag{182}$$

$$\mathbf{y}^{n+1} = \mathbf{y}^n(\rho_n). \tag{183}$$

Solve

$$\begin{cases} \mathbf{g}^{n+1} \in V_0;\ \forall \mathbf{z} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha \mathbf{g}^{n+1} \cdot \mathbf{z} + \nu \boldsymbol{\nabla} \mathbf{g}^{n+1} : \boldsymbol{\nabla} \mathbf{z}) dx = \alpha \int_\Omega \mathbf{y}^{n+1} \cdot \mathbf{z}\, dx + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{y}^{n+1} : \boldsymbol{\nabla} \mathbf{z}\, dx \\ \quad + \displaystyle\int_\Omega (\mathbf{u}^{n+1} \cdot \boldsymbol{\nabla}) \mathbf{z} \cdot \mathbf{y}^{n+1}\, dx + \int_\Omega (\mathbf{z} \cdot \boldsymbol{\nabla}) \mathbf{u}^{n+1} \cdot \mathbf{y}^{n+1}\, dx. \end{cases} \tag{184}$$

If $\displaystyle\int_\Omega (\alpha |\mathbf{g}^{n+1}|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^{n+1}|^2)\, dx / \int_\Omega (\alpha |\mathbf{g}^0|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^0|^2)\, dx \leq \varepsilon^2$, *take* $\mathbf{u} = \mathbf{u}^{n+1}$; *else, compute*

$$\gamma_n = \int_\Omega (\alpha |\mathbf{g}^{n+1}|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^{n+1}|^2)\, dx / \int_\Omega (\alpha |\mathbf{g}^n|^2 + \nu |\boldsymbol{\nabla} \mathbf{g}^n|^2)\, dx \tag{185}$$

and then

$$\mathbf{w}^{n+1} = \mathbf{g}^{n+1} + \gamma_n \mathbf{w}^n. \tag{186}$$

Do $n = n + 1$ *and return to (177).* $\square$

Remark 14 *The linearized advection–diffusion problem (87) can be also solved by a least–squares conjugate gradient method close to the one discussed in this section, but cheaper since the linearity of (87), we have to solve only 2 elliptic systems associated to* $\alpha I - \nu\triangle$ *per iteration. An interesting alternative is clearly use a preconditioned GMRES algorithm to solve (87), with* $\alpha I - \nu\triangle$ *as preconditioner.*

Remark 15 *We observe that each iteration of algorithm (173)-(186) requires the solution of three systems of mixed Dirichlet-Neumann boundary value problem associated with the elliptic operator* $\alpha I - \nu\Delta$*. This number is optimal for a nonlinear problem, since the solution of a linear problem by a least-squares conjugate gradient method requires the solution at each iteration of two linear systems associated with the preconditioning operator.*

Another important issue concerning algorithm (173)-(186) is their stopping criterion; we have used

$$J(\mathbf{u}^n)/J(\mathbf{u}^0) \leq \varepsilon \tag{187}$$

with ε *of the order of* 10^{-6}*.*

3.3 Solution of advection subproblem

Now we would like to consider the *pure advection problem*

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{V}\cdot\boldsymbol{\nabla})\mathbf{u} = \mathbf{0} \text{ in } \Omega \times (t^n, t^{n+1}), \\ \mathbf{u}(0) = \mathbf{u}_0,\ \mathbf{u} = \mathbf{g} \text{ on } \Gamma_- \times (t^n, t^{n+1}). \end{cases} \tag{188}$$

with $\boldsymbol{\nabla}\cdot\mathbf{V} = 0$ and $\partial\mathbf{V}/\partial t = \mathbf{0}$ in $\Omega\times(t^n, t^{n+1})$, $\Gamma_- = \{\mathbf{x} \mid \mathbf{x} \in \Gamma, \mathbf{V}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) < 0\}$ and $\partial\mathbf{g}/\partial t = 0$ on $\Gamma_- \times (t^n, t^{n+1})$. The above problem can be obtained when applying the operator–splitting scheme à la Marchuk-Yanenko to the Navier-Stokes equations, which we shall discuss later.

Solving the *pure advection problem* is a more delicate issue. Clearly, problem (188) can be solved by a *method of characteristics* (see, e.g., Pironneau 1989 [10] and Glowinski and Pironneau 1992 [102] and the references therein). An easy implementing alternative to the method of characteristics is provided by a *wave-like equation* method. It follows from that after translation and dilation on the time axis, each component of $\mathbf{u}$ is a solution of a *transport equation* of the following type:

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} + \mathbf{V}\cdot\boldsymbol{\nabla}\varphi = 0 \text{ in } \Omega \times (0,1), \\ \varphi(0) = \varphi_0,\ \varphi = g \text{ on } \Gamma_- \times (0,1), \end{cases} \tag{189}$$

Let us follow Dean, Glowinski, and Pan 1998 [101] to discuss the solution of the transport problem (189). Since each component of $\mathbf{u}$, in equation (188) verifies a transport equation such as (189), we shall focus on the solution of this last equation. The properties $\boldsymbol{\nabla} \cdot \mathbf{V} = 0$ and $\partial \mathbf{V}/\partial t = \mathbf{0}$ on $\Omega \times (0,1)$ (we also have $\partial g/\partial t = 0$ on $\Gamma_- \times (0,1)$) imply that problem (189) is "equivalent" to the (formally) wall-posed problem:

$$\begin{cases} \dfrac{\partial^2 \varphi}{\partial t^2} - \boldsymbol{\nabla} \cdot ((\mathbf{V} \cdot \boldsymbol{\nabla} \varphi)\mathbf{V}) = 0 \ in \ \Omega \times (0,1), \\ \varphi(0) = \varphi_0, \ \dfrac{\partial \varphi}{\partial t}(0) = -\mathbf{V} \cdot \boldsymbol{\nabla} \varphi_0, \\ \varphi = g \ on \ \Gamma_- \times (0,1), \ \mathbf{V} \cdot \mathbf{n} \ (\dfrac{\partial \varphi}{\partial t} + \mathbf{V} \cdot \boldsymbol{\nabla} \varphi) = 0 \ on \ (\Gamma \setminus \Gamma_-) \times (0,1). \end{cases} \tag{190}$$

Solving the *wave-like equation* (190) by a classical finite element/time stepping method is quite easy since a variational formulation of (190) is given by

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial^2 \varphi}{\partial t^2} v \, d\mathbf{x} + \int_\Omega (\mathbf{V} \cdot \boldsymbol{\nabla} \varphi)(\mathbf{V} \cdot \boldsymbol{\nabla} v) \, d\mathbf{x} \\ \qquad + \displaystyle\int_{\Gamma \setminus \Gamma_-} \mathbf{V} \cdot \mathbf{n} \varphi_t v \, d\Gamma = 0, \ \forall v \in W_0, \ a.e. \ on \ (0,T), \\ \varphi(0) = \varphi_0, \ \dfrac{\partial \varphi}{\partial t}(0) = -\mathbf{V} \cdot \boldsymbol{\nabla} \varphi_0, \\ \varphi = g \ on \ on \ \Gamma_- \times (0,1), \end{cases} \tag{191}$$

with the test function space W_0 defined by

$$W_0 = \{v | v \in H^1(\Omega), \ v = 0 \ on \ \Gamma_-\}.$$

We observe that $\mathbf{V} \cdot \mathbf{n} \geq 0$ on $\Gamma \setminus \Gamma_-$, implying that the boundary term in (191) is *dissipative*.

Let H_h^1 be a C^0 - conforming *finite element* subspace of $H^1(\Omega)$ as discussed in, e.g., Ciarlet 1978 and 1991 [24, 81]. We define W_{0h} by $W_{0h} = H_h^1 \cap W_0$; we suppose that $\lim_{h \to 0} W_{0h} = W_0$ in the usual finite element sense (see Ciarlet 1978 and 1991 [24, 81]). Next, we define $\tau_1 > 0$ by $\tau_1 = \Delta t / Q_1$, where Q_1 is a positive integer and we discretize problem (191) by

$$\varphi^0 = \varphi_{0h} (\simeq \varphi_0), \tag{192}$$

$$\begin{cases} \displaystyle\int_\Omega (\varphi^{-1} - \varphi^1) v \, d\mathbf{x} = 2\tau_1 \int_\Omega (\mathbf{V}_h \cdot \boldsymbol{\nabla} \varphi^0) v \, d\mathbf{x}, \ \forall v \in W_{0h}, \\ \varphi^{-1} - \varphi^1 \in W_{0h}, \end{cases} \tag{193}$$

and for $q = 0, \ldots, Q_1 - 1$,

$$\begin{cases} \varphi^{q+1} \in H_h^1, \ \varphi^{q+1} = g_h \ on \ \Gamma_-, \\ \displaystyle\int_\Omega \frac{\varphi^{q+1} + \varphi^{q-1} - 2\varphi^q}{\tau_1^2} v \, d\mathbf{x} + \int_\Omega (\mathbf{V}_h \cdot \boldsymbol{\nabla} \varphi^q)(\mathbf{V}_h \cdot \boldsymbol{\nabla} v) \, d\mathbf{x} \\ + \displaystyle\int_{\Gamma \setminus \Gamma_-} \mathbf{V}_h \cdot \mathbf{n} (\frac{\varphi^{q+1} - \varphi^{q-1}}{2\tau_1}) v \, d\Gamma = 0, \ \forall v \in W_{0h}, \end{cases} \tag{194}$$

where, in (193) and (194), $\mathbf{V}_h$ and g_h approximate $\mathbf{V}$ and g respectively. Scheme (192)-(194) is a centered scheme which is formally second-order accurate with respect to space and time discretizations. To be stable, scheme (192)-(194) has to verify a condition such as

$$\tau_1 \leq ch, \tag{195}$$

with c of the order of $1/||\mathbf{V}||$. If one chooses an appropriate numerical integration method to compute the first and third integrals in (194), the above scheme becomes *explicit*, i.e. φ^{q+1} is obtained via the solution of a linear system with a *diagonal* matrix.

Remark 16 *Scheme (192)-(194) does not introduce numerical dissipation, unlike the upwinding schemes commonly used to solve transport problems like (188) and (189).*

Remark 17 *Since the wave equation in (190) is, for arbitrary data, a model for simultaneous transport phenomena in the directions* $\mathbf{V}$ *and* $-\mathbf{V}$*, both playing the same role, one has to be aware that the initial condition and the boundary conditions have to be treated very accurately in order to keep at a small level the transport phenomenon taking place in the* $-\mathbf{V}$ *direction, which is here a numerical artifact.*

Remark 18 *In order to show that this projection/wave-like equation method is closely related to the Chorin's projection method [39]. Let us consider the homogeneous boundary condition,* $\mathbf{u}|_\Gamma = \mathbf{0}$ *for all* t *and set* $Q_1 = 1$ *in (192)-(194). Then we have the following scheme:*

$$\mathbf{u}^0 = \mathbf{u}_{0h} \text{ is given;} \tag{196}$$

for $n \geq 0$, $\mathbf{u}^n$ *being known,*

$$\left\{ \begin{array}{l} \displaystyle\int_\Omega \frac{\mathbf{u}^{n+1/3} - \mathbf{u}^n}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p^{n+1/3} \boldsymbol{\nabla} \cdot \mathbf{v}\, d\mathbf{x} = 0,\ \forall \mathbf{v} \in V_{0h}, \\ \displaystyle\int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u}^{n+1/3}\, d\mathbf{x} = 0,\ \forall q \in L^2_h; \\ \mathbf{u}^{n+1/3} \in V_{0h}, p^{n+1/3} \in L^2_{0h}, \end{array} \right. \tag{197}$$

$$\left\{ \begin{array}{l} \displaystyle\int_\Omega \frac{\mathbf{u}^{n+2/3} - \mathbf{u}^{n+1/3}}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} + \int_\Omega (\mathbf{u}^{n+1/3} \cdot \boldsymbol{\nabla}) \mathbf{u}^{n+1/3} \cdot \mathbf{v}\, d\mathbf{x} \\ \displaystyle = -\frac{\triangle t}{2} \int_\Omega (\mathbf{u}^{n+1/3} \cdot \boldsymbol{\nabla}) \mathbf{u}^{n+1/3} (\mathbf{u}^{n+1/3} \cdot \boldsymbol{\nabla}) \mathbf{v}\, d\mathbf{x},\ \forall \mathbf{v} \in V_{0h}; \\ \mathbf{u}^{n+2/3} \in V_{0h}. \end{array} \right. \tag{198}$$

$$\int_\Omega \frac{\mathbf{u}^{n+1} - \mathbf{u}^{n+2/3}}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{u}^{n+1} : \boldsymbol{\nabla} \mathbf{v}\, d\mathbf{x} = 0,\ \forall \mathbf{v} \in V_{0h}; \mathbf{u}^{n+1} \in V_{0h}. \tag{199}$$

The difference between the above scheme and the Chorin's projection method is a right-hand-side term in (198) which is a naturally built-in diffusion term only acting in the direction of streamlines. This extra term is also close to the one introduced in streamline-diffusion methods (e.g., see Johnson 1986 [103]).

4 Iterative solution of the Stokes type sub–problem

At each full time step of scheme (77)-(85), we have to solve twice the following *generalized Stokes problem*:

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \ in\ \Omega, \\ \nabla\cdot\mathbf{u} = 0 \ in\ \Omega, \\ \mathbf{u} = \mathbf{g}_0 \ on\ \Gamma_0,\ \nu\dfrac{\partial\mathbf{u}}{\partial n} - \mathbf{n}p = \mathbf{g}_1 \ on\ \Gamma_1 \end{cases} \tag{200}$$

with α and ν two positive parameters. Our main goal in this section is to discuss *iterative methods* for the solution of *generalized Stokes problem* (200).

4.1 Mathematical properties of the generalized Stokes problem

We suppose that in above (200), Ω is a *bounded* domain of $\mathbb{R}^d$ (with $d = 2$ or 3, in practice), $\alpha \geq 0$, $\nu > 0$, $\Gamma_0 \cap \Gamma_1 = \emptyset$, $\overline{\Gamma_0 \cup \Gamma_1} = \Gamma$; we suppose also that $\mathbf{f} \in (L^2(\Omega))^d$, $\mathbf{g}_0 = \tilde{\mathbf{g}}_0|_{\Gamma_0}$ with $\tilde{\mathbf{g}}_0 \in (H^1(\Omega))^d$, $\mathbf{g}_1 \in (L^2(\Gamma_1))^d$. If (200) has a solution $\{\mathbf{u}, p\}$ belonging to $(H^1(\Omega))^d \times L^2(\Omega)$, this solution verifies clearly

$$\begin{cases} \mathbf{u} \in V_{g_0}, p \in L^2(\Omega), \\ \displaystyle\int_\Omega (\alpha\mathbf{u}\cdot\mathbf{v} + \nu\nabla\mathbf{u} : \nabla\mathbf{v})dx - \int_\Omega p\nabla\cdot\mathbf{v}dx = \\ \displaystyle\int_\Omega \mathbf{f}\cdot\mathbf{v}dx + \int_{\Gamma_1} \mathbf{g}_1\cdot\mathbf{v}d\Gamma,\ \forall\mathbf{v} \in V_0, \\ \nabla\cdot\mathbf{u} = 0, \end{cases} \tag{201}$$

where

$$V_0 = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d,\ \mathbf{v} = \mathbf{0} \ on\ \Gamma_0\}, \tag{202}$$

$$V_{g_0} = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d,\ \mathbf{v} = \mathbf{g}_0 \ on\ \Gamma_0\}. \tag{203}$$

Actually, things would be no more complicated if, in (201), one replaces the linear functional

$$\mathbf{v} \to \int_\Omega \mathbf{f}\cdot\mathbf{v}dx + \int_{\Gamma_1} \mathbf{g}_1\cdot\mathbf{v}d\Gamma$$

by $L : (H^1(\Omega))^d \to \mathbb{R}$, defined as follows

$$L(\mathbf{v}) = \int_\Omega \mathbf{f}_0\cdot\mathbf{v}dx + \sum_{i=1}^{d}\int_\Omega \mathbf{f}_i\cdot\frac{\partial\mathbf{v}}{\partial x_i}dx + \int_{\Gamma_1}\mathbf{g}_1\cdot\mathbf{v}d\Gamma, \tag{204}$$

with $\mathbf{f}_i \in (L^2(\Omega))^d$, $\forall i$=0, 1, ... d; functional L is clearly *linear* and *continuous* over $(H^1(\Omega))^d$. We have then the following theorem of *uniqueness*:

Theorem 2 *Suppose that the above hypotheses on* $\alpha, \nu, L, \mathbf{g}_0, \mathbf{g}_1$ *hold and that* $\{\mathbf{u}, p\}$ *is a solution to*

$$\begin{cases} \mathbf{u} \in V_{g_0},\ p \in L^2(\Omega), \\ \displaystyle\int_\Omega (\alpha \mathbf{u} \cdot \mathbf{v} + \nu \nabla \mathbf{u} : \nabla \mathbf{v}) dx - \int_\Omega p \nabla \cdot \mathbf{v} dx = L(\mathbf{v}),\ \forall \mathbf{v} \in V_0, \\ \nabla \cdot \mathbf{u} = 0. \end{cases} \tag{205}$$

Then $\{\mathbf{u}, p\}$ *is unique in* $V_{g_0} \times L^2(\Omega)$ *(resp., in* $V_{g_0} \times (L^2(\Omega)/\mathbb{R})$*) if* $\displaystyle\int_{\Gamma_i} d\Gamma > 0,\ \forall i = 0, 1$ *(resp., if* $\Gamma_0 = \Gamma$*, i.e.,* $\Gamma_1 = \emptyset$*).*

The proof of the above theorem can be found in, e.g., Glowinski 2003 [4]. *Existence* results for problem (205) will be discussed in Section 4.3.

Remark 19 *If* $\alpha > 0$*, then Theorem 2 still holds if* $\Gamma_1 = \Gamma$*.*

4.2 The Stokes operator.

We suppose from now on that in addition to Ω *bounded*, we also have $\nu > 0$ and $\alpha \geq 0$ (resp., $\alpha > 0$) if $\displaystyle\int_{\Gamma_0} d\Gamma > 0$ (resp., $\Gamma_0 = \emptyset$). We call, then, *Stokes operator* the linear operator from $L^2(\Omega)$ into $L^2(\Omega)$ defined by

$$Aq = \nabla \cdot \mathbf{u}_q,\ \forall q \in L^2(\Omega), \tag{206}$$

where, in (206), $\mathbf{u}_q$ is the *unique* solution (from the Lax-Milgram Theorem (e.g., see Section 14 in Glowinski 2003 [4] or Ciarlet 1978 [24] (Chapter 1)) of the following *linear variational problem* in V_0 (space V_0 is defined by (202)):

$$\begin{cases} \mathbf{u}_q \in V_0, \\ \displaystyle\int_\Omega (\alpha \mathbf{u}_q \cdot \mathbf{v} + \nu \nabla \mathbf{u}_q : \nabla \mathbf{v}) dx = \int_\Omega q \nabla \cdot \mathbf{v} dx,\ \forall \mathbf{v} \in V_0. \end{cases} \tag{207}$$

If function q is sufficiently smooth (say $q \in H^1(\Omega)$), then $\mathbf{u}_q$ and q are related by

$$\begin{cases} \alpha \mathbf{u}_q - \nu \Delta \mathbf{u}_q + \nabla q = \mathbf{0}\ in\ \Omega, \\ \mathbf{u}_q = \mathbf{0}\ on\ \Gamma_0,\ \nu \dfrac{\partial \mathbf{u}_q}{\partial n} - \mathbf{n} q = \mathbf{0}\ on\ \Gamma_1 \end{cases} \tag{208}$$

(use the *divergence theorem* to derive (208) from (207)).

Next, we define the (pressure) space P as follows:

$$P = L^2_0(\Omega) (= \{q | q \in L^2(\Omega),\ \int_\Omega q dx = 0\})\ if\ \Gamma_0 = \Gamma, \tag{209}$$

$$P = L^2(\Omega)\ if\ \int_{\Gamma_1} d\Gamma > 0. \tag{210}$$

One of the key results of this section is provided by the following:

Theorem 3 *Operator A is a strongly elliptic, symmetric automorphism of P (i.e., is a strongly elliptic, symmetric isomorphism from P onto itself).*

PROOF: See, e.g., Glowinski 2003 [4].

Remark 20 *The spectral properties of the Stokes operator A, and of related operators, are thoroughly discussed (in the particular case $\alpha = 0$ and $\Gamma_0 = \Gamma$) in two beautiful papers Crouzeix 1974 and 1997 [61, 62]; the main motivation of the above two references is to provide a detailed analysis of the convergence properties of some of the iterative methods, for solving (200), to be discussed in the following subsections.*

4.3 Existence results for the generalized Stokes problem (205).

We can complete, now, the *uniqueness* Theorem 2; we have thus

Theorem 4 *Suppose that the pressure space P is defined by (209) or (210); suppose also that*

$$\alpha > 0 \ if \ \Gamma_1 = \Gamma, \tag{211}$$

$$\int_\Gamma \mathbf{g}_0 \cdot \mathbf{n} d\Gamma = 0 \ if \ \Gamma_0 = \Gamma. \tag{212}$$

Then the generalized Stokes problem (205) has a unique solution in $V_{g_0} \times P$.

PROOF: Let us consider first the following *linear* variational problem

$$\begin{cases} \mathbf{u}_0 \in V_{g_0}, \\ \displaystyle\int_\Omega (\alpha \mathbf{u}_0 \cdot \mathbf{v} + \nu \nabla \mathbf{u}_0 : \nabla \mathbf{v}) dx = L(\mathbf{v}), \ \forall \mathbf{v} \in V_0; \end{cases} \tag{213}$$

problem (213) has a *unique* solution. Suppose now that problem (205) has a solution $\{\mathbf{u}, p\}$ in $V_{g_0} \times P$ (necessarily *unique*, from Theorem 2) and define $\bar{\mathbf{u}}$ by

$$\bar{\mathbf{u}} = \mathbf{u} - \mathbf{u}_0. \tag{214}$$

By subtraction between (205) and (213), the pair $\{\bar{\mathbf{u}}, p\}$ verifies, necessarily

$$\begin{cases} \bar{\mathbf{u}} \in V_0, \ p \in P, \\ \displaystyle\int_\Omega (\alpha \bar{\mathbf{u}} \cdot \mathbf{v} + \nu \nabla \bar{\mathbf{u}} : \nabla \mathbf{v}) dx = \int_\Omega p \nabla \cdot \mathbf{v} dx, \ \forall \mathbf{v} \in V_0, \\ \nabla \cdot \bar{\mathbf{u}} = -\nabla \cdot \mathbf{u}_0; \end{cases} \tag{215}$$

system (214), (215) is clearly equivalent to the generalized Stokes problem (205). Actually, it follows from (215) and from the results of previous subsection that the pressure p (if it exists) verifies

$$Ap = -\nabla \cdot \mathbf{u}_0. \tag{216}$$

Conversely, if equation (216) has a solution p in P and if $\bar{\mathbf{u}}$ is the corresponding solution of problem (207) (i.e., $\bar{\mathbf{u}} = \mathbf{u}_p$) then, the pair $\{\bar{\mathbf{u}} + \mathbf{u}_0, p\}$ is the unique solution of problem (205) in $V_{g_0} \times P$. Thus, the proof of the theorem will be complete if we can show that equation (216) has a solution in P. Since operator A is, from Theorem 3, an isomorphism from P onto P, equation (216) will have a unique solution in P if we can show that its right hand side $-\nabla \cdot \mathbf{u}_0$ belongs to P. If condition (210) holds, this is obviously the case since $\mathbf{u}_0 \in V_{g_0} \subset (H^1(\Omega))^d$ implies $\nabla \cdot \mathbf{u}_0 \in L^2(\Omega) (= P$, in that case). If condition (209) holds we still have $\nabla \cdot \mathbf{u}_0 \in L^2(\Omega)$, and also, from (212), $\displaystyle\int_\Omega \nabla \cdot \mathbf{u}_0 dx = \int_\Gamma \mathbf{g}_0 \cdot \mathbf{n} d\Gamma = 0$, i.e., $\nabla \cdot \mathbf{u}_0 \in L^2_0(\Omega) (= P$, here). The proof of the theorem is complete.

Remark 21 *As we shall see in the following subsections, it is possible to solve the generalized Stokes problem (205), via the iterative solution of equation (216), without knowing explicitly operator A; similarly, it will not be necessary to know the vector valued function* $\mathbf{u}_0$*, to solve (205), via (216). All we shall need, is to be able to compute* $Aq + \nabla \cdot \mathbf{u}_0$, $\forall q \in L^2(\Omega)$*; this can be done relatively easily since, from (213) and previous subsection, we have*

$$Aq + \nabla \cdot \mathbf{u}_0 = \nabla \cdot \mathbf{U}_q,$$

where $\mathbf{U}_q$ *is the unique solution of*

$$\begin{cases} \mathbf{U}_q \in V_{g_0}, \\ \displaystyle\int_\Omega (\alpha \mathbf{U}_q \cdot \mathbf{v} + \nu \nabla \mathbf{U}_q : \nabla \mathbf{v}) dx = \int_\Omega q \nabla \cdot \mathbf{v} dx + L(\mathbf{v}), \ \forall \mathbf{v} \in V_0, \end{cases}$$

i.e., of an elliptic system for the operator $\alpha I - \nu \Delta$.

4.4 A saddle-point interpretation of the generalized Stokes problem.

As we shall see in a moment, any pair $\{\mathbf{u}, p\}$ solution of the generalized Stokes problem (205) can be viewed as a *saddle-point* of a well chosen *Lagrangian functional*, defined over $(H^1(\Omega))^d \times L^2(\Omega)$. This interpretation is not necessary to prove the convergence of the various iterative methods to be discussed in the following subsections; what matters really there are the properties of the Stokes operator A defined in Section 4.2.

Let X and Y be two *non-empty* sets and let f be a mapping from $X \times Y$ into $\overline{\mathbb{R}}$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. We suppose that f is *proper*, i.e., there exists at least one pair $\{x, y\} \in X \times Y$ so that $f(x, y)$ is *finite*.

Definition 19.1 A pair $\{a, b\}$ is called a *saddle-point* of the functional f over $X \times Y$ if

$$\begin{cases} \{a, b\} \in X \times Y, \ f(a, b) \in \mathbb{R}, \\ f(a, y) \leq f(a, b) \leq f(x, b), \ \forall \{x, y\} \in X \times Y. \end{cases} \tag{217}$$

We associate with the generalized Stokes problem (205) the *Lagrangian functional*

$$\mathcal{L}(\mathbf{v}, q) = \frac{1}{2} \int_\Omega (\alpha |\mathbf{v}|^2 + \nu |\nabla \mathbf{v}|^2) dx - L(\mathbf{v}) - \int_\Omega q \nabla \cdot \mathbf{v} dx; \tag{218}$$

functional $\mathcal{L}$ is C^∞ on $(H^1(\Omega))^d \times L^2(\Omega)$. We have then the following:

Theorem 5 *Suppose that functional $\mathcal{L}$ has a saddle-point $\{\mathbf{u}, p\}$ over $V_{g_0} \times L^2(\Omega)$, i.e.,*

$$\begin{cases} \{\mathbf{u}, p\} \in V_{g_0} \times L^2(\Omega), \\ \mathcal{L}(\mathbf{u}, q) \leq \mathcal{L}(\mathbf{u}, p) \leq \mathcal{L}(\mathbf{v}, p), \ \forall \{\mathbf{v}, q\} \in V_{g_0} \times L^2(\Omega). \end{cases} \tag{219}$$

Then $\{\mathbf{u}, p\}$ is a solution of the Stokes problem (205). Conversely, any solution of (205) belonging to $V_{g_0} \times L^2(\Omega)$ is a saddle-point of $\mathcal{L}$ over $V_{g_0} \times L^2(\Omega)$.

4.5 A gradient method for the generalized Stokes problem.

It follows from Theorem 5 of the previous subsection that any solution of the generalized Stokes problem (205) is also a saddle–point over $V_{g_0} \times L^2(\Omega)$ of the Lagrangian functional defined by (218); conversely, any saddle-point of $\mathcal{L}$ over $V_{g_0} \times L^2(\Omega)$ is also a solution of the Stokes problem (205). This *equivalence* property implies, among other things, that it makes sense to attempt solving problem (205) by solving the saddle–point problem (219), by the Uzawa's algorithm:

$$p^0 \in L^2(\Omega), \ \textit{given}; \tag{220}$$

for $n \geq 0, p^n \in L^2(\Omega)$ being known, we obtain $\mathbf{u}^n$ and p^{n+1} via

$$\begin{cases} \mathbf{u}^n \in V_{g_0}, \\ \displaystyle\int_\Omega (\alpha \mathbf{u}^n \cdot \mathbf{v} + \nu \nabla \mathbf{u}^n : \nabla \mathbf{v}) dx = \int_\Omega p^n \nabla \cdot \mathbf{v} dx + L(\mathbf{v}), \ \forall \mathbf{v} \in V_0, \end{cases} \tag{221}$$

$$p^{n+1} = p^n - \rho \nabla \cdot \mathbf{u}^n. \tag{222}$$

Remark 22 *Problem (221) is a system for the elliptic operator $\alpha I - \nu \Delta$. If*

$$L(\mathbf{v}) = \int_\Omega \mathbf{f} \cdot \mathbf{v} dx + \int_{\Gamma_1} \mathbf{g}_1 \cdot \mathbf{v} d\Gamma,$$

with, for example, $\mathbf{f} \in (L^2(\Omega))^d$ and $\mathbf{g}_1 \in (L^2(\Gamma_1))^d$, respectively, then problem (221) is equivalent to solving in V_{g_0} the elliptic system

$$\begin{cases} \alpha \mathbf{u}^n - \nu \Delta \mathbf{u}^n = \mathbf{f} - \nabla p^n \ \textit{in} \ \Omega, \\ \mathbf{u}^n = \mathbf{g}_0 \ \textit{on} \ \Gamma_0, \ \nu \dfrac{\partial \mathbf{u}^n}{\partial n} = \mathbf{g}_1 + \mathbf{n} p^n \ \textit{on} \ \Gamma_1 \end{cases} \tag{223}$$

(the boundary condition on Γ_1 makes sense only if p^n has a trace on Γ_1). □

Concerning the convergence of algorithm (220)-(222) we have then

Theorem 6 *Suppose that the parameter ρ in (222) satisfies*

$$0 < \rho < 2\nu/d. \tag{224}$$

We have then the following convergence properties for algorithm (220)-(222):

$$\lim_{n\to+\infty} \mathbf{u}^n = \mathbf{u} \text{ in } (H^1(\Omega))^d, \tag{225}$$

$$\lim_{n\to+\infty} p^n = p \text{ in } L^2(\Omega), \text{ if } P = L^2(\Omega), \tag{226}$$

$$\lim_{n\to+\infty} p^n = p + (\int_\Omega p^0 dx)/meas.(\Omega) \text{ in } L^2(\Omega), \text{ if } P = L^2_0(\Omega), \tag{227}$$

where, in (225)-(227), $\{\mathbf{u}, p\}$ is the unique solution of the generalized Stokes problem (205) in $V_{g_0} \times P$.

The proof of the above theorem can be found in, e.g., Glowinski 2003 [4], or Glowinski 1984 [8] (Theorem 5.11).

Algorithm (220)-(222) can be interpreted as a gradient method by using operator A introduced in Section 4.2. If $\Gamma_0 = \Gamma$, we suppose for simplicity that, in (220), we take $p^0 \in L^2_0(\Omega)(= P$ in that case), implying (from (222), (212)) that $p^n \in L^2_0(\Omega),\ \forall n \geq 0$. Proceeding as in Section 4.3, we define $\mathbf{u}_0$ by

$$\begin{cases} \mathbf{u}_0 \in V_{g_0}, \\ \displaystyle\int_\Omega (\alpha \mathbf{u}_0 \cdot \mathbf{v} + \nu \nabla \mathbf{u}_0 : \nabla \mathbf{v}) dx = L(\mathbf{v}),\ \forall \mathbf{v} \in V_0. \end{cases} \tag{228}$$

Subtracting (228) to (221) we obtain

$$\begin{cases} \mathbf{u}^n - \mathbf{u}_0 \in V_0, \\ \displaystyle\int_\Omega [\alpha(\mathbf{u}^n - \mathbf{u}_0) \cdot \mathbf{v} + \nu \nabla(\mathbf{u}^n - \mathbf{u}_0) : \nabla \mathbf{v}] dx = \int_\Omega p^n \nabla \cdot \mathbf{v} dx,\ \forall \mathbf{v} \in V_0, \end{cases}$$

which implies, from the definition of operator A (see Section 4.2) that

$$Ap^n = \nabla \cdot (\mathbf{u}^n - \mathbf{u}_0),$$

i.e.,

$$\nabla \cdot \mathbf{u}^n = Ap^n + \nabla \cdot \mathbf{u}_0,$$

which implies in turn that algorithm (220)-(222) is equivalent to

$$p^0 \in P \text{ is given}; \tag{229}$$

then for $n \geq 0, p^n \in P$ being known,

$$p^{n+1} = p^n - \rho(Ap^n + \nabla \cdot \mathbf{u}_0). \tag{230}$$

Algorithm (229)-(230) is clearly a *fixed point* method for solving problem (216), namely

$$Ap = -\nabla \cdot \mathbf{u}_0.$$

We introduce now the functional $J_* : P \to \mathbb{R}$ defined by

$$J_*(q) = \frac{1}{2}\int_\Omega (Aq)qdx + \int_\Omega \nabla \cdot \mathbf{u}_0 q dx, \ \forall q \in P. \tag{231}$$

The differential J'_* of functional J_* is given by

$$J'_*(q) = Aq + \nabla \cdot \mathbf{u}_0, \tag{232}$$

implying that algorithms (220)-(222) and (229)-(230) can also be written as follows:

$$p^0 \in P \ \textit{is given}; \tag{233}$$

and for $n \geq 0, \ p^n \in P$ *being known*

$$p^{n+1} = p^n - \rho J'_*(p^n); \tag{234}$$

algorithm (233)-(234) is clearly a *gradient algorithm, with constant step* ρ, applied to the solution of the minimization problem

$$\begin{cases} p \in P, \\ J_*(p) \leq J_*(q), \ \forall q \in P. \end{cases} \tag{235}$$

Equation (216) is the *Euler-Lagrange equation* associated to the minimization problem (235) and can also be written as

$$J'_*(p) = 0. \tag{236}$$

Actually, the minimization problem (235) is the *dual problem* associated with the *saddle-point problem*

$$\begin{cases} \{\mathbf{u}, p\} \in V_{g_0} \times P, \\ \mathcal{L}(\mathbf{u}, q) \leq \mathcal{L}(\mathbf{u}, p) \leq \mathcal{L}(\mathbf{v}, p), \ \forall \{\mathbf{v}, q\} \in V_{g_0} \times P, \end{cases} \tag{237}$$

with $\mathcal{L}$ still defined by (218); this follows from (Glowinski 2003 [4] (Chapter 5, Section 20)

Theorem 7 *The minimization problem (235) and the dual problem associated with problem (237) coincide.*

4.6 Conjugate gradient algorithms for the generalized Stokes problem.

To apply the *conjugate gradient algorithm* (103)-(110) to the solution of the minimization problem (216), (235), we first equip the space P with the classical scalar product of $L^2(\Omega)$, namely

$$\{q, q'\} \to \int_\Omega qq' dx, \ \forall \{q, q'\} \in P \times P, \tag{238}$$

and the corresponding norm and then obtain the following conjugate gradient algorithm, a variant of the *Uzawa's algorithm* (220)-(222):

$$p^0 \in P \text{ is given;} \tag{239}$$

solve

$$\begin{cases} \mathbf{u}^0 \in V_{g_0};\ \forall \mathbf{v} \in V_0 \text{ we have} \\ \displaystyle\int_\Omega (\alpha \mathbf{u}^0 \cdot \mathbf{v} + \nu \nabla \mathbf{u}^0 : \nabla \mathbf{v})\, dx = L(\mathbf{v}) + \int_\Omega p^0 \nabla \cdot \mathbf{v} dx, \end{cases} \tag{240}$$

compute

$$g^0 = \nabla \cdot \mathbf{u}^0 \tag{241}$$

and set

$$w^0 = g^0. \tag{242}$$

For $n \geq 0$, assuming that p^n, g^n, w^n are known, solve

$$\begin{cases} \bar{\mathbf{u}}^n \in V_0, \\ \displaystyle\int_\Omega (\alpha \bar{\mathbf{u}}^n \cdot \mathbf{v} + \nu \nabla \bar{\mathbf{u}}^n : \nabla \mathbf{v})\, dx = \int_\Omega w^n \nabla \cdot \mathbf{v} dx,\ \forall \mathbf{v} \in V_0, \end{cases} \tag{243}$$

compute

$$\bar{g}^n = \nabla \cdot \bar{\mathbf{u}}^n, \tag{244}$$

and then

$$\rho_n = \int_\Omega |g^n|^2 dx / \int_\Omega \bar{g}^n w^n dx. \tag{245}$$

Update $\mathbf{u}^n$, p^n and g^n by

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \rho_n \bar{\mathbf{u}}^n \tag{246}$$

$$p^{n+1} = p^n - \rho_n w^n, \tag{247}$$

$$g^{n+1} = g^n - \rho_n \bar{g}^n. \tag{248}$$

If $||g^{n+1}||_{L^2(\Omega)} / ||g^0||_{L^2(\Omega)} \leq \varepsilon$ take $p = p^{n+1}$; else, compute

$$\gamma_n = ||g^{n+1}||^2_{L^2(\Omega)} / ||g^n||^2_{L^2(\Omega)} \tag{249}$$

and update w^n via

$$w^{n+1} = g^{n+1} + \gamma_n w^n. \tag{250}$$

Do $n = n + 1$ and return to (243). □

The rate of convergence of the above conjugate gradient algorithm (239)-(250) has been studied in Glowinski 2003 [4] (Chapter V, Section 21).

Algorithms (220)-(222) and (239)-(250) may be slow in practice, particularly for flow at large Reynolds number where $\alpha \sim 1/\triangle t$ is taken very large (to follow the fast dynamics of such flow) and where ν is very small. To explain this behavior let us recall that, from the definition of operator A (see Section 4.2), we have $\mathbf{u}_q = -(\alpha I - \nu \Delta)^{-1} \nabla q$ via (207) and

$$Aq = -\nabla \cdot (\alpha I - \nu \Delta)^{-1} \nabla q,\ \forall q \in P. \tag{251}$$

Assuming that $(\alpha I - \nu\Delta)^{-1}$ and ∇ *commute* (which is not strictly true, in general) we obtain, from (251),

$$Aq = -\nabla \cdot \nabla(\alpha I - \nu\Delta)^{-1}q = -\Delta(\alpha I - \nu\Delta)^{-1}q,$$

i.e., $A = -\Delta(\alpha I - \nu\Delta)^{-1}$, which implies in turn that

$$A^{-1} = (\alpha I - \nu\Delta)(-\Delta)^{-1} = \alpha(-\Delta)^{-1} + \nu I. \tag{252}$$

Relation (252) shows that if

$$\nu >> \alpha \tag{253}$$

operator A behaves, essentially, like I/ν, explaining why algorithm (239)-(250) have good convergence properties if condition (253) holds. On the other hand, if

$$\alpha >> \nu, \tag{254}$$

we have $A \simeq -\dfrac{1}{\alpha}\Delta$ and therefore operator A is very far from being a multiple of the identity operator, explaining the very slow convergence of the above algorithms (we can expect operator A to have a condition number of the order of h^{-2}, after space discretization, if (254) holds). In Cahouet and Chabard 1988 [63], they have considered the case (*without boundary*) where $\Omega = \mathbb{R}^d$ and justified, in some sense, relation (252), whose derivation was quite heuristical. On the basis of these results, we shall assume that A^{-1} behaves like

$$\nu I, \quad if\ \nu >> \alpha, \tag{255}$$

and

$$\begin{cases} \alpha(-\Delta)^{-1}\ (\textit{for the homogeneous Dirichlet condition on}\ \Gamma_1 \\ \textit{and the homogeneous Neumann condition on}\ \Gamma_0), \quad if\ \alpha >> \nu. \end{cases} \tag{256}$$

Relation (256) implies that *preconditioning is necessary* if $\alpha >> \nu$. In order to have a preconditioning operator whose good properties remain uniform when the ratio α/ν varies from 0 to $+\infty$, we suggest to take as preconditioner (as done in Cahouet and Chabard 1988 [63], for the case $\Gamma_0 = \Gamma,\ \Gamma_1 = \emptyset$) the *isomorphism* S from P onto P defined by

$$S^{-1} = \alpha(-\Delta)^{-1} + \nu I; \tag{257}$$

in (257), the *Green operator* $(-\Delta)^{-1}$ is associated to the boundary conditions described in (256). The fact that the preconditioning operator S is defined by its inverse does not create practical problems as shown in the following section.

A preconditioned conjugate gradient algorithm

As already observed above, it follows from the properties of operators A that problems (216), (235) can be solved by the *conjugate gradient algorithms* when the Hilbert space P being the usual $L^2(\Omega)$-scalar product, namely

$$\{q, q'\} \to \int_\Omega qq'dx,\ \forall q, q' \in P. \tag{258}$$

In order to avoid the deterioration of the convergence properties, associated with large values of the ratio α/ν, and to keep the convergence as uniform as possible, we suggested to employ as scalar product on space P the one advocated in Cahouet and Chabard 1988 [63], namely

$$\{q, q'\} \rightarrow \int_\Omega (Sq)q'dx, \ \forall q, q' \in P, \quad \textit{with operator } S \textit{ defined, via } S^{-1}, \textit{ by } (257). \tag{259}$$

Using the scalar product (259) leads to the following conjugate gradient algorithm, a sophisticated variant of algorithm (239)-(250):

$$p^0 \in P \textit{ is given}; \tag{260}$$

solve

$$\begin{cases} \mathbf{u}^0 \in V_{g_0}; \ \forall \mathbf{v} \in V_0, \\ \displaystyle\int_\Omega [\alpha \mathbf{u}^0 \cdot \mathbf{v} + \nu \nabla \mathbf{u}^0 : \nabla \mathbf{v}] \, dx = L(\mathbf{v}) + \int_\Omega p^0 \nabla \cdot \mathbf{v} dx, \end{cases} \tag{261}$$

and set

$$r^0 = \nabla \cdot \mathbf{u}^0. \tag{262}$$

Solve now

$$\begin{cases} -\Delta \boldsymbol{\varphi}^0 = r^0 \ in \ \Omega, \\ \dfrac{\partial \boldsymbol{\varphi}^0}{\partial n} = 0 \ on \ \Gamma_0, \ \boldsymbol{\varphi}^0 = 0 \ on \ \Gamma_1, \end{cases} \tag{263}$$

if $\displaystyle\int_{\Gamma_i} d\Gamma > 0, \ \forall i = 0, 1;$ *or*

$$\begin{cases} -\Delta \boldsymbol{\varphi}^0 = r^0 \ in \ \Omega, \\ \dfrac{\partial \boldsymbol{\varphi}^0}{\partial n} = 0 \ on \ \Gamma, \ \displaystyle\int_\Omega \boldsymbol{\varphi}^0 dx = 0, \end{cases} \tag{264}$$

if $\Gamma_0 = \Gamma$; *or*

$$\begin{cases} -\Delta \boldsymbol{\varphi}^0 = r^0 \ in \ \Omega, \\ \boldsymbol{\varphi}^0 = 0 \ on \ \Gamma, \end{cases} \tag{265}$$

if $\Gamma_1 = \Gamma$. *Then set*

$$g^0 = \nu r^0 + \alpha \boldsymbol{\varphi}^0, \tag{266}$$

$$w^0 = g^0. \tag{267}$$

Then, for $n \geq 0$, *assuming that* p^n, r^n, g^n, w^n *are known, compute* p^{n+1}, r^{n+1}, g^{n+1}, w^{n+1} as follows:

Solve:

$$\begin{cases} \bar{\mathbf{u}}^n \in V_0; \ \forall \mathbf{v} \in V_0, \\ \displaystyle\int_\Omega [\alpha \bar{\mathbf{u}}^n \cdot \mathbf{v} + \nu \nabla \bar{\mathbf{u}}^n : \nabla \mathbf{v}] \, dx = \int_\Omega w^n \nabla \cdot \mathbf{v} dx, \end{cases} \tag{268}$$

and set

$$\bar{r}^n = \nabla \cdot \bar{\mathbf{u}}^n. \tag{269}$$

Compute

$$\rho_n = \int_\Omega r^n g^n dx / \int_\Omega \bar{r}^n w^n dx, \tag{270}$$

and then

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \rho_n \bar{\mathbf{u}}^n \tag{271}$$

$$p^{n+1} = p^n - \rho_n w^n, \tag{272}$$

$$r^{n+1} = r^n - \rho_n \bar{r}^n. \tag{273}$$

Solve, next,

$$\begin{cases} -\Delta\bar{\varphi}^n = \bar{r}^n \text{ in } \Omega, \\ \dfrac{\partial\bar{\varphi}^n}{\partial n} = 0 \text{ on } \Gamma_0, \ \bar{\varphi}^n = 0 \text{ on } \Gamma_1, \end{cases} \tag{274}$$

if $\int_{\Gamma_i} d\Gamma > 0, \ \forall i = 0, 1$; or

$$\begin{cases} -\Delta\bar{\varphi}^n = \bar{r}^n \text{ in } \Omega, \\ \dfrac{\partial\bar{\varphi}^n}{\partial n} = 0 \text{ on } \Gamma, \ \displaystyle\int_\Omega \bar{\varphi}^n dx = 0, \end{cases} \tag{275}$$

if $\Gamma_0 = \Gamma$; or

$$\begin{cases} -\Delta\bar{\varphi}^n = \bar{r}^n \text{ in } \Omega, \\ \bar{\varphi}^n = 0 \text{ on } \Gamma, \end{cases} \tag{276}$$

if $\Gamma_1 = \Gamma$. Then, compute

$$g^{n+1} = g^n - \rho_n(\nu\bar{r}^n + \alpha\bar{\varphi}^n). \tag{277}$$

If $\int_\Omega r^{n+1} g^{n+1} dx / \int_\Omega r^0 g^0 dx \leq \varepsilon$, *take* $p = p^{n+1}$; *else, compute*

$$\gamma_n = \int_\Omega r^{n+1} g^{n+1} dx / \int_\Omega r^n g^n dx, \tag{278}$$

and update w^n *by*

$$w^{n+1} = g^{n+1} + \gamma_n w^n. \tag{279}$$

Do $n = n + 1$ *and return to (268).* □

Remark 23 *Each iteration of algorithm (260)-(279) requires the solution of one elliptic system for the operator* $\mathbf{v} \to \alpha\mathbf{v} - \nu\Delta\mathbf{v}$. *As already mentioned, for flow at large Reynolds number where* $\alpha \sim 1/\Delta t$ *is large and* ν *is small, the discrete analogues to the above operator are fairly well conditioned, symmetric and positive definite matrices, making the iterative solution of the corresponding linear systems quite inexpensive. We also have to solve the Poisson problems (one among (263), (264), and (265), and another one among (274), (275), and (276)). We shall discuss this aspect of the practical implementation of algorithm*

(260)-(279) later. Actually, it follows from, e.g., Glowinski 2003 [4] (Chapter III, Sections 14.4 and 14.5), that the Poisson problems (263), (265) and (274), (276) are well-posed if Ω *is bounded. Suppose now that* $\Gamma_0 = \Gamma$*; assuming that relation (212) holds (which is necessary for problem (205) to have a solution), it follows from, e.g., Glowinski 2003 [4] (Chapter III, Section 14.3), that the Poisson-Neumann problem (264) is well-posed, since (212) implies*

$$\int_\Omega \nabla \cdot \mathbf{u}^0 dx = \int_\Gamma \mathbf{g}_0 \cdot \mathbf{n} d\Gamma = 0.$$

A similar result holds for the Poisson-Neumann problem (275), since $\bar{\mathbf{u}}^n \in V_0 (= (H_0^1(\Omega))^d$*, here) implies*

$$\int_\Omega \nabla \cdot \bar{\mathbf{u}}^n dx = \int_\Gamma \bar{\mathbf{u}}^n \cdot \mathbf{n} d\Gamma = 0, \ \forall n \geq 0.$$

Remark 24 *Algorithm (260)-(279) has proved to be quite effective for solving a large variety of Navier-Stokes problems, for a large range of Reynolds numbers. To be more precise, with* ε *of the order of* 10^{-14} *in the stopping criterion, it is very rare that more than ten iterations of algorithm (260)-(279) are needed to solve the generalized Stokes problem (205), even for complicated three dimensional flow problems, requiring several million of grid points for the space discretization. This high level of performances definitely justifies the choice of the operator* S *defined by (257), as preconditioner. From this facts, we feel obliged to quote Dennis and Schnabel 1989 [64] on the convergence of conjugate gradient algorithms (in this quotation* p *is the number of iterations necessary to achieve the convergence and* n *is the dimension of the optimization problem):*

> "It is not unusual for strictly convex quadratic arising from discretized partial differential equations to be solved with $p \sim n/10^3$. Such spectacularly successful preconditioning nearly always comes from deep insight into the problem and not from matrix theoretic considerations. They often come from discretizing and solving a simplified problem."

There is nothing to add to the above quotation.

References

[1] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer–Verlag, New York, 1988.

[2] M. Lesieur, *Turbulence in Fluids*, Kluwer, Dordrecht, 1990.

[3] E. Guyon, J.P. Hulin, and L. Petit, *Hydrodynamique Physique*, Intereditions/Editions du CNRS, Paris, 1991.

[4] R. Glowinski, Finite element methods for incompressible viscous flow, in *Handbook of Numerical Analysis*, **Vol. IX**, P.G. Ciarlet, J.-L. Lions (deceased) eds., North-Holland, Amsterdam, 2003.

[5] W. Prager, *Introduction to Mechanics of Continua*, Ginn and Company, Boston, MA, 1961.

[6] G.K. Batchelor, *An Introduction to Fluid Mechanics*, Cambridge University Press, Cambridge, U.K., 1967.

[7] A.J. Chorin and J.E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, Springer–Verlag, New York, 1990.

[8] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer–Verlag, New York, 1984.

[9] M.O. Bristeau, R. Glowinski, B. Mantel, J. Periaux, P. Perrier, Numerical Methods for incompressible and compressible Navier-Stokes problems, in *Finite Element in Fluids*, **Vol. 6**, R.H. Gallagher, G. Carey, J.T. Oden, and O.C. Zienkiewicz eds., J. Wiley, Chicester, 1985, 1–40.

[10] O. Pironneau, *Finite Element Methods for Fluids*, J. Wiley, Chicester, 1989.

[11] J. Leray, Sur le mouvement d'un liquide visqueux emplissant l'espace, *Acta Mathematica*, **63**(1934), 193–248.

[12] J. Leray, Essai sur les mouvements d'un liquide visqueux que limitent des parois, *J. Math. Pures et Appl.*, **13**(1934), 331–418.

[13] E. Hopf, Uber die Anfangswertaufgabe fur die hydrodynamischen Grundgleichungen, *Math. Nachrichten*, **4**(1951), 213–231.

[14] J. Leray, Aspects de la mécanique théorique des fluides, *La Vie des Sciences*, Comptes Rendus de l'Académie des Sciences, Paris, Série Générale, **11**(1994), 287–290.

[15] J.L. Lions and G. Prodi, Un théorème d'existence et d'unicité dans les équations de Navier-Stokes en dimension2, *C.R. Acad. Sci., Paris* **248**, 3519–3521.

[16] J.L. Lions, *Equations Différentielles Opérationnelles et Problémes aux Limites*, Springer–Verlag, Berlin, 1961.

[17] J.L. Lions, *Quelques Méthodes de Résolution des Problémes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[18] O. Ladysenskaya, *Theory and Numerical Analysis of the Navier-Stokes Equations*, Gordon and Breach, New York, NY, 1969.

[19] R. Temam, *The Mathematical Theory of Viscous Incompressible Flow*, North-Holland, Amsterdam, 1977.

[20] L. Tartar, *Topics in Nonlinear Analysis*, Publications Mathématiques d'Orsay, Université Paris-Sud, Départment de Mathématiques, Paris, 1978.

[21] H.O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, Boston, MA, 1989.

[22] P.L. Lions, *Mahtematical Topics in Fluid Mechanics, Vol I: Incompressible Models*, Oxford University, Oxford, UK, 1996.

[23] M. Marion and R. Temam, *Navier-Stokes Equations*, in *Handbook of Numerical Analysis*, **Vol. VI**, P.G. Ciarlet, J.-L. Lions (deceased) eds., North-Holland, Amsterdam, 1998, 503–689.

[24] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[25] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, NY, 1991.

[26] S.C. Brenner and L.R. Scott *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, NY, 1994.

[27] N.N. Yanenko, *The Method of Fractional Steps*, Springer-Verlag, Berlin, 1971.

[28] G.I. Marchuk, *Methods of Numerical Mathematics*, Springer-Verlag, New York, NY, 1975.

[29] G.I. Marchuk, Splitting and alternating direction methods. In Ciarlet, P.G., and Lions, J.L. (eds.) *Handbook of Numerical Analysis*, **Vol I**, North-Holland, Amsterdam, 1990, 197–462.

[30] M. Crouzeix and A. Mignot, *Analyse Numérique des Equations Différentielles Ordiuaires*, Masson, Paris, 1984.

[31] R. Glowinski and P. Le Tallec, *Augmented Lagrangians and Operator Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, PA, 1989.

[32] R. Glowinski, Viscous flow simulation by finite element methods and related numerical techniques. In Murman, E.M., and Abarbanel, S.S. (eds.) *Progress and Supercomputing in Computational Fluid Dynamics*, Birkhauser, Boston, MA, 1985, 173–210.

[33] R. Glowinski, Splitting methods for the numerical solution of the incompressible Navier-Stokes equations. In Balakrishnan, A.V., Dorodnitsyn, A.A., and Lions, J.L. (eds.) *Vistas in Applied Mathematics*, Optimization Software, New York, NY, 1986, 57–95.

[34] G. Strang, On the construction and comparison of difference schemes, *SIAM J. Num. Anal.*, **5**(1968), 506–517.

[35] J.T. Beale and A. Majda, Rates of convergence for viscous splitting of the Navier-Stokes equations, *Math. Comp.*, **37**(1981), 243–260.

[36] R. Leveque and J. Oliger, Numerical methods based on additive splitting for hyperbolic partial differential equations, *Math. Comp.*, **37** (1983), 243–260.

[37] P.A. Raviart and J.M. Thomas, *Introduction à l'Analyse Numérique des Equations aux Dérivées Partielles*, Masson, Paris, 1983.

[38] A.J. Chorin, A numerical method for solving incompressible viscous flow problems, *J. Comp. Phys.*, **2** (1967), 12–26.

[39] A.J. Chorin, Numerical solution of the Navier-Stokes equations, *Math. Comp.*, **23** (1968), 341–354.

[40] R. Temam, Sur l'approximation des équations de Navier-Stokes par la méthode des pas fractionnaires (I), *Arch. Rat. Mech. Anal.*, **32** (1969), 135–153.

[41] R. Temam, Sur l'approximation des équations de Navier-Stokes par la méthode des pas fractionnaires (II), *Arch. Rat. Mech. Anal.*, **33** (1969), 377–385.

[42] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.

[43] J. Daniel, *The Approximate Minimization of Functionals*, Prentice Hall, Englewood Cliffs, NJ, 1970.

[44] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, NY, 1971.

[45] M.R. Hestenes and E.L. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Bureau National Standards*, Section B, **49** (1952), 409–436.

[46] R.W. Freund, G.H. Golub, and N.M. Nachtigal, Iterative solution of linear systems, *Acta Numerica 1992*, Cambridge University Press, 1992, 57–100.

[47] J. Nocedal, Theory of algorithms for unconstrained optimization, *Acta Numerica 1992*, Cambridge University Press, 1992, 199–242.

[48] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, PA, 1995.

[49] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, MA, 1995.

[50] G.H. Golub and D.P. O'Leary, Some history of the conjugate gradient and Lanczos algorithms: 1948-1976, *SIAM Review*, **31** (1989), 50–102.

[51] E. Zeidler, *Nonlinear Functional Analysis and its Applications. Volume I: Fixed-Point Theorems*, Springer-Verlag, New York, NY, 1986.

[52] I. Ekeland and R. Teman, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[53] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.

[54] J.M. Ortega, and W.C. Rheinboldt, Local and global convergence of generalized linear iterations. In Ortega, J.M., and Rheinboldt, W.C. (eds.) *Numerical Solution of Nonlinear Problems*, SIAM, Philadelphia, PA, 1970.

[55] J.M. Ortega and W.C. Rheinboldt, A general convergence result for unconstrained minimization methods, *SIAM J. Num. Anal.*, **9** (1972), 40–43.

[56] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[57] M.J.D. Powell, Some convergence properties of the conjugate gradient method, *Math. Program.*, **11** (1976), 42–49.

[58] M.J.D. Powell, Restart procedures of the conjugate gradient method, *Math. Program.*, **12** (1977), 148–162.

[59] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[60] J.B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[61] M. Crouzeix, Etude d'une méthode de linéarisation. Résolution numérique des équations de Stokes stationaires. In *Approximations et Méthodes Itératives de Résolution d'Inéquations Variationelles et de Problèmes Non Linéaires,* Cahiers de l'IRIA, **12** (1974), 139-244.

[62] M. Crouzeix, On an operator related to the convergence of Uzawa's algorithm for the Stokes equation. In Bristeau, M.O., Etgen, G., Fitzgibbon, W., Lions, J.L., Périaux, J., and Wheeler, M.F. (eds.) *Computational Science for the 21st Century,* Wiley, Chichester, 1997, 242–259.

[63] J. Cahouet and J.P. Chabard, Some fast 3-D solvers for the generalized Stokes problem, *Int. J. Numer. Meth. in Fluids*, **8** (1988), 269–295.

[64] J.E. Dennis and R.B. Schnabel, A view of unconstrained optimization. In Newhauser, G.L., Rinnooy Kan, A.H.G., and Todd, M.J. (eds.) *Handbook in Operations Research and Management Science,* **Vol. 1***: Optimization,* North-Holland, Amsterdam, 1989, 1–66.

[65] F. Thomasset, *Implementation of Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, New York, NY, 1981.

[66] R. Peyret and T.D. Taylor, *Computational Methods for Fluid Flow*, Springer-Verlag, New York, NY, 1982.

[67] C. Cuvelier, A. Segal, and A. Van Steenhoven, *Finite Element Methods and Navier-Stokes Equations*, Reidel, Dordrecht, 1986.

[68] M. Fortin, Finite element solution of the Navier-Stokes equations, *Acta Numerica 1993*, Cambridge University Press, 1993, 239–284.

[69] M.D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, Boston, MA, 1989.

[70] C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics, Volume 1: Fundamental and General Techniques*, Springer-Verlag, Berlin, 1991.

[71] C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics, Volume 2: Specific Techniques for Different Flow Categories*, Springer-Verlag, Berlin, 1991.

[72] M.D. Gunzburger and R.A. Nicolaides (eds.), *Incompressible Computational Fluid Dynamics,* Cambridge University Press, New York, NY, 1993.

[73] L. Quartapelle, *Numerical Solution of the Incompressible Navier-Stokes Equations*, Birkhauser, Basel, 1993.

[74] F.K. Hebeker, R. Rannacher, and G. Wittum (eds.), *Numerical Methods for the Navier-Stokes Equations*, Vieweg, Braunschweig/Wiesbaden, 1994.

[75] P.M. Gresho and R.L. SANI, *Incompressible Flow and the Finite Element Method: Advection-Diffusion and Isothermal Laminar Flow,* J. Wiley, Chichester, 1998.

[76] E. Fernandez-Cara and M.M. Beltran, The convergence of two numerical schemes for the Navier-Stokes equations, *Numerische Mathematik*, **55** (1989), 33–60.

[77] P. Kloucek and F.S. Rys, On the stability of the fractional step-θ-scheme for the Navier-Stokes equations, *SIAM J. Num. Anal.*, **31** (1994), 1312–1335.

[78] P. Hood and C. Taylor, A numerical solution of the Navier-Stokes equations using the finite element technique, *Computers and Fluids*, **1** (1973), 73–100.

[79] M. Bercovier and O. Pironneau, Error estimates for finite element method solution of the Stokes problem in the primitive variables, *Numer. Math.*, **33** (1979), 211–224.

[80] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs, NJ, 1973.

[81] P.G. Ciarlet, Basic error estimates for elliptic problems. In Ciarlet, P.G., and Lions, J.L. (eds.) *Handbook of Numerical Analysis*, **Vol. II**, North-Holland, Amsterdam, 1991, 17–351.

[82] R. Glowinski, Finite element methods for the numerical simulation of incompressible viscous flow. Introduction to the control of the Navier-Stokes equations. In Anderson, C.R., and Greengard, C. (eds.) *Vortex Dynamics and Vortex Methods*, Lecture in Applied Mathematics, Vol. 28, American Mathematical Society, Providence, RI, 1991, 219–301.

[83] M.O. Bristeau, R. Glowinski, and J. Periaux, Numerical methods for the Navier-Stokes equations. Applications to the simulation of compressible and incompressible viscous flow, *Computer Physics Reports*, **6** (1987), 73–187.

[84] E.J. Dean, R. Glowinski, and C.H. Li, Supercomputer solution of partial differential equation problems in Computational Fluid Dynamics and in Control, *Computer Physics Communications*, **53** (1989), 401–439.

[85] R. Glowinski and O. Pironneau, Finite element methods for Navier-Stokes equations, *Annual Review of Fluid Mechanics*, **24** (1992), 167–204.

[86] T.J.R. Hughes, L.P. Franca, and M. Balestra, A new finite element formulation for Computational Fluid Dynamics: V. Circumventing the Babaska-Brezzi Condition; A stable Petrov-Galerkin formulation of the Stokes problem accomodating equal-order interpolation, *Comp. Meth. Appl. Mech. Eng.*, **59** (1986), 85–100.

[87] J. Douglas and J. Wang, An absolutely stabilized finite element method for the Stokes problem, *Math. Comp.*, **52** (1989), 495–508.

[88] Z. Cai and J. Douglas, An analytic basis for multigrid methods for stabilized finite element methods for the Stokes problem. In Bristeau, M.O., Etgen, G., Fitzgibbon, W., Lions, J.L., Periaux, J., and Wheeler, M.F. (eds.) *Computational Science for the 21st Century*, Wiley, Chichester, 1997, 113–118.

[89] R. Glowinski and C.H. Li, On the numerical implementation of the Hilbert Uniqueness Method for the exact boundary controllability of the wave equation, *C.R. Acad. Sc., Paris*, t. 311 (1990), Série I, 135-142.

[90] Glowinski, R., C.H. Li, and J.L. Lions, A numerical approach to the exact boundary controllability of the wave equation (I) Dirichlet controls: Description of the numerical methods, *Japan J. Applied Math.*, **7** (1990), 1–76.

[91] R. Glowinski and J.L. Lions, Exact and approximate controllability for distributed parameter systems, Part II, *Acta Numerica 1995*, Cambridge University Press, 1995, 159–333.

[92] M. Crouzeix and P.A. Raviart, Conforming and nonconforming finite element methods for solving the stationary Stokes equations, *Revue Française d'Automatique, Informatique et Recherche Opérationnelle*, **R3** (1973), 33–76.

[93] J.E. Roberts and J.M. Thomas, Mixed and hybrid methods. In Ciarlet, P.G., and Lions, J.L. (eds.) *Handbook of Numerical Analysis*, **Vol. II**, North-Holland, Amsterdam, 1991, 523–639.

[94] R. Verfurth, Error estimates for a mixed finite element approximation of the Stokes problem, *Revue Française d'Automatique, Informatique et Recherche Opérationelle, Anal. Numer.*, **18** (1984), 175–182.

[95] R. Glowinski, T.W. Pan, T.I. Hesla, and D.D. Joseph, A distributed Lagrange multiplier/fictitious domain method for particulate flow, *Int. J. Multiphase Flow*, **25** (1999), 755–794.

[96] R. Glowinski, T.W. Pan, T.I. Hesla, D.D. Joseph, and J. Periaux , A distributed Lagrange multiplier/fictitious domain method for flows around moving rigid bodies: Application to particulate flow, *Int. J. Numer. Meth. in Fluids,* **30** (1999), 1043–1066.

[97] R. Glowinski, T.W. Pan, T.I. Hesla, D.D. Joseph, and J. Periaux, A distributed Lagrange multiplier/fictitious domain method for the simulation of flow around moving rigid bodies: Application to particulate flow, *Comp. Meth. Appl. Mech. Eng.,* **184** (2000), 241–267.

[98] R. Glowinski, T.W. Pan, T.I. Hesla, D.D. Joseph, and J. Periaux, A fictitious domain approach to the direct numerical simulation of incompressible viscous fluid flow past moving rigid bodies: Application to particulate flow, *J. Comp. Phys.,* **169** (2001), 363–426.

[99] S. Turek, A comparative study of time-stepping techniques for the incompressible Navier-Stokes equations: from fully implicit non-linear schemes to semi-implicit projection methods, *Int. J. Num. Math. in Fluids*, **22**(1996), 987-1011.

[100] E.J. Dean, R. Glowinski. A wave equation approach to the numerical solution of the Navier-Stokes equations for incompressible viscous flow, *C.R. Acad. Sci. Paris*, t. 325, Série I, (1997), 783–791.

[101] E. Dean, R. Glowinski, T.-W. Pan. A wave equation approach to the numerical simulation of incompressible viscous fluid flow modeled by the Navier-Stokes equations, in *Mathematical and numerical aspects of wave propagation*, J. De Santo ed., SIAM, Philadelphia, 1998, 65–74.

[102] R. Glowinski, O. Pironneau, Finite Element Methods for Navier-Stokes Equations, *Annu. Rev. Fluid Mech.*, **24**(1992), 167–204.

[103] C. Johnson, Streamline diffusion methods for problems in fluid mechanics, in *Finite Element in Fluids 6*, R. Gallagher ed., Wiley, 1986.

[104] U. Ghia, K.N. Ghia, and C.T. Shin, High-Reynolds solutions for incompressible flow using Navier-Stokes equations and a multigrid method, *J. Comp. Phys.*, **48**(1982), 387–411.

[105] R. Schreiber, H.B. Keller, Driven cavity flow by efficient numerical techniques, *J. Comp. Phys.*, **40**(1983), 310–333.

[106] C.H. Bruneau and C. Jouron, Un nouveau schéma décentré pour le probléme de la cavité entraînée, *C.R. Acad. Sci. Paris*, t. 307, Série I (1988), 359–362.

[107] J. Shen, Hopf bifurcation of the unsteady regularized driven cavity flow, *J. Comp. Phys.*, **95**(1991), 228–245.

[108] O. Goyon, High-Reynolds number solutions of Navier-Stokes equations using incremental unknowns, *Comput. Methods Appl. Mech. Engrg.*, **130** (1996), 319–335.

[109] S. Fujima, M. Tabata, Y. Fukasawa, Extension to three-dimensional problems of the upwind finite element scheme based on the choice up- and downwind points, Comp. Meth. Appl. Mech. Eng., **112**(1994), 109–131.

[110] H.C. Ku, R.S. Hirsh, T.D. Taylor, A pseudospectral method for solution of the three-dimensional incompressible Navier-Stokes equations, *J. Comp. Phys.*, **70**(1987), 439–462.

[111] T.P. Chiang, W.H. Sheu, R.R. Hwang, Effect of Reynolds number on the eddy structure in a lid–driven cavity, *Int. J. Numer. Meth. in Fluids,* **26**(1998), 557–579.

[112] A.K. Prasad and J.R. Koseff, Renolds number and end-wall efftects on a lid-driven cavity flow, *Phys. Fluids*, **A 1** (1989), 208–218.

Bol. Soc. Esp. Mat. Apl. n°36(2006), 63–86

Solutions faibles et équations de Navier-Stokes. De Jean Leray à Pierre–Louis Lions.

Dédié à la naissance de la petite Isabelle Desjardins

Didier Bresch

LAMA, UMR5127, Université de Savoie, 73376 Le Bourget du Lac Cedex, France.

didier.bresch@univ-savoie.fr

Résumé

Ce papier est un très rapide survol au dessus des équations de Navier-Stokes incompressible ou compressible. Le but est de présenter, succintement, différents résultats d'existence globale de solutions faibles à la Leray connus sur ces systèmes tout en esquissant les méthodes qui ont permis d'y parvenir. Nous donnerons également quelques problèmes ouverts, relatifs à l'existence globale de solutions faibles pour certains fluides compressibles. Le but est de permettre aux lecteurs néophites d'aborder plus facilement les ouvrages spécialisés dans leur partie solutions faibles et de comprendre les difficultés spécifiques aux équations régissant un fluide compressible.

Mots clefs : Fluides incompressibles ou compressibles, solutions faibles.
AMS subject classification : 35Q30.

1 Introduction

En 1755, Euler écrit le premier modèle pour un fluide : un gaz compressible décrit par sa densité et sa vitesse $v = v(t,x)$ $(\in R^n)$ et sa pression $p = p(t,x)$ $(\in R)$. Les équations d'Euler compressible sont données par

$$\partial_t \rho + \mathrm{div}(\rho u) = 0, \tag{1}$$

$$\partial_t(\rho u) + \mathrm{div}(\rho u \otimes u) + \nabla p = 0, \tag{2}$$

Fecha de recepción: 6/12/2005

pour un fluide barotrope, p est une fonction donnée de ρ comme $p = a\rho^\gamma$ avec $\gamma > 1$ et $a > 0$. Les équations d'Euler incompressible sont déduite pour ρ constante en laissant π inconnue où $\nabla\pi \equiv a\nabla\rho^\gamma$. Si $a \to +\infty$ c'est-à-dire le nombre de Mach tend vers 0 alors ρ devient constant et $a\nabla\rho^\gamma \to \nabla\pi$.

$$\partial_t u + u\cdot\nabla u + \nabla\pi = 0, \tag{3}$$
$$\mathrm{div} u = 0. \tag{4}$$

En 1822, Navier (1785–1836) obtient un modèle qui prend en compte les effets de viscosité pour un fluide newtonien. Ce modèle a été étudié plus tard par Stokes (1819–1903) qui justifiera (physiquement) en 1849 l'ajout dans les équations de quantité de mouvement du terme $-\mu\Delta u - (\lambda+\mu)\nabla\mathrm{div} u$ donnant

$$\partial_t\rho + \mathrm{div}(\rho u) = 0, \tag{5}$$
$$\partial_t(\rho u) + \mathrm{div}(\rho u\otimes u) - \mu\Delta u - (\lambda+\mu)\nabla\mathrm{div} u + \nabla p = 0, \tag{6}$$

et aux équations de Navier-Stokes incompressible

$$\partial_t u - \nu\Delta u + \mathrm{div}(u\otimes u) + \nabla\pi = 0, \qquad \mathrm{div} u = 0. \tag{7}$$

où ν, μ, λ sont des constantes satisfaisant $\nu > 0$, $\mu > 0$ et $\lambda + 2\mu > 0$.

Nous nous intéresserons d'ailleurs en particulier à ces deux derniers systèmes dans cet article. Nous considérerons le cas d'un domaine borné Lipschitzien en dimension 3 d'espace par souci de simplicité et nous ne chercherons pas à donner des résultats optimo.

Plus précisemment, nous allons montrer le chemin parcouru sur ce que l'on appelle maintenant les solutions faibles à la Leray des équations de Navier–Stokes incompressible à densité constante aux équations de Navier-Stokes compressible. Nous exposerons également des résultats récents concernant le modèle complet : Navier-Stokes compressible avec conduction de chaleur. Comprendre ce chemin permet de voir les spécificités de chaque système en vu, notamment, d'améliorer les résultats encore incomplets sur Navier-Stokes compressible.

Le lecteur intéressé par un article de revue sur les résultats de base sur les équations de Navier-Stokes incompressible, et d'autres équations similaires comme certaines équations de type fluides non newtoniens, est renvoyé notamment à [23]. Le lecteur intéressé par un article de revue sur les résultats de base sur les équations de Navier-Stokes compressible est renvoyé, lui, notamment à [18] et [21]. En ce qui concerne les ouvrages classiques sur le sujet, nous renvoyons le lecteur à [40], [29] et [4] pour les fluides incompressibles et à [29], [20], [36] pour les fluides compressibles.

Il y a bien sûr d'autres systèmes en mécanique des fluides comme par exemple les fluides non newtoniens, fluides réactifs, interaction fluides/structures, modèles météorologiques, magnétohydrodynamique, écoulements multiphasiques, problèmes à frontière libre pour n'en citer que quelques uns. Nous ne rentrerons pas dans une telle discussion. Nous tenons également à préciser aux lecteurs que l'exposé qu'il trouve ici est loin d'être exhaustif. Comme cela est

écrit dans le résumé, il s'agit d'un très rapide survol au dessus des équations de Navier-Stokes incompressible et compressible. La littérature est assez vaste sur le sujet....

2 De J. Leray à P.–L. Lions *via* J. Simon.

2.1 Navier-Stokes incompressible homogène

Les équations de Navier–Stokes constituent un modèle mathématique de base pour décrire le mouvement d'un fluide incompressible. Dans le célèbre papier publié dans *Acta Mathematica* en 1934, Sur le mouvement d'un fluide visqueux emplissant l'espace, Jean Leray (1906–1998) montre l'existence d'une solution régulière jusqu'à un temps T qu'il caractérise. Il introduit, également, le concept de solution faible (et pour ce faire, il définit d'ailleurs ce que l'on appelle maintenant un espace de Sobolev), en donnant une définition précise de ce qu'est une solution irrégulière du système, et montre qu'il existe une telle solution faible sur Navier-Stokes. On appelle maintenant ces solutions de régularité minimale (énergie finie) : solutions à la Leray. Un théorème d'unicité fort-faible montre également que s'il existe une solution dans le sens classique et une solution faible alors ces deux coincident. Le lecteur intéressé par les divers travaux pionniers de J. Leray, en mécanique des fluides, est renvoyé à [27] ainsi quà l'article [13].

Même si l'existence globale de solutions faibles apporte assez peu sur le caractère bien posé du système, une telle analyse a pourtant de nombreux intérêts pratiques. En plus de la signification physique, car la régularité des données initiales supposée est fortement liée à des quantités physiques bien identifiée, les propriétés de stabilité de solutions faibles entraînent la stabilité des schémas numériques, qui le plus souvent ne préservent pas les estimations de régularité forte sur leur limite continue.

Nous allons décrire, ici, succinctement la méthode maintenant classique de preuve d'existence globale de solutions faibles. Elle est basée principalement sur une approximation de type Galerkin qui consiste à définir des solutions approchées en dimension finie d'espace, prouver des estimations sur les solutions de ce problème approché qui soient uniformes puis passer à la limite par un argument de compacité. La formulation faible utilisée ne fournit pas directement la pression, il faut alors la retrouver indirectement à la fin de la preuve (Lemme de De Rham). Pour passer à la limite dans le terme non linéaire $u_m \cdot \nabla u_m$, on utilise le fait que la suite $\{u_m\}_{m\in N}$ est bornée dans $L^\infty(0,T;H) \cap L^2(0,T;V)$ par l'estimation d'énergie et que $\{\partial_t u_m\}_{m\in N}$ est bornée dans $L^{4/3}(0,T;V')$ en utilisant la formulation faible où apparaît $\partial_t u_m$ et les informations de régularié faible disponibles. On note $V = \{v \in (H_0^1(\Omega))^3 : \mathrm{div} u = 0\}$ et $H = \{v \in (L^2(\Omega))^3 : \mathrm{div} v = 0, v \cdot n_{|\partial\Omega} = 0\}$.

On peut alors établir le résultat d'existence globale de solutions faibles, en domaine borné, suivant

Théorème 2.1 *Soit Ω un ouvert lipschitzien borné de R^3, avec $u_0 \in H$ et la*

force extérieure $f \in L^2(0,T;H^{-1}(\Omega))$ *alors il existe un couple* (u,p) *tel que*

$$u \in L^2(0,T;V) \cap L^\infty(0,T;H) \cap \mathcal{C}([0,T];L^2(\Omega) \text{ } \textit{faible}), \tag{8}$$

$$p \in W^{-1,\infty}(0,T;L^2_{\rm loc}(\Omega)), \qquad \nabla p \in W^{-1,1}((0,T)\times\Omega) \tag{9}$$

et

$$\partial_t u - \nu\Delta u + u\cdot\nabla u + \nabla p = f, \tag{10}$$

$$\mathrm{div} u = 0, \tag{11}$$

$$u|_{t=0} = u_0, \qquad u|_{\partial\Omega\times(0,T)} = 0. \tag{12}$$

De plus,

$$\frac{1}{2}\|u(t)\|^2_{L^2(\Omega)} + \nu\int_0^t \|\nabla u(\tau)\|^2_{L^2(\Omega)}\,d\tau \le \frac{1}{2}\|u_0\|^2_{L^2(\Omega)} + \int_0^t < f, v >_{H^{-1}(\Omega)\times H^1_0(\Omega)}\,d\tau.$$

2.2 Navier-Stokes incompressible non homogène

Les premiers résultats d'existence globale de solutions faibles pour les équations de Navier-Stokes non homogène, incompressible ont été obtenus par Alexandre V. Kazhikhov (1947-2005) dans le cas μ indépendant de ρ et ρ_0 loin de zéro. Ces résultats furent étendus par Jacques Simon (1947–··) en permettant à ρ_0 de s'annuler en supposant μ constant. Le cas où μ dépend de ρ a été étudié par la suite en supposant certaines propriétés sur μ, voir par exemple [28]. Le terme de diffusion s'écrit alors $-2\mathrm{div}(\mu(\rho)D(u))$ où $D(u) = (\nabla u + {}^t\nabla u)/2$ au lieu de $-\mu\Delta u$ dans les équations de Navier-Stokes incompressible.

Donnons ici, le résultat d'existence globale de solutions faibles dans le cas où μ est indépendant de ρ provenant de [38]. Plus précisemment, le résultat d'existence globale à la Leray est le suivant

Théorème 2.2 *Soit* $\Omega \subset R^3$ *un domaine Lipschitzien borné de* R^3 *et* $T > 0$. *Si* $u_0 \in H$, $\rho_0 \in L^\infty(\Omega)$ *avec* $\rho_0 \ge 0$ *et la force extérieure* $f \in L^1(0,T;L^2(\Omega))$ *alors il existe*

$$u \in L^2(0,T;V), \qquad p \in W^{-1,\infty}(0,T;L^2(\Omega)), \qquad \rho \in L^\infty((0,T)\times\Omega)$$

tel que

$$\rho u \in L^\infty(0,T;L^2(\Omega)) \cap N^{1/4,2}(0,T;W^{-1,3}(\Omega))$$

où le dernier espace est un espace de type Nikolskii,

$$\inf \rho_0 \le \rho \le \sup \rho_0$$

et

$$\partial_t(\rho u) + \mathrm{div}(\rho u\otimes u) + \nabla p - \nu\Delta u = \rho f,$$

$$\partial_t \rho + \mathrm{div}(\rho u) = 0 \tag{13}$$

$$\mathrm{div} u = 0 \tag{14}$$

avec

$$u|_{\partial\Omega\times(0,T)} = 0, \qquad \rho|_{t=0} = \rho_0, \qquad (\int_\Omega \rho u\cdot v)|_{t=0} = \int_\Omega \rho_0 u_0\cdot v \text{ pour tout } v \in V.$$

Le schéma de preuve est également principalement basé sur une approximation de type Galerkin qui consiste à définir des solutions approchées en dimension finie d'espace, prouver des estimations sur les solutions de ce problème approché qui soient uniformes puis passer à la limite par un argument de compacité. Dans cette construction, on pose comme condition initiale $\rho_m^0 = \rho_0 + 1/m$. Pour passer à la limite dans le terme non linéaire, on ne dispose plus de dérivée en temps sur u_m mais seulement sur $\rho_m u_m$. On sait également que $\rho_m u_m$ est bornée uniformément dans $L^\infty(0,T;(L^2(\Omega))^3) \cap L^2(0,T;(L^6(\Omega))^3)$. L'idée cruciale est alors d'essayer d'obtenir des informations de la forme

$$\|\tau_h(\rho_m u_m) - \rho_m u_m\|_{L^2(0,T;(W^{-1,3}(\Omega))^3)} \leq ch^{1/4}$$

où $\tau_h v = v(t+h)$. Ce qui donne que $\rho_m u_m$ est bornée dans l'espace de Nikolskii $N^{1/4,2}(0,T;(W^{-1,3}(\Omega))^3))$. Par lemme de compacité, *cf.* [38]–[39], cette estimation donnera la compacité de $\rho_m u_m$ dans $L^2(0,T;(W^{-1,\infty}(\Omega))^3$. Cette information suffira pour passer à la limite dans le terme non linéaire $\rho_m u_m \otimes u_m$ dans $L^1(0,T;(W^{-1,6}(\Omega))^9)$. Notons que J. Simon est parvenu à une telle estimation par une utilisation du théorème de Fubini que n'avait pas vu, par exemple, A. Kazhikhov. C'est cette étape qui avait alors contraint A. Kazhikhov à supposer $\rho_0 \geq c > 0$, *cf.* [26]. Notons également la condition initiale sur ρu satisfaite en un sens faible.

Nous terminerons cette partie en mentionant que les travaux de J. Simon ont été réalisés à la même période que les travaux de R.J. DiPerna et P.–L. Lions sur les équations de transport. Ces derniers simplifient énormément les choses et permettent par exemple d'obtenir rapidement la compacité forte sur $\rho_m u_m$ une fois obtenue l'estimation sur $\partial_t(\rho_m u_m)$. On peux également considérer plus aisément le cas de viscosité μ dépendant de ρ, voir par exemple [28].

2.3 Navier-Stokes compressible barotrope

Lorsque l'écoulement est compressible et barotrope, même l'existence de solutions faibles en dimension d'espace supérieure ou égale à deux est longtemps resté sans réponse. Il existe une vaste littérature sur cette question dans laquelle de nombreux auteurs, A. Matsumura et T. Nishida, D. Hoff, D. Serre, A.V. Weigant et A. Kazhikhov pour n'en citer que quelques uns, ont apporté des réponses partielles sous diverses contraintes plus ou moins restrictives sur les données initiales (régularité, petitesse) ou sur les coefficients de viscosité (dépendance très particulière de λ par rapport à densité).

La première approche rigoureuse de ce problème dans toute sa généralité est due en 1993 à Pierre–Louis Lions (1956–..) dans le cas des équations de Navier-Stokes compressible en régime isentropique (*i.e.* lorsque la loi d'état du fluide reliant la pression p à la densité ρ est du type $p = a\rho^\gamma$ où a est

une constante strictement positive et γ est la constante obtenue par le rapport entre la chaleur spécifique à pression constante et la chaleur spécifique à volume constant, constante dite adiabatique). Dans ses travaux, Pierre–Louis Lions a présenté une théorie complète permettant d'obtenir des résultats d'existence de solutions faibles globales en n dimensions d'espace ($n \geq 2$) et ce pour des données initiales générales. Notons que les ingrédients de P.–L. Lions puisent leurs sources dans les travaux de D. Hoff et D. Serre sur l'importance du flux effectif sur la stabilité, les travaux de R. Coifman, Y. Meyer pour les propriétés de régularisatés liés aux commutateurs et les travaux de R.J. Di-Perna et P.-L. Lions sur les propriétés de l'équation de transport.

Nous allons ici nous inspirer de [36] pour présenter dans une approche heuristique les arguments de compacité de P.–L. Lions et les comparer à ceux d'E. Feireisl qui a amélioré les puissances de γ considérées. Notons que pour la construction des solutions approchées, une régularisation parabolique de l'équation de la masse et un rajout de terme de pression sont effectués. Nous en esquisserons le système en fin de sous-section.

Les équations de Navier—Stokes, modélisant l'évolution temporelle de la densité ρ et de la vitesse u d'un fluide compressible en régime isentropique occupant une région bornée tridimensionnelle Ω, s'écrivent

$$\partial_t \rho + \mathrm{div}(\rho u) = 0, \tag{15}$$

$$\partial_t(\rho u) + \mathrm{div}(\rho u \otimes u) - \mu\Delta u - (\lambda+\mu)\nabla \mathrm{div} u + a\nabla\rho^\gamma = \rho f. \tag{16}$$

La première équation, communément appelée équation de continuité, provient du principe de conservation de la masse, tandis que la deuxième équation, communément appelée équation de mouvement, provient à elle, du principe de conservation de la quantité de mouvement.

Lorsque ρ et u sont régulières et satisfont l'équation de continuité, pour toute fonction $b \in \mathcal{C}^1([0,\infty))$, il est clair qu'elles sont également solutions de l'équation de continuité dite renormalisée, cette terminologie provenant de la théorie du transport de R.J. Diperna et P.–L. Lions. Cette équation s'écrit

$$\partial_t b(\rho) + \mathrm{div}(b(\rho)u) + (b'(\rho)\rho - b(\rho))\mathrm{div} u = 0. \tag{17}$$

Pour un temps $T \in (0,\infty)$, des forces f et des données initiales ρ_0 et m_0 satisfaisant certaines hypothèses techniques, on dira que le couple de fonctions (ρ, u) est une solution faible renormalisée à énergie bornée des équations s'il possède les propriété suivantes : $\rho \in L^\infty(0,T;L^\gamma(\Omega)) \cap \mathcal{C}^0([0,T], L^\gamma_{\text{faible}}(\Omega))$, $\rho \geq 0$ p.p. dans Ω, $\rho|_{t=0} = \rho_0$ p.p. dans Ω, $u \in L^2(0,T;H^1_0(\Omega))$, $\rho|u|^2 \in L^\infty(0,T;L^1(\Omega))$ et $\rho u \in \mathcal{C}^0([0,T];L^{2\gamma/(\gamma+1)}(\Omega)_{\text{faible}})$ et $(\rho u)|_{t=0} = m_0$ p.p. dans Ω ; si le prolongement par zéro de (ρ, u) dans $(0,T)\times R^3\backslash\Omega$ est solution dans $\mathcal{D}'((0,T)\times\Omega)$; si p.p.t. $\tau \in (0,T)$, (ρ, u) satisfait l'inégalité d'énergie

$$E(\rho,u)(\tau) + \int_0^\tau \int_\Omega (\mu|\nabla u|^2 + (\lambda+\mu)|\mathrm{div} u|^2) \leq E_0 + \int_0^\tau \int_\Omega \rho f \cdot u$$

et $b(\rho)$ satisfait (17) pour b avec certaines propriétés de croissance. Dans cette inégalité, $E(\rho,u)(\tau) = \big(\int_\Omega \rho|u|^2/2 + a\rho^\gamma/(\gamma-1)\big)(\tau)$ désigne l'énergie totale au temps τ et $E_0 = \int_\Omega |m_0|^2/2\rho_0 + a\rho_0^\gamma/(\gamma-1)$ désigne l'énergie totale initiale.

La théorie développée par P.–L. Lions pour démontrer l'existence de solutions faibles renormalisées à énergie bornée fait apparaître une limitation sur les valeurs autorisées pour la constante adiabatique γ, à savoir $\gamma \geq 9/5$ en dimension 3. Récemment E. Feireisl a généralisé cette approche pour pouvoir traiter les valeurs $\gamma > 3/2$ en dimension 3 et plus généralement $\gamma > n/2$ où n est la dimension d'espace. Nous allons ici nous borner à faire comprendre les différentes lignes.

La technique est de construire une suite de solutions approchées sur un système proche de celui considéré par théorèmes de point fixe, approximations de type Faedo-Galerkin. Ceci se fait en introduisant un, voir plusieurs paramètres, puis en modifiant le système d'origine de telle sorte que les solutions approchées du système obtenu tendent vers une solution du système d'origine lorsque le, voire les paramètres en question tendent vers leur valeur critique. Cette méthode fait continuellement apparaître le problème de compacité d'un ensemble borné de solutions approchées.

Notons qu'avec un bon choix de multiplicateur, que pour $\gamma > d/2$, il est possible d'établir l'estimation clef suivante sur la densité

$$\int_0^\infty dt \int_\Omega \rho^q \leq C(R,T) \text{ pour } q = (1+\frac{2}{n})\gamma - 1.$$

Théorème 2.3 *Si $\gamma > n/2$, si $\rho_0 \in L^\gamma(\Omega)$ et $\frac{|m_0|}{\rho_0} \in L^1(\Omega)$ alors il existe une solution globale faible (ρ, u) de Navier-Stokes compressible barotrope.*

Le couple (ρ^n, u^n) satisfaisant l'estimation

$$\rho^n \to \rho \text{ dans } L^\gamma \text{ faible }, \qquad u^n \to u \text{ dans } L^2(0,T;H^1_0(\Omega)) \text{ faible }.$$

Le rajout de la viscosité $-\nu\Delta u - (\lambda+\mu)\nabla \text{div} u$ implique de la régularité en espace. En fait, il n'y a donc plus de phénomène de compactification et (ρ, u) peut ne pas être solution. La viscosité régularise la vitesse et donc empêche les chocs et donc préserve les oscillations en densité qui peuvent se propager. Malgré tout si $\rho_0^n \to \rho_0$ dans $L^1(\Omega)$ alors

$$\rho^n \to \rho \text{ dans } \mathcal{C}([0,T];L^1(\Omega)) \text{ pour tout } T \in (0,\infty).$$

On peux montrer que la différence entre convergence forte et convergence faible décroit en temps.

$$\beta(\rho^n)[(\lambda+2\mu)\text{div}u^n - a(\rho^n)^\gamma] \to \overline{\beta}[(\lambda+2\mu)\text{div}u - a\overline{\rho^\gamma}]$$

et β est une fonction $\mathcal{C}^0$ arbitraire sur $[0,\infty)$ avec une croissance à l'infini suffisamment faible.

Tout d'abord, de l'inégalité d'énergie dans laquelle ρ_n, u_n, $E_{0,n}$ et 0 remplacent ρ, u, E_0 et f et g, il résulte que

$$\|\rho_n\|_{L^\infty(0,T;L^\gamma(\Omega))} + \|u_n\|_{L^2(0,T;H^1_0(\Omega))} + \|\rho_n|u_n|^2\|_{L^\infty(0,T;L^1(\Omega))} \leq c(T, E_{0,n}).$$

En accord avec ces bornes, il existe des fonctions ρ et u telles que, modulo l'extraction de sous-suites et lorque $n \to +\infty$, on a

$$\rho_n \to \rho \text{ dans } L^\infty(0,T;L^\gamma(\Omega)) \text{ faible}*, \qquad u_n \to u \text{ dans } L^2(0,T;H^1_0(\Omega)) \text{ faible}.$$

Ceci rends possible le passage à la limite $n \to +\infty$ dans l'équation de continuité et dans tous les termes qui apparaîssent dans l'équation de mouvement, excepté dans le terme de pression $P(\rho_n) = a\rho_n^\gamma$. En effet, pour s'assurer que celui converge vers $P(\rho) = a\rho^\gamma$, il est nécessaire d'avoir plus d'informations, comme par exemple la convergence $\rho_n \to \rho$ dans $L^1((0,T)\times\Omega)$.

Dans un premier temps, pour éviter que la limite de la suite $a\rho_n^\gamma$ ne soit une mesure, il est intéressant de posséder davantage d'information sur ρ_n. Pour cela on teste formellement l'équation de quantité de mouvement par

$$\varphi = \mathcal{B}(\rho_n^\theta - \int \rho_n^\theta)$$

où $\mathcal{B}$ est l'opérateur de Bogovskii sur Ω (un inverse de l'opérateur de divergence) et $0 < \theta < \gamma$. Après calculs, on obtient

$$\int_0^T \int_\Omega \rho_n^{\gamma+\theta} \leq C(T,\Omega,E_{0,n}), \qquad \theta = \frac{2}{3}\gamma - 1. \tag{18}$$

Cette observation est due à P.–L. Lions qui l'a obtenue d'une autre manière. Il faut noter que pour la démontrer, on a besoin de savoir que (ρ_n, u_n) est une solution de l'équation de continuité renormalisée dans laquelle $b(s) = s^\theta$.

Noter également que l'hypothèse de régularité de la région Ω est étroitement liée à l'utilisation de l'opérateur de Bogovskii qui n'est pas défini si Ω ne possède pas la régularité minimale d'avoir une frontière Lipschitzienne. Évidemment, une estimation locale du type précédent suffit pour assurer que la pression limite n'est pas une mesure. À l'aide de (18), il est alors possible de préciser certaines convergences, notamment que modulo l'extraction de sous-suites et lorsque $n \to +\infty$, on a

$$\rho_n \to \rho \text{ dans } \mathcal{C}^0([0,T], L^\gamma_{\text{faible}}(\Omega)), \tag{19}$$

$$\rho_n^\gamma \to \overline{\rho^\gamma} \text{ dans } L^{(\gamma+\theta)/\gamma}((0,T)\times\Omega) \text{ faible}, \tag{20}$$

$$\rho_n u_n \to \rho u \text{ dans } (\mathcal{C}^0([0,T]; L^{2\gamma/(\gamma+1)}_{\text{faible}}(\Omega)), \tag{21}$$

$$\rho_n u_n^i u_n^j \to \rho u^i u^j \text{ dans } \mathcal{D}'((0,T)\times\Omega), \quad i,j = 1,2,3. \tag{22}$$

Notons la convergence des termes non linéaires qui proviennent de la convergence forte de ρ_n déduite notamment de l'estimation uniforme sur $\partial_t \rho_n$ donnée par l'équation de la masse et de la convergence forte de $\sqrt{\rho}_n u_n$ déduite notamment de l'estimation uniforme sur $\partial_t(\rho_n u_n)$ donnée par l'équation de quantité de mouvement. En conséquence, les prolongements par zéro dans $(0,T)\times R^3/\Omega$ des fonctions ρ, u et $\overline{\rho^\gamma}$ encore notés ρ, u et $\overline{\rho^\gamma}$ satisfont les

équations

$$\partial_t \rho + \mathrm{div}(\rho u) + 0 \text{ dans } \mathcal{D}'((0,T)\times R^3), \tag{23}$$

$$\partial_t(\rho u) + \mathrm{div}(\rho u \otimes u) - \mu\Delta u - (\lambda+\mu)\nabla \mathrm{div} u + a\nabla\overline{\rho^\gamma} = 0 \text{ dans } \mathcal{D}'((0,T)\times\Omega). \tag{24}$$

La difficulté qui consiste à prouver que (ρ, u) est une solution faible renormalisée à énergie bornée reste entière et réside principalement dans la démonstration de $\overline{\rho^\gamma} = \rho^\gamma$ p.p. dans $(0,T)\times\Omega$). Ceci recquiere une forme de compacité de la suite des densités $\{\rho_n\}_{n\in N^*}$ qui n'est pas disponible au vu des seules estimations. C'est une observation de P.–L. Lions qui va comblée cette lacune. Celle-ci fait état de la chose suivante : la suite des quantités $\{a\rho_n^\gamma - (\lambda+2\mu)\mathrm{div}u_n\}_{n\in N^*}$, couramment appelée suite des flux effectifs visqueux ou suite des pressions effectives, possède une certaine forme de compacité faible : propriété identifiée antérieurement en dimension 1 par D. Hoff et D. Serre.

Plus exactement, on a le théorème suivant

Théorème 2.4 *Pour toute fonction $b \in \mathcal{C}^1([0,\infty))$ satisfaisant certaines conditions de croissance à l'infini, on a*

$$\lim_{n\to+\infty} \int_0^T \int_\Omega (a\rho_n^\gamma - (2\mu+\lambda)\mathrm{div}u_n) b(\rho_n)\varphi \, dxdt = \int_0^T \int_\Omega (a\overline{\rho^\gamma} - (2\mu+\lambda)\mathrm{div}u)\overline{b(\rho)}\varphi \, dxdt \tag{25}$$

où les quantités surmontées d'une barre désignent les limites faibles des suites correspondantes.

Jusque là, rien ne différencie réellement les approches de P.–L. Lions et E. Feireisl. En effet, ce dernier dans son approche aura besoin d'une version semblable au théorème précédent

Théorème 2.5 *Pour toute fonction $b \in \mathcal{C}^1([0,\infty))$ et tout $k > 0$, si on dénote par b_k la fonction définie par $b_k(s) = b(s)$ si $s \in [0,k)$ et $b_k(s) = b(k)$ si $s \in [k,+\infty)$, alors*

$$\lim_{n\to\infty} \int_0^T \int_\Omega (a\rho_n^\gamma - (\lambda+2\mu)\mathrm{div}u_n) b_k(\rho_n)\varphi \, dxdt = \int_0^T \int_\Omega (a\overline{\rho^\gamma} - (\lambda+2\mu)\mathrm{div}u)\overline{b_k(\rho)}\varphi \, dxdt \tag{26}$$

for all $\varphi \in \mathcal{D}((0,T)\times\Omega)$.

Que ce soit P.–L. Lions ou E. Feireisl, tous deux ont ensuite respectivement utilisé cette compacité faible avec $b(s) = s$. En réalité, le premier a avoir fait cette observation, à l'aide d'une écriture de chaque intégrale faisant apparaître des commutateurs, est D. Serre en dimension un d'espace. En dimension

supérieure d'espace, la preuve de P.–L. Lions est basée sur des résultats d'analyse harmonique de R. Coifman et Y. Meyer et tient compte des observations de D. Serre. La preuve de E. Feireisl est, quant à elle, basée sur le lemme div-rot de la théorie de la compacité par compensation introduite par F. Murat et L. Tartar et sur l'écriture de chaque intégrale faisant apparaître un autre commutateur.

Voyons tout d'abord comment P.–L. Lions a conclu à la convergence forte de la suite des densités $\{\rho_n\}_{n\in N^*}$ vers ρ. Pour somplifier la présentation, supposons $s(\gamma) > \gamma + 1$ c'est-à-dire $\gamma > 3$. Au vu de l'estimation sur ρ, il est clair que $\rho \in L^2((0,T)\times\Omega)$ si $\gamma \geq 9/5$. Dans ce cas, la théorie du transport de R.J. DiPerna et P.–L. Lions s'applique à l'équation de continuité pour garantir la validit{e de l'équation renormalisée. Alors si $b(s) = s\ln s$, le passage à la limite $n \to +\infty$ dans l'équation satisfaite par ρ_n, u_n, l'équation renormalisée et l'identité de compacité faible permettent l'obtention d'une équation d'évolution pour l'amplitude des oscillations de la suite des densités mesurées par $\{\overline{\rho\ln\rho} - \rho\ln\rho\}$. Celle-ci s'écrit

$$\begin{aligned}\partial_t(\overline{\rho\ln\rho} - \rho\ln\rho) \qquad &+\mathrm{div}((\overline{\rho\ln\rho} - \rho\ln\rho)u)\\ &= \frac{a}{2\mu+\lambda}(\overline{\rho^\gamma}\rho - \overline{\rho^{\gamma+1}}) \text{ dans } \mathcal{D}'((0,T)\times R^3). \quad (27)\end{aligned}$$

L'intégration formelle de cette équation sur $(0,T)\times\Omega$, la monotonie de la pression $p(\rho) = a\rho^\gamma$ et la convexité de la fonction $s \mapsto s\ln s$, $s \geq 0$, impliquent que $\overline{\rho\ln\rho} = \rho\ln\rho$ p.p. dans $(0,T)\times\Omega$. Autrement dit, la convergence faible commute avec la fonction strictement convexe, ce qui est totalement équivalent à la convergence forte de la suite des densités $\{\rho_n\}_{n\in N^*}$ vers ρ dans $L^1((0,T)\times\Omega)$ et achève la preuve du théorème dans le cas $\gamma > 3$.

Voyons maintenant où résident les améliorations de E. Feireisl permettant de traiter les valeurs γ pour lesquelles $\gamma < s(\gamma) \leq \gamma + 1$, *i.e.* $\gamma \in (3/2, 3]$. Pour cet intervalle de valeurs, on n'est pas toujours assuré que ρ soit de carré intégrable. En conséquence, il n'est pas possible d'appliquer directement la théorie du transport de R.J. DiPerna et P.–L. Lions. Ce manque d'information, E. Feireisl l'a comblé en remarquant qu'il était possible de contrôler l'amplitude des oscillations possibles en densité pour une norme L^p, avec p supérieur à 2, par le résultat suivant

Théorème 2.6 *Considérons $\{\rho_n\}_{n\in N}$ la suite de solutions approchées et ρ sa limite faible alors*

$$\mathrm{osc}_{\gamma+1}[\rho_n - \rho] = \sup_{k>0}\limsup_{n\to+\infty} \|T_k(\rho_n) - T_k(\rho)\|_{L^{\gamma+1}(\Omega)} \leq c(T,\Omega,E_{0,n})$$

où $T_k(z) = kT(z/k)$ pour $k \geq 1$ avec $T \in \mathcal{C}^1(\mathbb{R})$ une fonction paire telle que $T(z) = z$ pour $0 \leq z \leq 1$, $T(z) = 2$ pour $z \geq 3$ et T concave sur $[0,\infty)$.

La démonstration de ce résultat utilise le théorème de compacité faible. Noter également que celle-ci ne nécessite en aucun cas l'estimation sur les densités. Ce résultat doit être perçu de la manière suivante : bien que les

valeurs $\gamma \in (3/2, 9/5)$, même si la suite des densités $\{\rho_n\}_{n\in N^*}$ est seulement bornée dans $L^{\gamma+\theta}((0,T)\times\Omega)$, l'amplitude de ses oscillations, mesurée par le membre de gauche du théorème précédent l'est toujours dans un espace meilleur que $L^2((0,T)\times\Omega)$. Cette propriété va ensuite se substituer à l'information manquante $\rho \in L^2((0,T)\times\Omega)$ et permettre de démontrer, en particulier, que, pour les valeurs $\gamma \in (3/2, 9/5)$, le couple (ρ, u) est une solution de l'équation renormalisée sur la masse.

Pour conclure E. Feireisl a adapté les arguments de la fin d'approche de P.–L. Lions présenté plus haut. Disons simplement qu'il a utilisé une fonction $b(s) = L_k(s)$ telle que $sL_k'(s) - L_k(s) = \underline{T_k(s)}$ et $L_k(s) \to s\ln s$ quand $k \to \infty$ à la place de $b(s) = s\ln s$ pour récupérer $\overline{\rho\ln\rho} = \rho\ln\rho$ p.p. dans $(0,T)\times\Omega$. D'où la fin de la preuve.

Donnons maintenant à titre indicatif le modèle approché nécessaire à la construction de la suite de solutions. Il s'écrit ainsi :

$$\partial_t\rho + \mathrm{div}(\rho u) - \varepsilon\Delta\rho = 0, \tag{28}$$

$$\partial_t(\rho u) + \mathrm{div}(\rho u\otimes u) - \mu\Delta u - (\mu+\lambda)\nabla\mathrm{div}u + a\nabla\rho^\gamma + \delta\nabla\rho^\beta + \varepsilon\nabla\rho\cdot\nabla u = \rho f. \tag{29}$$

avec β suffisamment grand et avec comme condition aux bords supplémentaire $\partial_n\rho = 0$ sur la densité.

Les étapes de démonstration sont alors les suivantes : Existence globale de solutions faibles sur le modèle ci-dessus par méthode de Galerkin, passage à la limite quand ε tend vers 0 puis passage à la limite quand δ tend vers 0. Notons que l'on a ajouté le terme $a\nabla\rho^\beta$, pour β grand, dans le système approché pour effectuer le passage à la limite dans le terme $\varepsilon\nabla\rho\cdot\nabla u$ dans l'approximation de Galerkin. Le premier passage à la limite $\varepsilon \to 0$ utilisera les arguments de P.–L. Lions car la densité sera alors suffisamment intégrable alors que la dernière étape nécessitera le travail d'E. Feireisl dans le cas où $3/2 < \gamma < 9/5$.

Remarquons que la construction de solutions faibles pour Navier-Stokes compressible barotrope utilise fortement l'approximation par une suite de solutions telles que $(\rho_n, m_n)|_{t=0} = (\rho_0^n, m_0^n)$. La suite de densité initiale $\{\rho_0^n\}_n$ est construite telle qu'elle converge fortement dans $L^1(\Omega)$ vers la donnée initiale de départ ρ_0.

Il est important de savoir qu'une suite de solutions pour lesquelles les densités oscilleraient de plus en plus fort pourraît converger vers une paire (ρ, u) qui ne soit pas solution de Navier-Stokes compressible classique. L'étude des effets d'oscillations en densité a, par exemple, été étudié rigoureusement par D. Serre, *cf.* [37], dans le cas unidimensionnel. Le cas multidimensionnel ayant été étudié formellement. À la lumière des travaux de P.–L. Lions et E. Feireisl, M. Hillairet, *cf.* [25], a récemment donné une preuve rigoureuse à ces calculs formels multidimensionnels en étudiant le problème de Cauchy associé au système homogénéisé qui en résulte.

Pour finir cette section, nous indiquons de nouveau le très bon livre de A. Novotny et I. Straskarba, *cf.* [36] pour un exposé assez complet sur la théorie mathématique des écoulements compressibles dans le cas isentropique. On y

trouvera notamment énormément de résultats originaux dus à l'école tchèque : cas stationnaire, cas non borné, domaines extérieurs, conditions aux bords de type entrée-sortie, *etc*...

2.4 Viscosités anisotropes ou non constantes.

Il est important de remarquer que les deux démonstrations de P.–L. Lions et d'E. Feireisl utilisent fortement que les viscosités μ et λ sont constantes. Elle utilise également fortement une diffusion isotrope dans toutes les directions en espace.

Que se passe-t-il, par exemple, si les viscosités dépendent de la densité ? C'est-à-dire si l'on considère un opérateur de diffusion du type

$$-2\mathrm{div}(\mu(\rho)D(u)) - \nabla(\lambda(\rho)\mathrm{div}u)$$

avec $D(u) = (\nabla u + {}^t\nabla u)/2$?

Que se passe-t-il également si l'on considère une viscosité anisotrope comme cela est fait par exemple en océanographie sur les équations de Navier-Stokes incompressible ? C'est-à dire avec un opérateur du type

$$-\Delta_\mu u - \widetilde{\lambda}\nabla\mathrm{div}u$$

où $\nabla_\mu = (\mu\partial_x, \mu\partial_y, \mu_z\partial_z)$ où $\mu_z \neq \mu$ et $\widetilde{\lambda} \geq -\min(\mu, \mu_z)$.

Pour ce qui est de la dépendance des viscosités par rapport à la densité, nous y reviendrons dans la section suivante. Un résultat d'existence globale de solutions faibles dans le cas anisotrope semble quant à lui loin d'être obtenu. Une tentative, infructueuse, a pourtant été menée par l'auteur en collaboration avec B. Desjardins et D. Gérard-Varet, *cf.* [10]. La ligne d'échec est la suivante. Si l'on essaie de suivre la démonstration de P.–L. Lions, on est alors amené à l'égalité suivante

$$\overline{\rho\mathrm{div}u} = \overline{\rho}\mathrm{div}\overline{u} + \overline{\rho(A_\mu\Delta)\rho^\gamma} - \overline{\rho}\overline{(A_\mu\Delta)\rho^\gamma}$$

avec $A_\mu = (\Delta_\mu + \widetilde{\lambda}\Delta)^{-1}$ au lieu de

$$\overline{\rho\mathrm{div}u} = \overline{\rho}\mathrm{div}\overline{u} + \overline{\rho^{\gamma+1}} - \overline{\rho}\overline{\rho^\gamma}$$

comme c'est le cas normallement. Dans le dernier cas, on peut alors utiliser la convexité de $x \mapsto x^{\gamma+1}$ pour conclure que $\overline{\rho\mathrm{div}u} \geq \overline{\rho}\mathrm{div}\overline{u}$. Dans le premier cas, le dernier terme du membre de droite ne semble pas avoir de signe. Il ne semble donc pas possible de conclure.

Pour finir, dans le même ordre d'idée, on peut également se poser la question naturelle suivante : pour quelles équations de type fluides compressibles non newtoniens est-il possible d'établir un résultat d'existence globale de solutions faibles ? La réponse ne semble pas triviale, tout au moins pour l'auteur.

3 Solutions à la Leray sur Navier–Stokes compressible avec température ?

Cette partie correspond à une note aux comptes rendus de l'Académie des Sciences écrite, en collaboration, par l'auteur et B. Desjardins, *cf.* [8]. On présente un résultat de stabilité de solutions régulière approchées. Nous renvoyons le lecteur à [5] pour le détail de preuve. La construction des suites de solutions approchées est en cours dans [9]. Une fois cette construction effectuée, nous obtiendrons le premier résultat d'existence globale de solutions faibles sur le modèle complet de Navier-Stokes. Depuis les travaux fondateurs de P.–L. Lions sur les équations de Navier–Stokes compressible avec coefficients de viscosité constants, les modèles compressibles barotropes ont été abondamment étudiés. L'amélioration principale sur l'existence de solutions faibles a d'ailleurs été celle de E. Feireisl concernant les coefficients de γ et les classes de pression considérées. La compréhension des modèles compressibles complets avec équation sur la température, elle, est nettement moins avancée mis à part quelques résultats dans le cas de la dimension 1 d'espace ou dans celui de solutions locales en temps ou de solutions globales à données petites en dimension quelconque.

La principale difficulté dans la construction de solutions globales faibles à la Leray de ce système fortement non linéaire provient du manque d'estimations *a priori*. En effet, dans le cas par exemple du gaz parfait, les seules estimations a priori disponibles semblent être : ρ, $\rho|u|^2$, $\rho\theta$ et $\rho\log\rho$ bornés dans $L^\infty(0,T;L^1(\Omega))$ et $\kappa(\rho,\theta)\theta^{-2}|\nabla\theta|^2$ et $|D(u)|^2/\theta$ bornés dans $L^1((0,T)\times\Omega)$. Ces bornes ne sont pas suffisantes pour construire des solutions faibles et, en fait, elles ne sont même pas suffisantes pour donner un sens à l'équation d'énergie elle-même car $u(\rho|u|^2/2+\rho e+p)=u(\rho|u|^2/2(r+C_v)\rho\theta)$ ne semble alors pas être intégrable.

Dans le cas d'un domaine entier ou périodique, des travaux récents, *cf.* [5], ont, pourtant, permis de montrer qu'il est possible d'obtenir l'analogue de la théorie de Pierre–Louis Lions pour les équations de Navier–Stokes compressible complètes avec conduction de chaleur sous des hypothèses de compatibilité entre les viscosités λ et μ. Cette hypothèse de compatibilité permet, notamment, d'obtenir l'information manquante sur ρ, u, θ pour conclure à l'intégrabilité de chaque terme non linéaire de l'équation d'énergie.

Ce résultat d'existence globale de solutions à la Leray peut être vu comme une première réponse partielle à un problème complètement ouvert, sur les équations de Navier–Stokes compressible complètes, décrit dans le livre de P.–L. Lions [28]. Notons l'article récent de A. Mellet et A. Vasseur, [34], où un résultat concernant les écoulements barotropes est obtenu pour $\gamma>1$ et en toute dimension d'espace entre 1 et 3. Ce résultat utilise d'une part la nouvelle entropie mathématique découverte dans [7] mais également un multiplicateur adéquat pour mieux contrôler $\sqrt{\rho}u$ et pouvoir passer à la limite sans avoir recours aux termes de traînée utilisés dans [6] par exemple. Rappelons que sur le sujet de l'existence globale de solutions d'écoulements barotropes avec viscosités variables, seul un résultat d'existence globale de solutions fortes en dimension 2 d'espace entier avait été obtenu par V.A. Vaigant et A.V. Kazhikhov en

supposant $\mu = \text{cte}$ et $\lambda = b\rho^\beta$ avec $b > 0$ et $\beta \geq 3$. Notons qu'un domaine borné peut être considéré avec une condition aux bords de type Dirichlet homogène ou de type Navier pour étendre les résultats de [5]. Un travail a été réalisé en ce sens avec B. Desjardins et D. Gérard-Varet dans [11].

Un fluide compressible conducteur de chaleur gouverné par les équations de Navier–Stokes compressible satisfait le système suivant dans $\mathbb{R}_+ \times \Omega$

$$\partial_t \rho + \operatorname{div}(\rho u) = 0, \tag{30}$$
$$\partial_t(\rho u) + \operatorname{div}(\rho u \otimes u) = \operatorname{div}\sigma + \rho f, \tag{31}$$
$$\partial_t\left(\rho E\right) + \operatorname{div}\left(\rho u H\right) = \operatorname{div}((\sigma + p\mathrm{Id}) \cdot u) + \operatorname{div}(\kappa \nabla\theta) + \rho f \cdot u, \tag{32}$$
$$E = e + \frac{|u|^2}{2}, \quad H = h + \frac{|u|^2}{2}, \quad h = e + \frac{p}{\rho},$$

où u désigne le champ de vitesse du fluide tridimensionnel, ρ la densité, κ la conductivité thermique, σ le tenseur des contraintes, p le champ de pression, e l'énergie interne spécifique et h l'enthalpie spécifique. L'énergie totale spécifique est notée E et l'entalpie spécifique associée H. Finallement les forces extérieures sont données par un champ f. Les équations (30) (31) et (32) expriment respectivement la conservation de la masse, des moments et de l'energie totale. Afin de fermer le système, deux ingrédients supplémentaires sont nécessaires. Tout d'abord, le fluide est supposé Newtonien, c'est–à–dire qu'il existe μ et λ tels que

$$\sigma = 2\mu D(u) + (\lambda \operatorname{div} u - p)\,\mathrm{Id}, \tag{33}$$

où $D(u)$ désigne le taux de déformation, exprimé comme la partie symétrique du gradient de vitesse ∇u. Comme deuxième condition, une loi thermodynamique de fermeture donne la pression p et l'énergie interne e comme fonctions de la densité ρ et de la température θ. Par simplicté ici, on supposera l'écoulement d'un gaz parfait. Des lois générales sont considérées dans [5].

Le système (30)–(32) est complèté par des conditions initiales

$$\rho|_{t=0} = \rho_0, \qquad \rho u|_{t=0} = m_0, \qquad \rho E|_{t=0} = G_0. \tag{34}$$

Les fonctions ρ_0 et m_0 sont supposées satisfaire

$$\rho_0 \geq 0, \quad \text{et} \quad \frac{|m_0|^2}{\rho_0} = 0 \quad \text{sur} \quad \{x \in \Omega \,/\, \rho_0(x) = 0\}, \tag{35}$$

et G_0 est pris tel que

$$G_0 - \frac{|m_0|^2}{2\rho_0} \geq 0, \quad \text{p.p. dans } \Omega. \tag{36}$$

Dans la plupart des applications pratiques et industrielles, les coefficients de viscosité μ et λ, ainsi que le coefficient de conductivité thermique κ sont des fonctions données de la densité et température (la loi de Sutherland en est un exemple).

La question du caractère globalement bien posé du système précédent est l'un des grands défis depuis la fin de siècle dernier. Seuls des résultats très partiels sont disponibles, voir [22] et [20]. Dans ce papier remarquable, l'existence de solutions variationnelles est obtenue pour des lois de pression particulières, données par $p(\rho,\theta) = p_e(\rho) + \theta p_\theta(\rho)$ avec un comportement spécifique à l'infini pour p_e, et des restrictions sur la croissance de p_θ, qui doit croître plus lentement que la pression à température zéro p_e. Remarquons que la loi de pression pour les gaz parfaits ne satisfait pas les conditions données par les auteurs, même pour des densités grandes. En fait, la méthode utilisée dans [22], [20] exploite fortement le rôle dominant de la pression barotrope p_c, même loin du vide, pour obtenir un résultat d'existence. Cette hypothèse restrictive empêche de traiter les situations communément rencontrées dans les applications pratiques. De plus, dans [22], l'équation de température est satisfaite seulement comme une inégalité (ce qui justifie la notion de solutions variationnelles) ce qui ne semble pas satisfaisant d'un point de vue physique, même si le second principe de la Thermodynamique est préservé. Remarquons tout de même l'extension récente, très interessante, obtenue dans [20] où le cas de viscosités dépendants de la température est considéré. Ce travail est le seul résultat avec une telle dépendance importante pour les applications physiques.

Dans le résultat de [5], des solutions faibles classiques à la Leray sont obtenues (des solutions des équations de Navier–Stokes au sens des distributions). Pour cela, les coefficients de viscosité λ et μ sont supposés être respectivement des fonctions $C^0(\mathbb{R}_+^*)$ and $C^0(\mathbb{R}_+) \cap C^1(\mathbb{R}_+^*)$ de la densité seulement, telles que les contraintes suivantes sont satisfaites pour toute densité positive ρ : Il existe trois constantes positives c_0, c_1, A, tels que, pour tout $\tau > 0$,

$$\lambda(\tau) = 2(\tau\mu'(\tau) - \mu(\tau)), \tag{37}$$

$$\forall \tau < A, \quad \mu(\tau) \geq c_0\tau^n \text{ et } 3\lambda(\tau) + 2\mu(\tau) \geq c_0\tau^n, \tag{38}$$

$$\forall \tau \geq A, \quad c_1\tau^m \leq \mu(\tau) \leq \frac{\tau^m}{c_1} \text{ et } c_1\tau^m \leq 3\lambda(\tau) + 2\mu(\tau) \leq \frac{\tau^m}{c_1} \tag{39}$$

avec $n \in (2/3, 1)$ et $m > 1$. Rappelons que l'hypothèse (37) a été introduite dans [7] dans le cadre de fluides barotropes. L'extension à des coefficients de viscosité dépendant de la densité et de la température, qui permettrait par exemple de couvrir le cas de la loi de Sutherland malheureusement reste hors de portée des résultats présentés. On remarque également que l'hypothèse faite sur les coefficients ne couvrent ni le cas des coefficients constants, ni celui découvert par V.A. Veigant et A.V. Kazhikhov. Dans le cas de viscosités constantes, il est important de remarquer que l'on obtient aucun caractère diffusif du système de part le signe opposé de μ et λ.

On suppose également que le coefficient de conductivité thermique κ satisfait

$$\kappa(\rho,\theta) = \kappa_0(\rho,\theta)(\rho+1)(\theta^a+1), \tag{40}$$

avec $a \geq 2$ où κ_0 est une fonction $C^0(\mathbb{R}_+^2)$ telle que pour toute densité positive

ρ

$$c_3 \leq \kappa_0(\rho,\theta) \leq \frac{1}{c_3}, \tag{41}$$

pour une constante positive c_3.

On suppose également que l'équation d'état est celle d'un gaz parfait polytropic du type :

$$p = r\rho\theta + p_c(\rho), \qquad e = C_v\theta + e_c(\rho), \tag{42}$$

où r et C_v sont des coefficients strictement positifs. De plus la pression additionnelle p_c et l'énergie interne e_c sont associés à l'isotherme de Kelvin. On demande que e_c soit une fonction de classe $\mathcal{C}^2$ positive de R_+^* telle que

$$p_c(\rho) = \rho^2 \frac{de_c}{d\rho}(\rho).$$

On demande également qu'il existe $\rho_* > 0$, $\tau_* > 0$, $k > 1$, $\ell > 1$, $C_* > 1$, $C'_* > 0$, $C_{**} > 0$, $C'_{**} > 0$ tels que pour tout $\rho \in (0, \rho_*)$,

$$\frac{\rho^{-\ell-1}}{C_*} \leq p'_c(\rho) \leq C_*\rho^{-\ell-1}, \qquad \frac{\rho^{-\ell-1}}{C'_*} \leq e_c(\rho) \leq C'_*\rho^{-\ell-1},$$

où $\ell \geq \dfrac{2n(3m-2)}{m-1} - 1$ et pour tout $\rho > \rho_*$

$$-\frac{1}{\tau_*}\mu'(\rho) \leq p'_c(\rho) \leq C_{**}\rho^{k-1}, \qquad 0 \leq e_c(\rho) \leq C'_{**}\rho^{k-1},$$

où $k \leq (m - \dfrac{1}{2})\dfrac{5(\ell+1)-6n}{\ell+1-n}$. Remarquons que la composante froide de la pression et de l'énergie interne peut s'annuler pour ρ suffisamment grand. On retrouve ainsi l'équation d'état d'un gaz parfait loin du vide.

Definition de solutions faibles. On dira que (ρ, u, θ) est une solution faible sur $(0,T)$ de (30)–(34) si les trois conditions suivantes sont satisfaites :

- Les propriétés de régularité suivantes sont satisfaites

$$\rho e \ \text{ et } \ \rho|u|^2 \in L^\infty(0,T;L^1(\Omega)), \quad \frac{\nabla\mu(\rho)}{\sqrt{\rho}} \in L^\infty(0,T;(L^2(\Omega))^3), \tag{43}$$

$$(\rho^{n/2} + \rho^{m/2})\nabla u \in L^2(0,T;L^2(\Omega)^9), \tag{44}$$

$$(1+\sqrt{\rho})\nabla(\theta^{a/2} + \log\theta) \in L^2(0,T;(L^2(\Omega))^3), \tag{45}$$

Enfin, on a

$$\begin{aligned} &\rho \in C([0,T];H^{-s}(\Omega)),\\ &\rho u \in C([0,T];H^{-s}(\Omega)^3), \quad \rho E \in C([0,T];H^{-s}(\Omega)), \end{aligned} \tag{46}$$

avec s une constante positive.

- La condition initiale (34) est satisfaite dans $D'(\Omega)$.
- Les équations (30)–(32) sont satisfaites dans $(D'((0,T)\times\Omega))$.

On obtient alors le résultat suivant :

Théorème 3.1 *Soi une suite de solutions approchée régulières satisfaisant de manière uniforme (52), l'inégalité d'énergie et (50)–(51), associées à des approximations régulières des données initiales. Supposons (37)–(41) satisfaits et que les données initiales (ρ_0, m_0, G_0) satisfont (35) et (36), et sont prises telles que*

$$H(0) = \int_\Omega \left(G_0 + \frac{|m_0|^2}{2\rho_0} \right) dx < +\infty, \tag{47}$$

que la densité initiale ρ_0 et l'entropie initiale s_0 satisfont

$$\rho_0 - \rho_\infty \in L^1(\Omega), \quad \rho_0 \log \frac{\rho_\infty}{\rho_0} \in L^1(\Omega), \quad \rho_0 e_c(\rho_0) \in L^1(\Omega), \tag{48}$$

$$\frac{\nabla\mu(\rho_0)}{\sqrt{\rho}_0} \in (L^2(\Omega))^3, \quad \rho_0 s_0 \in L^1(\Omega) \tag{49}$$

Alors, pour un gaz satisfaisant 42 où $s_0 = C_v \log(\theta_0/\rho_0^\Gamma)$ avec C_v et Γ constantes liées au gaz, il existe une solution globale faible de (30)–(32) avec (34).

La preuve d'un tel résultat repose sur des résultats de compacité pour des suites de solutions approchées, qui sont en cours de construction dans [9].

Notons que des lois d'état plus générales peuvent être considérées. De tels résultats de compacité sont obtenus grâce aux relations suivantes mises en évidence formellement dans le cas barotrope dans [8]

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \rho|u|^2 + \int_\Omega 2\mu(\rho)D(u) : D(u) + \int_\Omega \lambda(\rho)|\mathrm{div} u|^2 = \int_\Omega p(\rho,\theta)\,\mathrm{div} u \tag{50}$$

et

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \rho|u + 2\nabla\varphi(\rho)|^2 + \int_\Omega 2\mu(\rho)A(u) : A(u) = \int_\Omega p(\rho,\theta)\mathrm{div} u - 2\nabla p(\rho,\theta)\cdot\nabla\varphi(\rho), \tag{51}$$

où $A(u) = (\nabla u - {}^t\nabla u)/2$ représente la partie antisymétrique de ∇u, et φ est défini à une constante près par $\varphi'(\tau) = \mu'(\tau)/\tau$ $(\tau > 0)$. Cette relation va permettre de contrôler ρ proche et loin du vide. La pression froide aidant l'information proche du vide. On utilise également la forme du coefficient de conductivité choisie avec l'inégalité d'entropie

$$\int_\Omega \frac{1}{\theta}(2\mu|D(u)|^2 + \lambda|\mathrm{div} u|^2) + \int_\Omega \frac{\kappa}{\theta^2}|\nabla\theta|^2 \le \frac{d}{dt}\int_\Omega \rho s \tag{52}$$

et une information supplémentaire sur $(1+\sqrt{\rho})\nabla\theta^{(a-c+1)/2}$ obtenue en multipliant formellement l'équation de température $C_v(\partial_t(\rho\theta) + \mathrm{div}\,(\rho\theta u) +$

$\Gamma\rho\theta$ div $u) = 2\mu D(u) : D(u) + \lambda|$ div $u|^2 +$ div $(\kappa\nabla\theta)$ par $1/(C_v\theta^c)$ pour c approprié avec $0 < c < 1$.

Dans la démonstration de [22], le manque d'information sur la densité ne permet pas d'établir l'intégrabilité *a priori* de certaines quantités pour définir une formulation faible. Elle ne permet pas non plus d'établir la compacité nécessaire sur θ pour un passage à la limite sur l'équation en température si celle-ci était potentiellement définie. Une formulation renormalisée sur la température est alors nécessaire. L'inéquation obtenue sur la température provient, dans le processus de passage à la limite, d'une mesure de défaut dont les auteurs savent caractériser le signe.

4 Faible nombre de Mach et solutions à la Leray.

Le résultat d'existence globale de solutions faibles à la Leray pour les équations de Navier-Stokes compressible barotrope, établi par P.–L. Lions, a ouvert un grand champ d'études mathématiques possibles. Peut-on justifier la convergence de solutions faibles à la Leray de Navier-Stokes compressible vers des solutions faibles à la Leray de Navier-Stokes incompressible quand le nombre de Mach Ma tend vers 0 ? Plus précisemment, peut on justifier le passage du modèle compressible

$$\partial_t \rho + \mathrm{div}(\rho u) = 0, \tag{53}$$

$$\partial_t(\rho u) + \mathrm{div}(\rho u \otimes u) - \mu\Delta u - (\mu+\lambda)\nabla \mathrm{div} u + \frac{\nabla \rho^\gamma}{\mathrm{Ma}^2} = 0, \tag{54}$$

au modèle incompressible

$$\partial_t u + u \cdot \nabla u - \mu\Delta u + \nabla\pi = 0, \tag{55}$$

$$\mathrm{divu} = 0 \tag{56}$$

en laissant le nombre de Mach Ma tendre vers 0 et en supposant la densité initiale ρ_0, dépendant de Ma, asymptotiquement égale à 1 quand le nombre de Mach tend vers 0 ? On est alors dans le cas d'une équation aux dérivées partielles, en les variables $(\Psi = (\rho - 1)/\mathrm{Ma}, m = \rho u)$, perturbée par opérateur singulier antisymétrique. Cette question a, par exemple, donné lieu à plusieurs articles car les méthodes employées et les résultats obtenus dépendent fortement du type de données initiales (mal ou bien préparées) ou/et du type de domaines considérés (tore, domaine entier ou borné). Rappelons au lecteur que le nombre de Mach est le rapport entre la vitesse caractéristique du fluide et la vitesse du son dans ce fluide. Notons que les estimations, pour parvenir au résultat de convergence, dépendent du type de solutions considérées : solution forte ou solutions faibles.

Difficultés inhérentes aux solutions faibles ? Les principales difficultés, pour une analyse asymptotique sur les solutions faibles, sont la possible dégénérescence de

la densité et la faible régularité des solutions (régularité d'énergie). Rappelons que certaines questions fondamentales comme l'unicité ou la stabilité (système bien posé au sens d'Hadamard) restent encore largement ouvertes sur les solutions faibles. Le grand avantage de la théorie des solutions faibles est qu'elle n'impose aucune restriction sur les données et que l'on a pas à se soucier de la possible dégénérescence du temps d'existence en fonction des coefficients comme c'est le cas lorque l'on s'intéresse aux solutions régulières. Dans le cas de solutions régulieres, obtenir une borne inférieure du temps d'existence, indépendante des coefficients, n'est pas toujours une tâche facile, voir par exemple les articles de G. Métivier, S. Schochet (*cf.* [35]) et de T. Alazard (*cf.* [2]).

Résultats connus sur les solutions faibles. L'étude mathématique de la limite incompressible, sur les solutions faibles à la Leray, remonte, dans le cas de données initiales mal préparées et avec domaine périodique, principalement aux travaux de P.–L. Lions (*cf.* [30]). Ces travaux ont été, par la suite, complétés par P.–L. Lions et N. Masmoudi (*cf.* [31]) qui ont écrit un article considérant successivement le domaine entier, le domaine périodique puis le domaine borné avec des conditions aux bords de type Navier.

Il y eu, ensuite, les travaux de B. Desjardins et E. Grenier (*cf.* [16]) qui ont montré la convergence forte locale en espace, avec des données mal préparées, dans l'espace entier en utilisant des inégalités de type Strichartz (effet dispersif de l'équation des ondes). Ils ont ainsi pu montrer que la partie potentielle de la vitesse converge localement fortement vers 0 de manière beaucoup plus simple que par les auteurs précédents. La convergence forte de la partie incompressible étant aisément obtenue par des méthodes de compacité standarts.

Puis un résultat très intéressant a été établi par les quatres auteurs, précédemment cités, (*cf.* [17], dans le cas d'un domaine borné avec données mal préparées et conditions aux bords de type Dirichlet homogène. Ils ont montré que les ondes acoustiques étaient amorties dans la couche visqueuse proche du bord sous hypothèse que le domaine ne soit pas une boule (résultat basé sur la conjecture de Schiffer). Ce résultat est basé sur un développement asymptotique du spectre de l'opérateur des ondes perturbé.

Un résultat de convergence locale a également été montré par P.–L. Lions et N. Masmoudi, *cf.* [32], par utilisation simple de lois de conservation pour les ondes. Ceci leur a permis notamment d'obtenir un résultat avec des applications en domaine extérieur, mais cet article a également donné une direction nouvelle pour l'étude du passage des équations de type Boltzmann vers les équations gouvernant les fluides incompressibles. On renvoie le lecteur intéressé par les limites hydrodynamiques de l'équation de Boltzmann à l'article de C. Villani ([42]). Nader Masmoudi, (*cf.* [33]), a également justifié que les solutions faibles de Navier-Stokes compressible isentropique convergent vers des solutions d'Euler incompressible quand le nombre de Mach et le nombre de Reynolds tendent vers 0 en considérant des conditions aux bords stables lors du passage à la limite.

Fluides compressibles en géophysique. Autour des limites asymptotiques de type

faible nombre de Mach, mentionnons les travaux de l'auteur, B. Desjardins et D. Gérard-Varet (*cf.* [10]) dans le cas d'un domaine cylindrique en supposant que le nombre de Froude Fr et le nombre de Rossby Ro tendent vers 0 avec Ro = Fr. Notons que dans ce cas, le nombre de Froude joue le même rôle que le nombre de Mach.

Ils montrent alors que si la convergence forte doit avoir lieu elle ne peut être que sous l'hypothèse que la base du cylindre soit un disque. Ce résultat est également basé sur la conjecture de Schiffer mais également sur une hypothèse de généricité spectrale. On rappelle que les auteurs n'ont pas obtenu l'existence globale de solutions faibles sur Navier-Stokes compressible avec viscosités anisotropes.

Il existe également un résultat de convergence de solutions faibles à la Leray sur les équations de Saint-Venant avec viscosités quand le nombre de Froude Fr et nombre de Rossby Ro tendent vers 0 avec Fr = Ro et en supposant $b = 1$. Les équations de Saint-Venant sont, en effet, des équations de type fluides compressibles. On obtient alors, asymptotiquement, les équations quasi-géostrophique standarts avec un terme supplémentaire, trace de la surface libre. On appelle équations quasi-géostrophique standarts les équations obtenues à partir de Navier-Stokes incompressibles tri-dimensionnelles avec force de Coriolis dans un tore par analyse asymptotique quand le nombre de Rossby tend vers 0. On renvoie le lecteur intéressé au livre récent de J.–Y. Chemin, B. Desjardins, I. Gallagher, E. Grenier (it cf. [14]). Notons que la viscosité sur Saint-Venant visqueux dépends de la hauteur du fluide. On n'est donc plus dans le cadre de solutions faibles dûes à P.–L. Lions mais dans le cadre de diffusion dégénérée. On utilise alors un résultat de l'auteur et B. Desjardins, (*cf.* [6]). Pour indication, l'équation de Saint-Venant bidimensionelle visqueuse qui a été considérée s'écrit

$$\partial_t h + \mathrm{div}(hu) = 0, \tag{57}$$

$$\partial_t (hu) + \mathrm{div}(hu \otimes u) + h\nabla \frac{(h-b)}{\mathrm{Fr}^2} - 2\nu \mathrm{div}(hD(u)) + \frac{hu^\perp}{\mathrm{Ro}} + r_0 u + r_1 h|u|u = 0 \tag{58}$$

avec $u = (u_1, u_2)$ où $u^\perp = (-u_2, u_1)$, $D(u) = (\nabla u + {}^t\nabla u)/2$, b est une fonction dépendante de x donnée qui correspond au fond et ν, r_0 et r_1 sont respectivements la viscosité, et les coefficients de frottement laminaire et turbulent.

Un dernier résultat concerne le lien entre les équations de Saint-Venant avec viscosité ci-dessus et les équations des lacs avec viscosité suivante

$$\partial_t (bu) + bu \cdot \nabla u + b\nabla p \tag{59}$$

$$-2\nu \mathrm{div}(bD(u)) + \frac{bu^\perp}{\mathrm{Ro}} + r_0 u + r_1 b|u|u = 0, \tag{60}$$

$$\mathrm{div}(bu) = 0. \tag{61}$$

On remarque qu'il s'agit d'une limite quand Fr tend vers 0 avec Ro fixé en supposant b une fonction de x avec $b = O(1)$. Cette étude a été mené récemment par l'auteur, M. Gisclon et C.K. Lin dans [12] dans le cas de données bien préparées et en supposant $b \geq c > 0$ avec b le fond.

Limite faible nombre de Mach et modèle de Navier-Stokes complet ? Maintenant que des résultats d'existence globale à la Leray sur le modèle complet ont été établi, une preuve de la convergence forte locale sur ce type de solutions vers des équations de Navier-Stokes non homogène et contrainte sur la divergence de u en rapport avec la température peut être envisagée dans le cas du domaine entier. Ce travail est en cours par Th. Alazard, l'auteur et B. Desjardins, *cf.* [3]. Il complèterait ainsi le résultat récent de Th. Alazard (*cf.* [1]) qui traite de la limite incompressible sur solutions fortes du modèle complet dans l'espace entier et qui apporte une réponse à une question difficile qui était, notamment, posée par A. Majda et P.–L. Lions. Le lecteur pourra d'ailleurs consulter [29] pour l'asymptotique formelle. Le résultat de T. Alazard prolonge également le travail de G. Métivier et S. Schochet (*cf.* [35]) qui traitait le cas où l'entropie est transportée.

La convergence dans un domaine périodique est, elle, loin d'être justifiée dans le cas de données mal préparées que ce soit sur les solutions faibles ou sur les solutions fortes avec transport d'entropie comme première approximation. Le problème provient d'une étude spectrale d'opérateurs à coefficients variables dépendants du temps et possibilité de croisement de valeurs propres. Un élément de réponse a été apporté par G. Métivier et S. Schochet en dimension finie en considérant l'entropie transportée et les problèmes identifiés sur le passage à la limite en la dimension.

Le lecteur intéressé par des articles de revues très bien écrits sur la limite incompressible, avec énoncés mathématiques des résultats, est renvoyé à [1], [15], [18] et [24].

Remerciements. L'auteur tient à remercier Enrique Fernández- Cara pour lui avoir proposé de rédiger cet article pour le bulletin de la SEMA. Il remercie également Thomas Alazard de Bordeaux I pour sa relecture et ses commentaires.

Pourquoi cette dédicace ? Les résultats exposés ici sont, en partie, le fruit de travaux en collaboration, depuis plusieurs années, entre l'auteur et Benoît Desjardins. Par cette dédicace, l'auteur tiens à lui témoigner son amitié ainsi qu'à Nancy.

Références

[1] Th. Alazard. Alentours de la limite incompressible. Séminaire : Équations aux dérivées partielles, 2004–2005, Exp. No. XXIV, 16pp., École Polytechnique, Palaiseau, (2005).

[2] Th. Alazard. Low Mach number limit of the full Navier-Stokes equations. *Arch. Rational Mech Anal.,* à paraître (2005).

[3] TH. ALAZARD, D. BRESCH, B. DESJARDINS. Low Mach numer limit of the full Navier-Stokes equations for weak solutions. En préparation.

[4] F. BOYER, P. FABRIE. *Éléments d'analyse pour l'étude de quelques modèles d'écoulements de fluides visqueux incompressibles.* Séries : Mathématiques et Applications, À paraître, vol. 22, (2006).

[5] D. BRESCH, B. DESJARDINS. On the existence of global weak solutions to the Navier–Stokes equations for viscous compressible and heat conducting fluids. A paraître dans *J. Math. Pures et Appliquées,* (2006).

[6] D. BRESCH, B. DESJARDINS. Existence of global weak solutions for a 2D viscous shallow water equations and convergence to the quasi-geostrophic model. *Commun. Math. Phys.,* **238**, 1–2, (2003), 211–223.

[7] D. BRESCH, B. DESJARDINS. Some diffusive capillary models of Korteweg type. *C. R. Acad. Sciences,* Paris, Section Mécanique, Vol. **332**, no11 (2004), 881–886.

[8] D. BRESCH, B. DESJARDINS. Stabilité de solutions faibles globales pour les équations de Navier-Stokes modélisant un fluide compressible conducteur de chaleur. *C.R. Acad. Sci.,* Paris, Section Mathématiques, vol.343, Issue 3, 219–224, (2006).

[9] D. BRESCH, B. DESJARDINS. On the construction of approximate solutions for the 2D viscous shallow water model and for compressible Navier-Stokes models. A paraître dans *J. Maths. Pures et Appliquées,* (2006).

[10] D. BRESCH, B. DESJARDINS, D. GÉRARD-VARET. Rotating fluids in a cylinder. *Discrete Contin. Dyn. Syst.,* (11)1 : 47–82, (2004).

[11] D. BRESCH, B. DESJARDINS, D. GÉRARD-VARET. On compressible Navier-Stokes equations with density dependent viscosities in bounded domains. A paraître dans *J. Maths Pures et Appliquées,* (2006).

[12] D. BRESCH, M. GISCLON, C.K. LIN. An example of low Mach (Froude) number effects for compressible flows with nonconstant density (height) limit. *Math. Mod. Numer. Anal.* 39(3), 477–486, (2005).

[13] J.–Y. CHEMIN. Jean Leray et Navier-Stokes. Numéro spécial de la Gazette des Mathématiciens, Société Mathématique de France, 71–82 (2000).

[14] J.–Y. CHEMIN, B. DESJARDINS, I. GALLAGHER, E. GRENIER. *Mathematical aspects of rotating fluids.* À paraître chez Oxford lecture series in Mathematics and its applications, (2006).

[15] R. DANCHIN. Low Mach number limit for viscous compressible flows. *M2AN Math. Model. Numer. Anal., special issue,* 39, No3 (2005).

[16] B. DESJARDINS, E. GRENIER. Low Mach number limit of viscous compressible flows in the whole space. *P. Soc. Lond. Proc. Ser. A. Math. Phys. Eng. Sci.* **455**, 2271–2279, (1999).

[17] B. DESJARDINS, E. GRENIER, P.–L. LIONS, N. MASMOUDI. Incompressible limit for solutions of the isentropic Navier-Stokes equations with Dirichlet boundary conditions. *J. Math. Pures Appl.,* **78**, (1999), no.5, 461–471.

[18] B. Desjardins, C.K. Lin. A survey of the compressible Navier–Stokes equations. *Taiwanese Journal of Mathematics,* vol. 33, No. 2, pp. 123–137, (1999).

[19] R.J. Diperna, P.–L. Lions. Ordinary differential equations, Sobolev spaces and transport theory. *Invent. Math.,* **98**, pp. 511–547, (1989).

[20] E. Feireisl. *Dynamics of viscous compressible fluids.* Oxford Science Publication, Oxford, (2004).

[21] E. Feireisl. Viscous and/or heat conducting compressible fluids. *Handbook of mathematical fluid dynamics,* vol. I, 307–371, North-Holland, Amsterdam, (2002).

[22] E. Feireisl. On the motion of a viscous, compressible, and heat conducting fluid. *Indiana Univ. Math. J.,* **53**, 1707–1740, (2004).

[23] E. Fernández-Cara. A review of basic theoretical results concerning the Navier-Stokes and other similar equations. *Bol. Soc. Esp. Mat. Apl.,* No32, (2005), 45–73.

[24] I. Gallagher. Résultats récents sur la limite incompressible. *Séminaire Bourbaki,* (2003)–(2004), no.926.

[25] M. Hillairet. Propagation of density-oscillations in solutions to the barotropic compressible Navier-Stokes system. To appear in *J. Math. Fluid. Mech.,* (2005).

[26] A. Kazhikhov. Communication personnelle, (2002).

[27] J. Leray. Selected papers. Oeuvres scientifiques. Vol. II. Fluid dynamics and real partial differential equations. With an introduction by Peter Lax. Edited by Paul Malliavin. *Springer-Verlag, Berlin ; Société Mathématique de France, Paris,* (1998).

[28] P.-L. Lions. *Mathematical topics in fluid dynamics, Vol.1, incompressible models.* Oxford Science Publication, Oxford, (1998).

[29] P.-L. Lions. *Mathematical topics in fluid dynamics, Vol.2, compressible models.* Oxford Science Publication, Oxford, (1998).

[30] P.–L. Lions. Limites incompressible et acoustique pour des fluides visqueux, compressibles et isentropiques. *C. R. Acad. Sci. Paris, Ser I Math.,* **316**, (1993), 1335–1340.

[31] P.–L. Lions, N. Masmoudi. Incompressible limit for a viscous compressible fluid. *J. Math. Pures Appl.,* **77**, (1998), no.6, 585–627.

[32] P.–L. Lions, N. Masmoudi. Une approche locale de la limite incompressible. *C.R. Acad. Sci.* Paris, Série Math., 329, 5, (1999), 387-392.

[33] N. Masmoudi. Incompressible, inviscid limit of the compressible Navier-Stokes system. *Ann. Inst. H. Poincaré Anal. Non Linéaire,* **18**, (2001), no.2, 199–224.

[34] A. Mellet, A. Vasseur. On the isentropic compressible Navier-Stokes equations. A paraître *Comm. Partial Diff. Eqs* (2006).

[35] G. MÉTIVIER, S. SCHOCHET. The incompressible limit of the non-isentropic Euler equations. *Arch. Ration. Mech. Anal.,* **158**, (2001), no. 1, 61–90.

[36] A. NOVOTNY, I. STRASKRABA. *Introduction to the mathematical theory of compressible flow.* Oxford lecture series in Mathematics and its applications, (2004).

[37] D. SERRE. Variations de grande amplitude pour la densité d'un fluide visqueux compressible. *Phys. D,* 48(1) :113-128, (1991).

[38] J. SIMON. Non-homogeneous viscous incompressible fluids : existence of velocity, density and pressure. *SIAM J. Math. Anal.,* **21**, (1990).

[39] J. SIMON. Compact sets in the space $L^p(0,T;B)$. *Annali. Mat. Pura ed Appl.,* (IV), Vol. CXLVI, pp. 65–96, (1987).

[40] R. TEMAM. *Navier-Stokes equations.* 3rd edition, North-Holland, Amsterdam, (1984).

[41] V.A. VAIGANT, A.V. KAZHIKHOV. On the existence of global solutions of two-dimensional Navier-Stokes equations of a compressible viscous fluid. *Siberian Math. J.* 36 (1995), no. 6, 1108–1141.

[42] C. VILLANI. Limite hydrodynamiques de l'équation de Boltzmann (d'après C. Bardos, F. Golse, C.D. Levermore, P.–L. Lions, N. Masmoudi et L. Saint-Raymond). *Séminaire Bourbaki,* 53ème année, (2000)–(2001), no893.

Bol. Soc. Esp. Mat. Apl. nº 36(2006), 87–96

An algorithm to share secrets based on memory cellular automata

A. Martín del Rey [1]

Departamento de Matemática Aplicada, E.P.S. de Ávila
Universidad de Salamanca

delrey@usal.es

Resumen

The main goal of this work is to explore the possibilities of a particular type of discrete dynamical systems called cellular automata in order to design cryptographic algorithms to share secrets. Specifically, a (n, n)-threshold secret sharing scheme is presented. It is proved to be perfect and ideal. Moreover a security analysis is performed.

Palabras clave: *Cryptography, Cellular Automata, Discrete Dynamical Systems, Secret Sharing Schemes*
Clasificación por materias AMS: *68P25, 68Q80, 94A60*

1 Introduction

The cryptographic algorithms designed to share a secret among a set of participants are called secret sharing schemes. The main characteristic of these schemes is that only some qualified subsets of participants can recover the original secret.

Secret sharing schemes were proposed by Shamir ([19]) and Blakley ([5]). The original motivation of these authors was to safeguard cryptographic keys from loss, deterioration or robbery. Nevertheless, these algorithms have many applications nowadays in different areas such as access control, opening a safety deposit box, etc.

The foremost secret sharing scheme is the (k, n)-*threshold scheme*, where $k, n \in \mathbb{Z}$, and $1 \leq k \leq n$. For this scheme, the collaboration of a Trusted Third Party (TTP) is necessary. The TTP computes n shares, $S_1, \ldots, S_n$, from an

Fecha de recepción: 2/1/2006

initial secret S, and distributes them in a secure way into n participants, in such a way any set of k or more of these participants can easily recover the original secret by pooling their shares. Moreover, any group of $k-1$ or less participants knowing only $k-1$ or fewer shares are unable to recover the secret.

The most basic case of this type of cryptographic protocols is obtained if $k = n$. This particular scheme is very important since it can be applied to many situations. Mainly when, by security requirements, only one qualified set of participants is appropriate.

A secret sharing scheme is called ideal if the size of every share is equal to the size of the secret to be shared. Moreover, secret sharing schemes satisfying the additional property that subsets of $k-1$ or fewer participants can obtain absolutely no information about the secret are called perfects ([15, 21, 22]).

In general, secret sharing schemes are based on mathematical tools and developments. For example, Shamir's algorithm is based on polynomial interpolation, and Blakley's scheme is based on the intersection of affine hyperplanes. Moreover, the algorithm proposed by Asmuth and Bloom in [3] uses prime numbers, the one proposed by Karnin, Greene and Hellman is based on matrix multiplication ([11]), etc.

In this work we are interested in the use of a particular type of discrete dynamical systems called memory cellular automata in order to design secret sharing schemes. In [13] this problem is tackled. The protocol proposed in this work is a degenerate (k, n)-threshold scheme with a non-complete access structure. The main goal of this work is to introduce an (n, n)-threshold scheme without the disadventages of the first one. Furthermore, this scheme exhibit some advantages over the classic ones: It is easier implemented on computer since CA are very suitable for hardware and software implementation.

Roughly speaking, memory cellular automata are delay discrete dynamical systems formed by a finite number of identical objects called cells, which are endowed with a state or value that changes at every discrete step of time according to a deterministic rule. The variables of this deterministic rule are the states of a set of cells at previous time steps (see [1, 2]). As is well known, discrete dynamical systems have several computational advantages. For example, the algorithms based on them can be easily implemented in a sequential or in a parallel way.

Cellular automata are simple models of computation capable to simulate complex behaviour. The use of cellular automata to design cryptographic protocols goes back to middle eighties when Wolfram proposed the cellular automaton with rule number 30 as a pseudorandom bit generator for cryptographic purposes ([24, 25]). Since then, many cryptosystems based on cellular automata have been proposed (see, for example, [4, 6, 7, 8, 9, 10, 14, 16, 23]), but up to our knowledge the possibilities of cellular automata in the design of secret sharing schemes have not been studied in the literature.

The rest of this work is organized as follows: In Section 2, some basic aspects about memory cellular automata is recalled. The algorithm to share secrets is introduced in Section 3, and an analysis about its security is carried out. In Section 4, an example of the use of this scheme is presented and, finally, the

conclusions are shown in Section 5.

2 A review of memory cellular automata

One-dimensional cellular automata (CA) are discrete dynamical systems formed by a finite one-dimensional array of N identical objects called cells, in such a way that each one of them can assume a state from a finite set $S = \mathbb{Z}_k$. The i-th cell is denoted by $\langle i \rangle$, and the state of this cell at time t is $s_i^{(t)}$. The CA evolves deterministically in discrete time steps, changing the states of all cells according to a local transition function. The updated state of each cell depends on the variables of the local transition function, which are the previous states of a set of cells which constitutes its neighbourhood. It is defined by an ordered set of integers $V = \{\alpha_1, \ldots, \alpha_q\} \subset \mathbb{Z}$, such that the neighbourhood of the cell $\langle i \rangle$ is given by:

$$V_i = \{\langle i + \alpha_1 \rangle, \ldots, \langle i + \alpha_q \rangle\}.$$

Consequently, the local transition function is of the following form:

$$s_i^{(t+1)} = f\left(s_{i+\alpha_1}^{(t)}, \ldots, s_{i+\alpha_q}^{(t)}\right), \quad 0 \leq i \leq N-1,$$

or equivalently,

$$s_i^{(t+1)} = f\left(V_i^{(t)}\right),$$

where $V_i^{(t)}$ stands for the states of the neighbour cells of $\langle i \rangle$ at time t. The vector

$$C^{(t)} = \left(s_0^{(t)}, \ldots, s_{N-1}^{(t)}\right) \in S^n,$$

is called the configuration at time t of the CA, where $C^{(0)}$ is the initial configuration of the CA. Moreover, the sequence $\{C^{(t)}\}_{0 \leq t \leq l}$ is called the evolution of order l of the CA, and if $\mathcal{C}$ is the set of all possible configurations of the CA, then $|\mathcal{C}| = k^N$. As the number of cells is finite, boundary conditions must be considered in order to assure the well-defined dynamics of the CA. In this work periodic boundary conditions are taken: If $i \equiv j (\mathrm{mod}\, N)$, then $s_i^{(t)} = s_j^{(t)}$.

Furthermore, we will consider those CA whose local transition function are of the following form:

$$s_i^{(t+1)} = \sum_{j=1}^{q} \sum_{\substack{\beta_1 < \ldots < \beta_j \\ \beta_1 \in V, \ldots, \beta_j \in V}} s_{i+\beta_1}^{(t)} \cdots s_{i+\beta_j}^{(t)} (\mathrm{mod}\, k),$$

where $0 \leq i \leq N-1$.

The global function of the CA is a linear transformation,

$$\Phi \colon \mathcal{C} \to \mathcal{C},$$

that yields the configuration at the next time step during the evolution of the CA, that is,

$$C^{(t+1)} = \Phi\left(C^{(t)}\right).$$

If Φ is bijective then there exists another cellular automaton, called its inverse, with global function Φ^{-1} (see [17]). When such inverse cellular automaton exists, the cellular automaton is called reversible and the evolution backwards is possible.

The standard paradigm for CA states that the state of every cell at time $t+1$ depends on the state of some cells (its neighbourhood) at time t. Nevertheless, one can consider CA for which the state of every cell at time $t+1$ not only depends on the states of some cells at time t but also on the states of (possible) another different groups of cells at times $t-1$, $t-2$, etc. This is the basic idea of memory cellular automata, MCA for short(see [20]). If the configuration $C^{(t+1)}$ of the MCA depends on the configurations $C^{(t)}, \ldots, C^{(t-k+1)}$, then it is called a k-th order MCA and, as a consequence, its local transition function is as follows:

$$s_i^{(t+1)} = f_0\left(V_i^{(t)}\right) + f_1\left(V_i^{(t-1)}\right) + \ldots + f_{k-1}\left(V_i^{(t-k+1)}\right) (\mathrm{mod}\, k), \qquad (1)$$

where $0 \leq i \leq N-1$, and f_i stands for a local transition function of a particular CA. Remark that, in this case, it is necessary to know k configurations, $C^{(0)}, \ldots, C^{(k-1)}$, in order to compute the evolution of the k-th order MCA.

Note that, in this case, the neighbourhoods of the i-th cell at times $t, t-1, \ldots, t-k+1$ can be different.

A basic point is to decide whether or not a k-th order MCA is reversible. In this sense, the following result holds (see [13]):

Proposición 1 *If* $f_{k-1}\left(V_i^{(t-k+1)}\right) = s_i^{(t-k+1)}$, *then the* k*-th order MCA with local transition function given in* (1) *is a reversible MCA, whose inverse CA is another* k*-th order MCA with local transition function:*

$$s_i^{(t+1)} = \sum_{j=0}^{k-2} f_{k-j-2}\left(V_i^{(t-j)}\right) + s_i^{(t-k+1)} (\mathrm{mod}\, k),$$

for $0 \leq i \leq N-1$.

3 The algorithm to share secrets

In this section we propose a MCA-based algorithm to share secrets consisting of a (n, n)-threshold secret sharing scheme based on the use of an n-th order MCA. As is stated above, numerous cryptographic protocols based on cellular automata have appeared in the literature. Usually, these algorithms use non-memory (or classic) cellular automata but in recent years the study of MCA as cryptographic tools have been proposed (see, for example [12, 18]). In the case of secret sharing schemes, classic cellular automata are used as follows: If we

want to share a secret into n participants then the initial configuration of the CA is divided into n blocks and one of them is given by the secret. Subsequently, several iterations of the CA must be computed in order to increase the diffusion. The configuration obtained is divided among another n blocks and these are the shares to be distributed into the participants. In the case of MCA, the text to be shared, S, is the initial configuration of a n-th order MCA, for example $S = C^{(0)}$, and the remaining $n-1$ configurations, $C^{(1)}, \ldots, C^{(n-1)}$, are $n-1$ random boolean vectors of the same size than S. The shares to be distributed among the n participants are n consecutive configurations of the evolution of the MCA. The algorithm based on MCA exhibit some advantages over classic CA since the use of MCA greatly increases the diffusion achieved per iteration. Moreover, the lenght of the configurations are smaller.

The MCA-proposed algorithm is formed by three phases which are presented in the following subsection.

3.1 Structure of the scheme

As it is mentioned above, the structure of the procedure to share secrets by means of MCA is divided into three phases: The setup phase, in which the MCA used is defined as well as its initial conditions; the sharing phase, in which the evolution of the MCA is computed and, consequently, the shares to be distributed among the participants are obtained; and finally, the recovery phase, which allows the participants to recover the shared secret.

3.1.1 The setup phase

1. The TTP computes $n-1$ random integer numbers, $q_0, \ldots, q_{n-2}$, and considers the following sets of integers:

$$\begin{aligned} V_0 &= \{-q_0, \ldots, 0, \ldots, q_0\}, \\ &\vdots \\ V_{n-2} &= \{-q_{n-2}, \ldots, 0, \ldots, q_{n-2}\}, \end{aligned}$$

   which stand for the sets defining the neighbourhood at each time of the n-th order MCA. The numbers $q_0, \ldots, q_{n-2}$ can be publicly known.

2. The TTP constructs the reversible n-th order MCA with local transition function given by:

$$s_i^{(t+1)} = \sum_{m=0}^{n-2} f_m\left(V_i^{(t-m)}\right) + s_i^{(t-n+1)} (\mathrm{mod}\, k), \quad 0 \leq i \leq N-1,$$

   where $V_i^{(t-m)}$ is the neighbourhood of the i-th cell given by the set of

indices V_m, and by the function:

$$f_m\left(V_i^{(t-m)}\right) = \sum_{j=1}^{2q_m+1} \sum_{\substack{\beta_1<\ldots<\beta_j \\ \beta_1\in V_m,\ldots,\beta_j\in V_m}} s_{i+\beta_1}^{(t)} \cdots s_{i+\beta_j}^{(t)} (\mathrm{mod}\, k).$$

3. The boolean vector representing the secret to be shared is considered as the initial configuration, $C^{(0)}$. The TTP computes the remaining $n-1$ configurations: $C^{(1)}, \ldots, C^{(n-1)}$, by using a random bit generator. Note that, once the shares are computed, these $n-1$ configurations, $C^{(1)}, \ldots, C^{(n-1)}$, can be destroyed.

3.1.2 The sharing phase

1. The TTP chooses a random integer number, p, such that $p \geq n$. This number can be publicly known.

2. Starting from the configurations $C^{(0)}, \ldots, C^{(n-1)}$, the TTP computes the $(n+p-1)$-th order evolution of the MCA:

$$\left\{C^{(0)}, \ldots, C^{(n-1)}, C^{(n)}, \ldots, C^{(p)}, \ldots, C^{(p+n-1)}\right\}.$$

3. The shares to be distributed among the n participants are the last n configurations computed: $S_1 = C^{(p)}, \ldots, S_n = C^{(n+p-1)}$.

Remark that $p \geq n$ is considered to avoid overlappings between the initial configurations and the shares.

3.1.3 The recovery phase

1. To recover the secret, $S = C^{(0)}$, the n shares $C^{(p)}, \ldots, C^{(p+n-1)}$ are needed.

2. The participants construct the inverse n-th order MCA by using the numbers $q_0, \ldots, q_{n-2}$ and the sets $V^{(0)}, \ldots, V^{(n-2)}$.

3. Taking
$$\tilde{C}^{(0)} = C^{(p+n-1)}, \ldots, \tilde{C}^{(n-1)} = C^{(p)},$$
and iterating p times the inverse n-th order MCA, the participants obtain the secret: $C^{(0)} = S$.

3.2 Security analysis

As the bit length size of every distributed share is equal to the bit length size of the secret (both are different configurations of the same n-th order MCA), the proposed scheme is ideal.

Furthermore, the scheme is also perfect because if only one share is unknown, say for example $S_n = C^{(n+p-1)}$, then the remaining $n-1$ participants can not obtain any information about the configuration $C^{(p-1)}$ as the evolution of the n-th order inverse MCA is given by the following linear system:

$$s_i^{(t+1)} = b_i + s_i^{(t-n+1)} (\mathrm{mod}\, k), \quad 0 \leq i \leq N-1,$$

where

$$b_i = f_0\left(V_i^{(t)}\right) + \ldots + f_{n-2}\left(V_i^{(t-n+2)}\right).$$

Consequently, as it is formed by n equations with $2n$ unknown variables: $s_i^{(t+1)}$, $s_i^{(t-k+1)}$, where $0 \leq i \leq N-1$, then it can not be solved and, obviously, no information about the configuration $C^{(t+1)} = \left(a_i^{(t+1)}\right)$, $0 \leq i \leq N-1$, is obtained. Note that a similar result holds if the number of unknown configurations is greater than one.

As a consequence, for the secret sharing scheme proposed it is impossible to recover the secret starting from $n-1$ (or less) shares.

Finally, the computational time needed both to obtain the shares and to obtain the secret is the same since both protocols consist of the iteration of a similar MCA (a given reversible MCA and its inverse). Specifically, they goes roughly as $O\left(2log\left(p(n-1)\right) + 8logN + Nlog2\right) \subseteq O\left(N + logp\right)$.

4 An example

Let us define a $(4,4)$-threshold secret sharing scheme.

For the sake of simplicity, let us consider the integers $q_0 = 7, q_1 = 5, q_2 = 2$; then, the sets defining the neighbourhoods are:

$$\begin{aligned} V_0 &= \{-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7\}, \\ V_1 &= \{-5,-4,-3,-2,-1,0,1,2,3,4,5\}, \\ V_2 &= \{-2,-1,0,1,2\}. \end{aligned}$$

Consequently, a 4-th order MCA can be constructed with $S = \mathbb{Z}_k$, $N = 128$ and the following local transition function:

$$s_i^{(t+1)} = f_0\left(V_i^{(t)}\right) + f_1\left(V_i^{(t-1)}\right) + f_2\left(V_i^{(t-2)}\right) + s_i^{(t-3)} (\mathrm{mod}\, 2),$$

with $0 \leq i \leq 127$, where

$$f_m\left(V_i^{(t-m)}\right) = \sum_{j=1}^{2q_m+1} \sum_{\substack{\beta_1 < \ldots < \beta_j \\ \beta_1 \in V_m, \ldots, \beta_j \in V_m}} s_{i+\beta_1}^{(t)} \cdots s_{i+\beta_j}^{(t)} (\mathrm{mod}\, 2),$$

with $0 \leq m \leq 2, 0 \leq i \leq 127$.

Suppose that the 128-bit length secret to be shared is given by the hexadecimal code

$$\texttt{eebd97edc22d0ec7d2a3623b6ec37447}.$$

Note that its binary expression stands for the configuration $C^{(0)}$ of the cellular automata. Moreover, let us randomly compute the remaining three initial conditions, $C^{(1)}, C^{(2)}$ and $C^{(3)}$, given respectively, by the following hexadecimal codes:

$$\texttt{ce64b017086c838c52f28d4a11adf0f6},$$
$$\texttt{3bc37a3093442cb286835e8b2bc9b528},$$
$$\texttt{87d16dadf6176bd9db94dde36e106ea0}.$$

As a consequence, if we take $p = 4$, then the evolution of the last defined 4-th order MCA is as follows:

$$C^{(4)} = \texttt{11426b923dd2e1382d5c9dc4913c8bb8},$$
$$C^{(5)} = \texttt{319b4fe8f7937c73ad1d72b5ee520f08},$$
$$C^{(6)} = \texttt{c43c85cf6cbbd34d797ca174d4344ad3},$$
$$C^{(7)} = \texttt{f82e925209e89427246b221c91ef915e}.$$

Consequently, the shares distributed among the participants are $S_1 = C^{(4)}$, $S_2 = C^{(5)}$, $S_3 = C^{(6)}$ and $S_4 = C^{(7)}$.

5 Conclusions

In this paper the use of cellular automata in the design of cryptographic algorithms to share secrets is studied. Moreover, a new (n, n)-threshold scheme based on memory cellular automata for text sharing is presented. It is shown to be ideal and perfect since the size of the shares to be distributed and the size of the secret are equal, and no information about the secret is obtained if $n - 1$ or less shares are known.

Acknowledgments. This work has been supported by the Consejería de Educación y Cultura of Junta de Castilla y León (Spain), and by Ministerio de Educación y Ciencia (Spain) under grant SEG2004-02418

Referencias

[1] R. Alonso-Sanz and M. Martín, One-dimensional cellular automata with memory: patterns from a single site seed, *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 12: 205–226, 2002.

[2] R. Alonso-Sanz and M. Martín, Elementary cellular automata with memory, *Complex Systems*, 14: 99–126, 2003.

[3] C. Asmuth and J. Bloom, A modular approach to key safeguarding, *IEEE T. Inform. Theory*, 29: 208–210, 1983.

[4] P.H. Bardell, Analysis of cellular automata used as pseudorandom pattern generators, *Proc. of 1990 International Test Conference*, 762–768, 1990.

[5] G.R. Blakley, Safeguarding cryptographic keys, *AFIPS Conference Proceedings*, 48: 313–317, 1979.

[6] K. Cattell and J.C. Muzio, An explicit similarity transform between cellular automata and LFSR matrices, *Finite Fields Appl.*, 4: 239–251, 1998.

[7] R. Díaz Len, A. Hernández Encinas, L. Hernández Encinas, S. Hoya White, A. Martín del Rey, G. Rodríguez Sánchez, and I. Visus Ruíz, Wolfram cellular automata and their cryptographic use as pseudorandom bit generators, *Internat. J. Pure Appl. Math.*, 4: 87–103, 2003.

[8] C. Fraile Rubio, L. Hernández Encinas, S. Hoya White, A. Martín del Rey, and G. Rodríguez Sánchez, The use of linear hybrid cellular automata as pseudorandom bit generators in cryptography, *Neural Parallel Sci. Comput.* 12: 175–192, 2004.

[9] P. Guan, Cellular automaton public-key cryptosystem, *Complex Systems* 1: 51–57, 1987.

[10] H.A. Gutowitz, Cryptography with dynamical systems, *Proc. of the NATO Advanced Study Institute: Cellular Automata and Cooperative Systems*, 237–274, 1993.

[11] E.D. Karnin, J.W Greene, and M.E. Hellman, On sharing secret systems, *IEEE T. Inform. Theory* 29: 35–41, 1983.

[12] A. Martín del Rey, Design of a Cryptosystem Based on Reversible Memory Cellular Automata, *Proc. of 10th IEEE Symposium on Computers and Communications*, 482–486, 2005.

[13] A. Martín del Rey, J. Pereira Mateus, G. Rodríguez Sánchez, A secret sharing scheme based on cellular automata, *Appl. Math. Comput.*, 170: 1356–1364, 2005.

[14] W. Meier and O. Staffelbach, Analysis of pseudorandom sequences generated by cellular automata, *Proc. EuroCrypt'91, LNCS*, 547: 186–189, 1991.

[15] A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of applied cryptography*, CRC Press, Boca Raton, FL, 1997.

[16] S. Nandi, B.K. Kar, and P.P Chaudhuri, Theory and applications of cellular automata in cryptography, *IEEE Trans. Comput.* 43: 1346–1357, 1994.

[17] D. Richardson, Tessellation with local transformations, *J. Comput. Syst. Sci.* 6: 373–388, 1972.

[18] M. Seredynski, P. Bouvry, Block encryption using reversible cellular automata, Proc. of ACRI 2004, *LNCS*, 3305: 785–792, 2004.

[19] A. Shamir, How to share a secret, *Commun. ACM* 22: 612–613, 1979.

[20] M. Sipper, *Evolution of Parallel Cellular Machines: The Cellular Programming Approach*, Springer-Verlag, Heidelberg, 1997.

[21] D.R. Stinson, An explication of secret sharing schemes, *Des. Codes Cryptogr.* 2: 357–390, 1992.

[22] D.R. Stinson, *Cryptography Theory and Practice, Second Edition*, CRC Press, Boca Raton, FL, 2002.

[23] M. Tomassini and M. Perrenoud, Cryptography with cellular automata, *Appl. Software Comput.* 1: 151–160, 2001.

[24] S. Wolfram, Random sequence generation by cellular automata, *Adv. Appl. Math.* 7: 123–169, 1986.

[25] S. Wolfram, Cryptography with cellular automata, *Advances in Cryptology: Crypto'85 Proceedings, LNCS*, 218: 429–432, 1986.

Bol. Soc. Esp. Mat. Apl. nº 36(2006), 97–123

A. Calsina y S. Cuadrado

Dinámica de poblaciones estructuradas y evolución fenotípica

Departament de Matemàtiques, Universitat Autònoma de Barcelona

acalsina@mat.uab.es,silvia@mat.uab.es

Resumen

La formulación clásica mediante ecuaciones en derivadas parciales de los problemas de la dinámica de poblaciones estructuradas adolece a menudo de dificultades relacionadas con el tratamiento de distribuciones de población singulares o concentradas. El presente trabajo ofrece una visión de conjunto de una formulación alternativa llamada acumulativa que consiste en considerar distribuciones de población en el espacio de las medidas, la noción de condiciones ambientales determinadas por la propia población y tales que supuestas prescritas convierten las ecuaciones en lineales, y finalmente, la substitución de los sistemas dinámicos basados en tasas instantáneas de cambio por otros construidos a partir de ingredientes referidos a intervalos finitos del tiempo. Se presentan algunos ejemplos de modelización en este contexto, en particular de dinámica de poblaciones estructuradas por variables fisiológicas y fenotípicas (o ecuaciones de selección y mutación) y finalmente se sugiere cómo la formulación acumulativa ofrece un marco unificado para considerar problemas estacionarios y dinámicos para estas ecuaciones en el límite 0 de la tasa de mutación o el tamaño de ésta.

Palabras clave: *Dinámica de poblaciones estructuradas, dinámica adaptativa, formulación acumulativa, ecuaciones de selección y mutación.*

1 Introducción. Ejemplos y características peculiares de la dinámica de poblaciones

El objeto de estudio de la dinámica de poblaciones estructuradas (y a tiempo continuo) son los (semi)sistemas dinámicos continuos cuyas variables de estado

Fecha de recepción: 5/8/2006

(los *estados poblacionales*) son distribuciones de población con respecto a variables (llamadas *estados individuales*) que pueden ser internas, como la edad, el tamaño, el sexo, rasgos fenotípicos diversos que pueden distinguir a especies biológicas, a variedades o incluso simplemente a individuos genéticamente distintos, etc. así como externas como el espacio.
Con más precisión, se considera un espacio de Haussdorf localmente compacto Ω al que se denomina espacio de los estados individuales, siendo éstos, elementos $x \in \Omega$ que reciben el nombre de estados individuales y pueden corresponder, como se ha dicho ya, a edad, tamaño, posición espacial, fenotipo, etc. El espacio de estados (estados poblacionales) del sistema dinámico mencionado, al que designaremos por X, es el espacio vectorial $M(\Omega)$ de las medidas reales acotadas sobre Ω. Ciertamente sólo las medidas positivas tendrán significado biológico, de forma que, para todo subconjunto medible $\omega \subset \Omega$ y toda medida positiva $V \in X$,

$$\int_\omega dV = V(\omega)$$

representa, para una determinada distribución o estado poblacional V, el número total de individuos cuyo estado individual pertenece a ω.

1.1 Estructura discreta

El ejemplo más sencillo de la situación esbozada es el correspondiente a un espacio de estados finito (o más en general, numerable) $\Omega = \{x_1, x_2, .., x_n\}$ (o bien $\Omega = \{x_1, x_2, ..\}$). En este caso, una medida V sobre Ω queda determinada por los pesos de los subconjuntos elementales:

$$V(\{x_i\}) =: v_i$$

de manera que el espacio de estados $M(\Omega)$ es isomorfo a $\mathbb{R}^n$ en el caso finito y a l^1 (el espacio de las sucesiones sumables de números reales) en el caso infinito. Los estados individuales x_i (o más sencillamente los subíndices i) designan especies o bien grupos de edad o de sexo aunque también pueden corresponder al tamaño o número de partículas en modelos de agregados como el descrito brevemente más abajo. Entre los modelos para distintas especies más conocidos recordemos por ejemplo el sistema de Alfred Lotka y Vito Volterra para las poblaciones de presas v_1 y depredadores v_2 en un sistema ecológico:

$$\begin{cases} v_1'(t) &= (a - bv_2(t))v_1(t), \\ v_2'(t) &= (-c + dv_1(t))v_2(t). \end{cases}$$

Aquí los coeficientes a, b, c y d son constantes positivas.
Como ejemplo de estructura con respecto a la edad escribamos el modelo siguiente para las poblaciones de jóvenes (infértiles) v_1 y adultos (fértiles) v_2:

$$\begin{cases} v_1'(t) &= b(x)v_2(t) - xv_1(t) - m_1(v_1(t))v_1(t), \\ v_2'(t) &= xv_1(t) - m_2v_2(t). \end{cases} \tag{1}$$

En este sistema de ecuaciones, x es la tasa de transición del estado inmaduro al maduro, de forma que su inverso T es el valor esperado de la edad de maduración, la fertilidad b se supone función decreciente de x (es decir, creciente con la edad de maduración) y m_1 y m_2 son las mortalidades de jóvenes y adultos respectivamente. El modelo se inspira en el ciclo vital de los insectos con metamorfosis completa en el que cabe esperar una fertilidad creciente con el tamaño al que se alcanza la madurez (por tanto también con la edad en la que se produce ésta), una acusada interacción de competencia por los recursos en la fase larvaria (es decir, una función m_1 creciente) y, en cambio, una mortalidad elevada y relativamente insensible a la densidad de población m_2 en la fase de imago. En modelos como el precedente una pregunta natural que será abordada más adelante en este trabajo es cuál es el valor que la selección natural determinará para el parámetro T supuestos constantes los demás parámetros. En efecto, observemos que un valor demasiado pequeño de T llevará asociada una fertilidad también pequeña mientras que un valor demasiado grande redundará en una tasa de promoción pequeña con lo que el número de adultos será insuficiente. Es de esperar pues la selección de un valor intermedio pero no es claro en este momento un criterio que permita responder a la pregunta planteada.

Los dos sistemas anteriores contienen solamente dos ecuaciones, aunque es claro que pueden generalizarse sin ningún esfuerzo adicional de modelización. Terminemos la subsección sobre ejemplos de estructura discreta citando un modelo para la encefalopatía espongiforme bovina o mal de las vacas locas. Al parecer las proteínas responsables de la temida enfermedad, los priones "enfermos", actúan convirtiendo priones "sanos" en priones "enfermos" los cuales forman agregados de muchas proteínas y son estos agregados proteínicos los culpables de los daños en el tejido nervioso (incidentalmente añadamos que agregados proteínicos como los de las enfermedades causadas por los priones están también presentes en otras enfermedades neurodegenerativas como el Alzheimer). El modelo siguiente ([45]) consta de infinitas ecuaciones, siendo x la densidad de priones sanos e y_i la densidad de agregados de i proteínas enfermas:

$$\begin{cases} x'(t) &= \lambda - dx(t) - \Sigma_{i=1}^{\infty}\beta_i x(t) y_i(t), \\ \\ y_i'(t) &= \beta_{i-1}x(t)y_{i-1}(t) - \beta_i x(t) y_i(t) - a y_i + \\ & \quad \Sigma_{j=i+1}^{\infty}(b_{j,i} + b_{j,i-j})y_j - \Sigma_{j=1}^{i-1} b_{i,j} y_i, \quad i = 1, 2... \end{cases}$$

Aparte de una producción a ritmo constante λ de priones sanos y de la degradación metabólica de priones sanos (con tasa d) y de agregados patógenos (con tasa a mucho menor que d), en las ecuaciones aparecen términos asociados a la incorporación de priones sanos a agregados de tamaño i (con tasa β_i) y a la ruptura de los agregados de tamaño i en dos trozos de tamaño j e $i-j$, con tasa $b_{i,j}$.

1.2 Estructura continua

Por dinámica de poblaciones estructurada a menudo se entiende sólo aquélla en la que la estructura es con respecto a una variable continua y esta situación es la que vamos ahora a considerar de modo introductorio. Consideremos como espacio de estados individuales Ω un dominio de $\mathbb{R}^N$. Los estados individuales en este contexto pueden interpretarse como coordenadas espaciales si la estructura es externa o, cuando ésta es interna, como la edad, el tamaño (longitud, peso, altura, dependiendo de la situación biológica que se quiera retratar), un rasgo morfológico continuo como la longitud del pico de un pinzón en el clásico ejemplo de Charles Darwin en las Islas Galápagos, la proporción de tiempo dedicado por los individuos a conseguir determinado recurso en detrimento de otro ([40], [9]), cierta estrategia etológica, una característica genéticamente fijada del ciclo biológico como en el ejemplo de la edad de maduración ([13], [14]), etc. Generalmente se supone que el estado poblacional es una medida sobre Ω absolutamente continua con respecto a la medida de Lebesgue (aunque después veremos que no siempre es éste el caso). Así podemos poner $dV = v(x)\mathrm{d}x$ siendo v una función de $L^1(\Omega)$.

Seguramente el modelo más conocido de dinámica de poblaciones con estructura espacial sean las ecuaciones de reacción y difusión. Por ejemplo la ecuación siguiente

$$\frac{\partial v}{\partial t}(x,t) = \alpha \Delta v(x,t) + r\Big(1 - \frac{v(x,t)}{K(x)}\Big)v(x,t), \tag{2}$$

modeliza la distribución espacial de la población de una especie sometida a interacción por competencia dada por un término de reacción de tipo logístico, y a difusión aleatoria. Aquí se ha supuesto que la capacidad del medio K depende de la posición espacial $x \in \Omega$, que es un dominio acotado o no de $\mathbb{R}^2$ o $\mathbb{R}^3$. Sobre modelos como el anterior se han publicado multitud de trabajos. Por ejemplo, y en el caso de dos especies e interacción presa-depredador, se han obtenido soluciones estables no constantes por el mecanismo de morfogénesis propuesto ya por Turing para el desarrollo embrionario en [52].

Un modelo clásico de estructura continua con respecto a la edad es el propuesto por Gurtin y McCamy en [32]. Se trata de la siguiente ecuación en derivadas parciales de primer orden no lineal y con términos no locales para una densidad $v(a,t)$:

$$\begin{aligned} \tfrac{\partial v}{\partial t}(a,t) + \tfrac{\partial v}{\partial a}(a,t) + m(a,P(t))v(a,t) &= 0, \\ v(0,t) &= \textstyle\int_0^\infty \beta(a,P(t))v(a,t)\mathrm{d}t, \\ P(t) &:= \textstyle\int_0^\infty v(a,t)\mathrm{d}a \end{aligned} \tag{3}$$

En este modelo el espacio de estados individuales (de edades de los individuos) se toma $\Omega = [0,\infty)$. A diferencia del modelo de difusión espacial, en este segundo caso la suposición de que la distribución de población viene dada

por una densidad, es decir, que es absolutamente continua, es algo artificial, pues tiene perfecto sentido la consideración de un número finito de individuos con exactamente el mismo estado (de la misma edad). Por ejemplo, una medida

$$V = p\delta_{a_0},$$

donde δ_{a_0} es la medida de Dirac en a_0 , representa una población de p individuos de edad a_0 . En el caso anterior de la distribución espacial, una acumulación de individuos en el mismo punto del espacio no suele tener sentido aunque sí lo puede tener en algunos casos. Por ejemplo piénsese en el célebre proceso de quimiotaxis de *Dyctiostelium discoideum* en el que una distribución en dimensión dos culmina con la formación de cuerpos verticales.

1.3 Características especiales de la dinámica de poblaciones

Una de las características peculiares de la dinámica de poblaciones es la conveniencia (a menudo la necesidad) de distinguir entre las causas de los cambios en las distribuciones de población, es decir, entre las diferentes tasas relativas de mortalidad, fertilidad, de transición entre los grupos, velocidades de crecimiento individual, etc. No se hace sin embargo esta distinción en los modelos clásicos de Verhulst (o ecuación logística), de Lotka y Volterra o de competencia incluso cuando éstos incorporan estructura espacial (véase el modelo de reacción y difusión anterior). En cambio tal distinción es imprescindible en dinámica de poblaciones con estructura fisiológica (véase el modelo de Gurtin y McCamy) porque las variables internas como la edad, el tamaño, etc. afectan de distinta forma las diferentes tasas vitales.

Otra de las características típicas de la dinámica de poblaciones es la *identificación* de un conjunto de variables de interacción que reciben a menudo el nombre de *ambiente*, y que tienen la propiedad de que, supuestas conocidas como funciones del tiempo, convierten el sistema dinámico para el estado poblacional en un sistema lineal (no autónomo). Conviene decir enseguida que la palabra *ambiente* no tiene en este contexto el significado habitual de conjunto de condiciones ambientales externas sólo incorporadas como parámetros en el modelo. Aquí por el contrario, el *ambiente* queda determinado por el propio estado poblacional, es decir, es función de la variable de estado. En otras palabras, es un *output* de la distribución de población por un lado, mientras que por otro, determina las tasas vitales jugando el papel de *input* en el sistema lineal.

Por ejemplo, en (3) el ambiente se reduce a la población total $P(t)$ pues supuesta conocida esta función, (3) es un sistema lineal no autónomo. En (2) en cambio, el ambiente es la densidad $v(\cdot, t)$, es decir, el propio estado poblacional, puesto que si se supone conocida esta función y se sustituye su valor en la fracción con denominador $K(x)$, se obtiene una ecuación lineal mientras que no hay ninguna posibilidad esencialmente distinta de escribir una ecuación lineal para v. Son dos situaciones muy diferentes porque en el primer ejemplo la interacción entre individuos queda resumida en una función escalar mientras

en el segundo no hay resumen en absoluto. En lo que sigue supondremos generalmente que el output queda definido por una función lineal

$$L : X \longrightarrow Z$$

donde Z es un espacio de dimensión finita. La hipótesis de linealidad no es restrictiva pero en cambio la finito dimensionalidad de Z sí como hemos visto ya en (2).

Con este planteamiento resulta natural obtener la solución del problema de valor inicial $V_0 \in X$ para un modelo no lineal como la solución del modelo lineal (el *semigrupo lineal para el estado poblacional* $S_I(t)V_0$) para un input (o ambiente) $I \in C([0,T],Z)$ que sea un punto fijo del operador *input* $\rightarrow$ *semigrupo lineal para el estado poblacional* $\rightarrow$ *output* o más formalmente,

$$I \in C([0,T],Z) \rightarrow LS_I(t)V_0 \in C([0,T],Z),$$

véase [21], [22], [24].

1.4 Estructura del trabajo

Lo que resta del trabajo se distribuye como sigue. La sección 2 ofrece una brevísima introducción a una formulación de la dinámica de poblaciones estructuradas, alternativa a la clásica, que evita la consideración de ecuaciones en derivadas parciales y de espacios funcionales complicados y permite obtener sistemas dinámicos en el espacio de las medidas. Las secciones 3 y 4 presentan varios ejemplos de uso de esta nueva formulación en modelos para poblaciones estructuradas por variables internas fisiológicas (como edad y tamaño) y fenotípicas respectivamente. La sección 5 empieza con la formulación acumulativa de dos modelos con estructura fisiológica y fenotípica. En la sección 5.1 se destacan algunas propiedades de la dinámica fisiológica-fenotípica en ausencia de mutación, en particular subrayando la posibilidad de trabajar con distribuciones de población singulares o concentradas, en 5.2 se presenta un resumen de la llamada *dinámica adaptativa* y 5.3 recoge la formulación acumulativa de algunos resultados sobre equilibrios de *ecuaciones de selección y mutación* con la intención de mostrar que la formulación acumulativa ofrece un marco unificado que podría iluminar algunos problemas planteados anteriormente en el límite de tasa o tamaño de la mutación pequeños de las ecuaciones de selección y mutación y su relación con la dinámica adaptativa. Finalmente la sección 6 aventura algunos caminos que sería interesante proseguir.

2 Formulación acumulativa

Trabajos de varios autores han abordado el problema de formular una teoría general de la dinámica de poblaciones. Es especialmente original el enfoque de O. Diekmann *et al* ([21], [22], [26]). En estos artículos los modelos no se formulan en términos de ecuaciones diferenciales o en derivadas parciales, es

decir, en términos de tasas instantáneas de cambio de las distribuciones de población como son los modelos de los ejemplos anteriores, sino mediante los dos ingredientes siguientes:

$u_I(x,\omega) \cong$ probabilidad de que un individuo con estado individual inicial $x \in \Omega$, viva todavía y tenga estado individual en $\omega \subset \Omega$ al final del *input* I.

$\Lambda_I(x,\omega) \cong$ número esperado durante el *input* I de descendientes directos con estado individual en el momento del nacimiento en $\omega \subset \Omega$, de un individuo que inicialmente tiene estado $x \in \Omega$.

u_I y Λ_I son, para cada x, medidas sobre Ω. Aunque en los ejemplos que veremos en las secciones siguientes obtendremos u_I y Λ_I a partir de modelos escritos en la forma más familiar de sistemas de ecuaciones diferenciales, mencionemos que es posible modelizar directamente en términos de la probabilidad u y el valor esperado Λ. El costo añadido de tener que pensar en términos de cantidades acumulativas en vez de tasas instantáneas viene compensado por el hecho de que la formulación acumulativa permite considerar distribuciones de población sin ninguna regularidad (medidas sobre el espacio de estados individuales) mientras que el tratamiento clásico mediante ecuaciones en derivadas parciales exige, para las demostraciones de existencia y unicidad de solución, la pertenencia de las condiciones iniciales a ciertos espacios funcionales más o menos regulares. Además potencialmente permite continuar soluciones inicialmente regulares más allá del tiempo de permanencia en $L^1_{loc.}$.

Supuestos conocidos los dos ingredientes u_I y Λ_I, se pueden obtener, de forma constructiva ([21], [22]), otros dos *núcleos* que son cantidades acumulativas de todas las generaciones:

$\Lambda^c_I(x,\omega) \cong$ número esperado durante el *input* I de descendientes *de todas las generaciones* (hijos, nietos, bisnietos, etc.) con estado individual en el momento del nacimiento en $\omega \subset \Omega$, de un individuo que inicialmente tiene estado $x \in \Omega$.

$u^c_I(x,\omega) \cong$ número esperado de descendientes de todas las generaciones (incluyendo a éste) de un individuo que inicialmente tiene estado $x \in \Omega$, que viven todavía y tienen estado individual en $\omega \subset \Omega$ al final del *input* I.

El superíndice c de estos núcleos es la inicial de la palabra *clan* o descendencia de todas las generaciones. Ahora puede escribirse en forma explícita un semigrupo lineal para la distribución de población durante un input I:

$$(S_I V_0)(\omega) = \int_\omega u^c_I(x,\omega) dV_0(x).$$

Dada una distribución de población inicial (una medida sobre Ω) V_0, la fórmula anterior proporciona el número de individuos con estado en ω después

de transcurrido un tiempo $l(I)$ si la población ha estado sometida al input (al ambiente) I, que es una función continua definida en el intervalo $[0, l(I)]$. El estado poblacional a tiempo $t \in [0, l(I)]$ viene dado por

$$S_I(t)V_0 := S_{\rho(t)I}V_0,$$

donde $\rho(t)I$ se define como la restricción al intervalo $[0, t]$ del input I.

El conjunto $\{S_I : I \in \bigcup_{s=0}^{\infty} C([0, s], Z)\}$ es un semigrupo para la composición, con elemento neutro (la aplicación identidad) correspondiente al input de duración nula, y satisfaciendo la propiedad $S_{I_2 \circ I_1} = S_{I_2} S_{I_1}$ donde $I_2 \circ I_1$ debe interpretarse como el input que se obtiene de la concatenación temporal de los inputs I_1 e I_2.

Como se ha dicho al final de la sección anterior la solución del problema no lineal se aborda ahora como un problema de punto fijo. En efecto, si para una medida inicial V_0 y para algún $T_{V_0} > 0$, I_{V_0} es un punto fijo del operador de input-output:

$$\begin{aligned} C([0, T_{V_0}], Z) &\longrightarrow C([0, T_{V_0}], Z) \\ I &\longrightarrow (t \in [0, T_{V_0}] \to LS_{\rho(t)I}V_0 \in Z), \end{aligned}$$

entonces $S(t, V_0) := S_{\rho(t)I_{V_0}}V_0$ es un semigrupo no lineal en el espacio de las medidas X que proporciona la distribución de población (el estado poblacional) a tiempo $t < T_{V_0}$ si la población inicial es V_0, el modelo de dinámica de poblaciones no lineal viene definido por los ingredientes u y Λ, y la interacción por la función lineal L del espacio de estados X en el espacio de ambientes Z.

Aunque la motivación principal del desarrollo de la formulación acumulativa de la dinámica de poblaciones son las poblaciones estructuradas por variables internas como las que consideramos en las secciones siguientes, mencionemos que también es aplicable a poblaciones con estructura espacial en las que los individuos efectúan movimientos aleatorios y que clásicamente han sido modelizadas mediante ecuaciones en derivadas parciales del tipo de las de reacción y difusión ([21], Sect. 8.3).

3 Poblaciones estructuradas fisiológicamente

La distribución de poblaciones con respecto a variables internas de tipo fisiológico como la edad o el tamaño ha sido objeto de numerosos trabajos, desde el modelo lineal de Sharpe-Lotka [51], pasando por el ya citado (y celebrado) artículo de Gurtin y MacCamy [32], el libro de Metz y Diekmann [43], la monografía sobre poblaciones estructuradas por la edad de Webb [53], así como las más recientes de Cushing [20] y de Iannelli [34], así como muchos artículos en los que desde la pasada década se abordan problemas de fundamentación teórica (de existencia y unicidad de solución del problema de valor inicial) de modelos expresados clásicamente (por medio de ecuaciones en derivadas parciales de primer orden con términos integrales) con velocidad de crecimiento dependiente de la densidad (es decir, expresado en términos clásicos de ecuaciones en

derivadas parciales, con parte principal no lineal -y no local-) ([11], [38], [39], [35]).

Los modelos que aparecen en los últimos trabajos citados toman generalmente formas del tipo siguiente

$$\begin{aligned} &\tfrac{\partial v}{\partial t}(x,t) + \tfrac{\partial}{\partial x}(g(x,I(t))v(x,t)) + m(x,I(t))v(x,t) = 0, \\ &v(x_b,t) = \textstyle\int_{x_b}^{x_m} \beta(x,I(t))v(x,t)\mathrm{d}x, \quad I(t) = \int_{x_b}^{x_m} v(x,t)\mathrm{d}x, \\ &v(x,0) = v_0(x) \end{aligned} \tag{4}$$

Aquí por tanto se supone que la distribución de población con respecto a la variable tamaño $x \in \Omega = [x_b, x_m]$ es una medida absolutamente continua con función densidad $v(\cdot,t) \in L^1(x_b, x_m)$. En el sistema precedente, las tasas vitales tienen el mismo significado que en (3), mientras que el ambiente tiene dimensión 1 para simplificar la exposición. Sin embargo advirtamos que en las referencias citadas Z es n-dimensional e incluso en [38] y [39] es de dimensión infinita. En el lenguaje de la formulación acumulativa, el estado individual es el tamaño y se trata de retratar una situación de crecimiento determinista en la que la velocidad de crecimiento individual g depende del estado individual x y del ambiente I (nótese que si suponemos que I es una función conocida del tiempo entonces (4) se convierte en un sistema lineal.) Así, el movimiento individual en el espacio Ω (el crecimiento individual) está regido por el problema de valor inicial

$$\begin{cases} x'(t) &= g(x(t), I(t)), \\ x(0) &= x_0 \end{cases} \tag{5}$$

cuya solución denotamos $x_I(t; x_0)$. La formulación acumulativa en el espacio de las medidas del problema (4) se basa en la función probabilidad de supervivencia definida por

$$\Pi_I(t) = \exp(-\int_0^t m(x_I(s;x_0), I(s))\mathrm{d}s),$$

que nos da la probabilidad de que un individuo con estado/tamaño inicial x_0 en tiempo 0, sometido a un input I sobreviva hasta tiempo t. Entonces tendremos

$$u_I(x,\omega) = \Pi_I(l(I))\delta_{x_I(l(I);x_0)}(\omega)$$

porque la probabilidad de que un individuo con estado/tamaño inicial x_0 viva todavía al final del input I y tenga estado individual (tamaño) en ω es igual a la probabilidad de supervivencia a tiempo $l(I)$ si ω contiene el tamaño que seguro alcanzará (determinado por la solución de (5)) y vale 0 en caso contrario. Por otra parte, el número esperado de hijos con tamaño en el momento inicial en ω, durante el input I, que tendrá un individuo con tamaño inicial x_0 es

$$\Lambda_I(x,\omega) = \int_0^{l(I)} \beta(x_I(s;x_0), I(s))\Pi_I(s)\mathrm{d}s\delta_{x_b}(\omega)$$

siendo ahora la presencia de la medida de Dirac en x_b debida a que el tamaño en el nacimiento es siempre x_b.

4 Poblaciones estructuradas fenotípicamente

Una categoría de modelos de dinámica de poblaciones se dedica al estudio de la evolución biológica. El interés general por este tema no es sólo teórico y ligado a escalas de tiempo muy grandes como lo demuestra cada año el siempre cambiante virus de la gripe, o las enormes dificultades que representa el descubrimiento de una vacuna contra el VIH o las cada día más preocupantes resistencias bacterianas a los antibióticos En esta clase pueden incluirse desde los conocidos modelos de competencia que justifican el principio de exclusión competitiva (véase por ejemplo [33]), pasando por los modelos de genética de poblaciones debidos entre otros a R.A Fisher que sustentaron el neodarwinismo como síntesis de la teoría de la evolución por mutación aleatoria y selección natural y la teoría de la herencia de Georg Mendel (véase [27]), hasta la mucho más reciente dinámica adaptativa ([44], [28]). También pertenecen a esta clase de modelos de dinámica de poblaciones los trabajos sobre las llamadas ecuaciones de selección y mutación en los que se consideran distribuciones de población con respecto a variables fenotípicas continuas genéticamente fijadas tales como características anatómicas particulares (recuérdese el ejemplo clásico de la longitud del pico de los pinzones en las Islas Galápagos) o bien parámetros definitorios del ciclo biológico como la edad de maduración, o la edad de cambio de sexo (en especies hermafroditas secuenciales), o características etológicas como por ejemplo el tiempo dedicado a la búsqueda de cierto recurso alimenticio (en detrimento de otro), o la tendencia a cierto comportamiento especial como por ejemplo el canibalismo, o combinaciones de estas variables. En el contexto de la genética de poblaciones pueden encontrarse ejemplos de ecuaciones para densidades de población con respecto al genotipo modelizando la mutación en trabajos de Crow y Kimura [19], [36] y más tarde de R. Bürger y colaboradores ([5]) y en un contexto más ecológico, en el que a menudo aparece explícita la interpretación fenotípica de la variable evolutiva y las ecuaciones son más propiamente de selección y mutación, en [8], [9], [10], [41], [49], [14], [15], [25], [16], [46], [31], [17], [18].

Un ejemplo de formulación clásica (mediante densidades de población) de un problema de dinámica de poblaciones estructuradas (únicamente) por el fenotipo es la ecuación siguiente (véase [16])

$$\begin{aligned} \frac{\partial v}{\partial t}(x,t) &= (1-\varepsilon(x))b(x)v(x,t) + \int_0^1 \beta(x,y)\varepsilon(y)b(y)v(y,t)\mathrm{d}y - \\ &\quad m(x,I(t))v(x,t), \qquad (6) \\ I(t) &= \int_0^1 v(x,t)\mathrm{d}x, \quad v \in L^1(0,1) \end{aligned}$$

Aquí el espacio de estados individuales Ω se reduce al intervalo $[0,1]$,

$\varepsilon(x) \in [0,1]$ es la probabilidad de mutación en cada reproducción de un individuo de fenotipo (o estado individual) x, $b(x)$ es la fertilidad específica para el fenotipo x, $b(x)$ es la fertilidad específica para el fenotipo x, y $\beta(\cdot, y)$ es la densidad de probabilidad del fenotipo del hijo de un individuo de fenotipo y en el caso en que se produce una mutación.

La formulación acumulativa del modelo precedente en el espacio X de medidas sobre $[0,1]$ se basa en los dos ingredientes mencionados en la sección 2. Primeramente, en

$$u_I(x,\omega) = \exp(-\int_0^{l(I)} m(x, I(s))\mathrm{d}s)\delta_x(\omega),$$

cuya interpretación es la de la probabilidad de que un individuo de fenotipo (en el instante inicial) x, viva todavía (el primer factor no es más que la probabilidad de supervivencia) y tenga fenotipo en ω después de un input I. Nótese que, a diferencia de la sección 3, la presencia de la medida de Dirac obedece aquí a que un estado fenotípico es invariable a lo largo de la vida de un individuo. Y en segundo lugar, en

$$\Lambda_I(x,\omega) = \int_0^{l(I)} b(x) e^{-\int_0^\tau m(x,I(s))\mathrm{d}s} \mathrm{d}\tau((1-\varepsilon(x))\delta_x(\omega) + \varepsilon(x)\int_\omega \beta(y,x)\mathrm{d}y),$$

que da el número esperado de hijos con fenotipo en ω durante un input I, de un individuo de fenotipo (inicial) x. En esta segunda fórmula la delta de Dirac corresponde a los nacimientos sin mutación y $(1-\varepsilon(x))\delta_x(\omega)+\varepsilon(x)\int_\omega \beta(y,x)\mathrm{d}y$ es la medida de probabilidad del fenotipo de los descendientes directos de un individuo de fenotipo x.

5 Poblaciones estructuradas fisiológica y fenotípicamente

En lo que resta del trabajo vamos a concentrarnos en los modelos de dinámica de poblaciones que consideran densidades o, más en general, distribuciones de población, con respecto a ambas estructuras internas, la fisiológica y la fenotípica. Este es sin duda el marco adecuado para el estudio de la evolución de los parámetros de los ciclos biológicos (edad de maduración, edad de cambio de sexo, etc.) pero también resulta imprescindible si el efecto de la característica fenotípica depende de forma relevante de alguna característica fisiológica (éste es por ejemplo el caso de la tendencia al canibalismo, cuyos efectos sobre la población dependen obviamente de la estructura en tamaño de ésta).

Distinguiremos pues dos espacios de estados individuales, Ω_0 y Ω_1, para las estructuras fisiológica y fenotípica respectivamente de forma que el espacio (completo) de los estados individuales será $\Omega = \Omega_0 \times \Omega_1$. Por supuesto el modelo (6) puede considerarse en este contexto con estructura fisiológica trivial, es decir, siendo Ω_0 un conjunto de un solo elemento, y $\Omega_1 = [0,1]$. Análogamente, las poblaciones estructuradas fisiológicamente corresponden a espacios Ω_1 de un solo elemento.

Un ejemplo de estructura fisiológica-fenotípica en sentido más propio es un modelo para la edad de maduración contenido en [15]. Se distinguen en este modelo dos grupos de edad, jóvenes inmaduros y adultos, de forma que podemos poner $\Omega_0 = \{1, 2\}$ y una variable fenotípica continua que es (por conveniencia) el inverso de la edad de maduración o, equivalentemente, la tasa de transición del estado juvenil al estado adulto (por lo tanto, por ejemplo, $\Omega_1 = [0, \infty)$). Si suponemos que con respecto a esta segunda variable la distribución de población es absolutamente continua, entonces se puede escribir el sistema siguiente para las densidades de jóvenes y adultos respectivamente

$$\begin{cases} \frac{\partial v_1}{\partial t}(x,t) &= \int_0^\infty \beta^\varepsilon(x,y) b(y) v_2(y,t) \mathrm{d}y - (x + m_1(I_1(t))) v_1(x,t), \\ \frac{\partial v_2}{\partial t}(x,t) &= x v_1(x,t) - m_2(I_2(t)) v_2(x,t), \end{cases} \tag{7}$$

con $I_1(t) = \int_0^\infty v_1(y,t)\mathrm{d}y$ y $I_2(t) = \int_0^\infty v_2(y,t)\mathrm{d}y$ las poblaciones totales. El sistema (7) es una versión con estructura fenotípica de (1), aunque allí se suponía constante m_2. El núcleo $\beta^\varepsilon(\cdot, y)$ es la densidad de probabilidad del fenotipo x del hijo de un individuo de tipo y y el parámetro ε puede interpretarse como una medida del tamaño medio de las mutaciones (por ejemplo si ε es el valor esperado de $|x - y|$, que se supone independiente de y). La formulación acumulativa del modelo (7) pasa por las siguientes definiciones, todas basadas en el hecho de que (7) supone implícitamente distribuciones exponenciales e independencia estocástica:

$u_I((1,x), \{1\} \times \omega) \cong$ probabilidad de que un individuo joven de fenotipo (inicial) x permanezca joven (y vivo) y su fenotipo pertenezca a ω al final de un input $I = (I_1, I_2) =$

$$e^{-xl(I)} e^{-\int_0^{l(I)} m_1(I_1(s))\mathrm{d}s} \delta_x(\omega),$$

$u_I((1,x), \{2\} \times \omega) \cong$ probabilidad de que un individuo joven de fenotipo (inicial) x haya madurado, viva todavía y su fenotipo pertenezca a ω después de un input $I = (I_1, I_2) =$

$$\int_0^{l(I)} x e^{-xs} e^{-\int_0^s m_1(I_1(\sigma))\mathrm{d}\sigma} e^{-\int_s^{l(I)} m_2(I_2(\sigma))\mathrm{d}\sigma} \mathrm{d}s \cdot \delta_x(\omega),$$

$u_I((2,x), \{1\} \times \omega) = 0$ por razones obvias,

$u_I((2,x), \{2\} \times \omega) \cong$ probabilidad de que un adulto de fenotipo (inicial) x, viva todavía y su fenotipo pertenezca a ω después de un input $I = (I_1, I_2) =$

$$e^{-\int_0^{l(I)} m_2(I_2(s))\mathrm{d}s} \delta_x(\omega).$$

$\Lambda_I((1,x),\{1\}\times\omega) \cong$ número esperado de hijos con fenotipo (edad de maduración) en ω de un individuo joven de fenotipo x durante un input $I =$

$$b(x)\int_0^{l(I)}\int_0^{l(I)-t} xe^{-xs}e^{-\int_0^s m_1(I_1(\sigma))\mathrm{d}\sigma}e^{-\int_s^{s+t} m_2(I_2(\sigma))\mathrm{d}\sigma}\mathrm{d}s\mathrm{d}t\cdot B_x^\varepsilon(\omega),$$

$\Lambda_I((2,x),\{1\}\times\omega) \cong$ número esperado de hijos con fenotipo (edad de maduración) en ω de un adulto de fenotipo x durante un input $I =$

$$b(x)\int_0^{l(I)} e^{-\int_0^t m_2(I_2(\sigma))\mathrm{d}\sigma}\mathrm{d}t\cdot B_x^\varepsilon(\omega).$$

En las últimas fórmulas, el número total esperado de hijos de un individuo es el producto de la fertilidad b por unidad de tiempo por el valor esperado del tiempo de vida como adulto T_a *durante* el input I y éste se calcula integrando la función $P(T_a \geq t)$ sobre el intervalo de duración de I. Por otro lado, B_x^ε es la medida de probabilidad del estado individual en el nacimiento del hijo de un individuo de fenotipo x. Como en (7) se suponía que (la segunda componente) de esta variable aleatoria era absolutamente continua con densidad $\beta^\varepsilon(\cdot,x)$, tendríamos de hecho aquí que $\mathrm{d}B_x^\varepsilon(y) = \beta^\varepsilon(y,x)\mathrm{d}y$. Sin embargo tienen sentido otras posibilidades como ya se ha observado en (6), donde teníamos $B_x^\varepsilon(\omega) = (1-\varepsilon(x))\delta_x(\omega) + \varepsilon(x)\int_\omega \beta(y,x)\mathrm{d}y$ pues suponíamos una probabilidad positiva $1-\varepsilon(x)$ de reproducción sin error. Aquí y en todo lo que sigue $\varepsilon = 0$ corresponde a que $B_x^0 = \delta_x$, es decir a que la reproducción es siempre sin error o mutación.

Finalmente, $\Lambda_I((1,x),\{2\}\times\omega) = \Lambda_I((2,x),\{2\}\times\omega) = 0$ por razones obvias como antes.

Un segundo ejemplo de población estructurada fisiológica y fenotípicamente, biológicamente muy parecido a (7), es, en formulación clásica, el sistema (véase [46])

$$\begin{gathered}\tfrac{\partial v}{\partial t}(a,x,t) + \tfrac{\partial v}{\partial a}(a,x,t) + m(a,I_1(t),I_2(t))v(a,x,t) = 0\\ v(0,x,t) = \int_0^\infty\int_y^\infty b(a,y)\beta^\varepsilon(x,y)v(a,y,t)\mathrm{d}a\mathrm{d}y,\\ I_1(t) = \int_0^\infty\int_0^y v(a,y,t)\mathrm{d}a\mathrm{d}y,\quad I_2(t) = \int_0^\infty\int_y^\infty v(a,y,t)\mathrm{d}a\mathrm{d}y.\end{gathered}\tag{8}$$

Aquí los estados individuales (a,x) pertenecen a $\Omega = [0,\infty)^2$, siendo a la edad fisiológica y x la edad de maduración (no su inverso como en (7)). Además, con respecto a (7) hemos pasado de la consideración de sólo dos grupos de edad a considerar ésta como una variable continua. Pensamos de momento en distribuciones de población absolutamente continuas, o equivalentemente, en densidades $v(\cdot,\cdot,t)$ en $L^1(\Omega)$. A la vista de (7), el único término de (8) que merece algún comentario es el del número de nacimientos por unidad de tiempo o condición de frontera en $a = 0$ (la segunda ecuación de (8)). Ahora la fertilidad

$b(a,x)$ se supone nula para $a < x$ y función creciente de la edad de maduración como en (1) y en (7). El ambiente es bidimensional, como en (7) y viene definido por las poblaciones totales de jóvenes y de adultos I_1 e I_2 respectivamente.

La formulación acumulativa de (8) en el espacio de estados poblacionales X de las medidas sobre Ω es como sigue:

$u_I((a,x),\omega_0 \times \omega_1) \cong$ probabilidad de que un individuo de edad inicial a y fenotipo-edad de maduración (inicial) x viva todavía, tenga edad en ω_0 y fenotipo en ω_1 después de un input $I = (I_1, I_2) =$

$$e^{-\int_0^{l(I)} m(a+s,I_1(s),I_2(s))\mathrm{d}s} \cdot \delta_{a+l(I)}(\omega^0) \cdot \delta_x(\omega^1).$$

En esta expresión y debido al "movimiento" determinista de los individuos en el espacio de edades y a la "inmovilidad" en el de los fenotipos, simplemente se multiplica la probabilidad de supervivencia por las medidas de Dirac.

$\Lambda_I((a,x),\omega_0 \times \omega_1) \cong$ número esperado durante el input I de hijos con edad en el nacimiento en ω_0 y fenotipo (inicial) en ω_1, de un individuo que inicialmente tiene edad a y fenotipo $x =$

$$\int_0^{l(I)} e^{-\int_0^t m(a+s,I_1(s),I_2(s))\mathrm{d}s} b(a+t,x)\mathrm{d}t \cdot \delta_0(\omega^0) \cdot B_x^\varepsilon(\omega_1). \tag{9}$$

Aquí en cambio, el número total esperado de hijos (el primer factor) puede obtenerse de la fórmula

$$E(f(T)) = \int_0^\infty P(T \geq t) f'(t)\mathrm{d}t$$

para el valor esperado de $f(T)$ donde T es una variable aleatoria positiva (en nuestro caso el tiempo de vida a partir del instante inicial y durante el input I) y $f(T)$ una función nula en 0 y creciente (en nuestro caso el número de hijos si el individuo vive hasta el instante T). El segundo factor simplemente refleja el hecho que todos los individuos nacen con edad cero y el tercero es la probabilidad de que un hijo de un individuo con fenotipo x tenga fenotipo en ω_1.

5.1 Subespacios invariantes triviales de la dinámica fisiológica-fenotípica. Ausencia de mutación

Un caso especial de dinámica fisiológica-fenotípica se da en ausencia de mutación, es decir, cuando ε (la medida de la mutación o la probabilidad de ésta) es 0. En la notación de antes, tendremos $B_x^0 = \delta_x$ y, si S^ε es el semigrupo de dinámica de poblaciones estructuradas fisiológica-fenotípicamente, hablaremos en esta sección y en la siguiente de S^0. En este caso, para todo subconjunto medible ω_1 del espacio fenotípico Ω_1, una población inicial cuyos miembros tengan todos su estado individual inicial fenotípico en ω_1 conservará esta

propiedad a lo largo del tiempo pues sin mutación no hay aparición de nuevos fenotipos. En otras palabras, el subespacio vectorial X_{ω_1} del espacio de estados X de las medidas con soporte en $\Omega_0 \times \omega_1$ es invariante por la dinámica fisiológica-fenotípica si $\varepsilon = 0$ (la dada por S^0). En particular, si ω_1 se reduce a un único punto $x \in \Omega_1$, $X_{\{x\}}$ es obviamente isomorfo al espacio de medidas sobre Ω_0 y corresponde al espacio de estados de la dinámica fisiológica (también llamada en algunos contextos puramente ecológica o de selección natural), mientras el rasgo evolutivo (el fenotipo) x juega el papel de un parámetro del sistema. Así, si además el espacio de estados fisiológicos Ω_0 es un conjunto de m elementos, el modelo (formulado clásicamente) se reduce a un sistema de m ecuaciones diferenciales ordinarias. De forma análoga, si $\omega_1 = \{x^1, x^2, .., x^n\}$, entonces el subespacio invariante por S_0, X_{ω_1}, es isomorfo a $(M(\Omega_0))^n$ y la dinámica corresponde a un modelo de selección pura con n fenotipos (especies) diferentes, que, como antes, se reduce a un sistema de $m \times n$ ecuaciones diferenciales ordinarias si Ω_0 es un conjunto de m elementos.

Sin mutación y por tanto en el contexto de esta subsección citemos los trabajos de A. Ackleh *et al* ([1], [2]) quienes consideran un modelo de tipo logístico con estructura (sólo) fenotípica. La variable fenotípica es bidimensional y sus componentes x_1 y x_2 vienen a coincidir esencialmente con las tasas relativas de fertilidad y mortalidad respectivamente. Es decir, que, en nuestra notación, $\Omega_0 = \{1\}$ y $\Omega_1 = [0, 1]^2$. En el segundo de los trabajos se considera el problema de valor inicial en el espacio de las medidas X y para cada subconjunto cerrado ω_1 contenido en el interior de Ω_1 se prueba que, para condiciones iniciales V_0 en X_{ω_1}, la solución tiende cuando el tiempo tiende a infinito a una medida de Dirac concentrada en el punto de máximo x^* en ω_1 del cociente entre las tasas de fertilidad y mortalidad respectivamente $(\frac{x_1}{x_2})$ si la medida por V_0 de todo entorno de x^* es positiva. El cociente $\frac{x_1}{x_2}$ viene a ser la ratio de reproducción y por ello este resultado tiene muchos puntos en común con otros que se expondrán en la sección 5.3.

5.2 Dinámica adaptativa. Una visión particular

Una respuesta muy interesante a preguntas como la que nos hacíamos en la sección 1.1 sobre el valor que la selección natural determina para las variables evolutivas que aparecen como parámetros en los sistemas ecológicos fue descubierta por Maynard-Smith y colaboradores ([42], [40]), basándose en conceptos procedentes de la teoría de juegos. En términos biológicos se define una *estrategia evolutivamente estable* global (abreviadamente ESS por *evolutionarily stable strategy*) como un valor $\hat{x} \in \Omega_1$ de un fenotipo o rasgo evolutivo, tal que una población (*residente*) cuyos individuos son *todos* del tipo $\hat{x}$ y se encuentra en equilibrio asintóticamente estable (más en general, en el atractor del sistema ecológico) no puede ser invadida por una población *mutante pequeña* cuyos individuos son *todos* de un tipo distinto y. Una $\hat{x}$ se llama ESS local si lo anterior es cierto para y en un entorno en Ω_1 de $\hat{x}$.

En el lenguaje de las secciones anteriores, podemos decir que un estado individual $\hat{x} \in \Omega_1$ es una ESS (global) si existe una medida $V_{\hat{x}} \in M(\Omega_0)$ tal

que $V_{\hat{x}} \times \delta_{\hat{x}} \in M(\Omega)$ es un estado poblacional de equilibrio del semigrupo S^0 de dinámica de poblaciones estructuradas fisiológica-fenotípicamente con $\varepsilon = 0$ (es decir, sin mutación) que es asintóticamente estable restringido al subespacio invariante $X_{\{\hat{x},y\}}$ cualquiera que sea $y \in \Omega_1$. Puede probarse, por lo menos en casos sencillos para los que valga el principio de estabilidad lineal (por ejemplo si Ω_0 es finito), que esto es equivalente a exigir que $V_{\hat{x}} \times \delta_{\hat{x}}$ sea asintóticamente estable como equilibrio del semigrupo S^0 restringido al subespacio $X_{\{\hat{x}\}}$ (es decir, para la dinámica fisiológica o puramente ecológica) y además que, si llamamos $\hat{I} = L(V_{\hat{x}} \times \delta_{\hat{x}})$, es decir, si $\hat{I}$ es el ambiente determinado por la población residente, entonces $S^0_{\hat{I}}(t)$ (usando la notación de la sección 2, nótese sin embargo que aquí $\hat{I}$ es una función constante) restringido al subespacio invariante (también por $S^0_{\hat{I}}(t)$) $X_{\{y\}}$, es un semigrupo lineal fuertemente continuo con cota de crecimiento negativa (es decir, es exponencialmente estable) cualquiera que sea $y \neq \hat{x}$. Son obvios los cambios necesarios en el párrafo precedente si se consideran ESS locales.

Definamos ahora la llamada función *fitness* $\tau(x, y)$ en Ω_1^2 como la cota de crecimiento del semigrupo lineal al que se alude en el parágrafo anterior , es decir, la cota de crecimiento de $S^0_{I_x}(t)$ restringido a $X_{\{y\}}$, donde $I_x = L(V_x \times \delta_x)$, suponiendo ahora la existencia del estado de equilibrio para la dinámica fisiológica V_x para cualquier $x \in \Omega_1$. La interpretación biológica de la función *fitness* sería la tasa de crecimiento relativa máxima de la población invasora de fenotipo y sometida a un input o ambiente constante determinado por el equilibrio (asintóticamente estable) de la población residente con fenotipo x. En particular tendremos que, como $V_x \times \delta_x$ es un estado de equilibrio,

$$V_x \times \delta_x = S(t, V_x \times \delta_x) = S_{I_x}(t)(V_x \times \delta_x).$$

Esto puede interpretarse en el sentido de que, para cualquier $t > 0$, $V_x \in M(\Omega_0)$ es un vector propio positivo de valor propio 1 de un operador positivo en $M(\Omega_0)$. Bajo ciertas hipótesis de irreducibilidad y compacidad vale en esta situación el teorema de Perron-Frobenius en dimensión infinita (véase [[50], [3]) y como consecuencia el radio espectral del operador $S_{I_x}(t)$ restringido a $X_{\{x\}}$ es 1 y su cota de crecimiento, que es el logaritmo del radio espectral, 0. Así resulta $\tau(x, x) = 0$ para todo $x \in \Omega_1$.

De la definición de ESS se deduce que la función $y \to \tau(\hat{x}, y)$ tiene un (único) punto de máximo absoluto estricto en $\hat{x}$ si y sólo si $\hat{x}$ es una ESS (pues $\tau(\hat{x}, y) < \tau(\hat{x}, \hat{x}) = 0$ si $y \neq \hat{x}$). Por lo tanto $(\hat{x}, \hat{x})$ es un punto crítico, de hecho un punto de silla, de la función $\tau(x, y)$ si $\hat{x}$ es una ESS.

Los trabajos [28], [44] llevan a cabo un estudio de la interpretación en términos de adaptabilidad biológica del carácter de los puntos críticos regulares de τ sobre la bisectriz de Ω_1^2, en especial cuando Ω_1 es unidimensional.

En el contexto de la *dinámica adaptativa*, un fenotipo x tal que (x, x) es un punto crítico de τ recibe el nombre de estrategia singular. Así toda ESS es una estrategia singular pero hay estrategias singulares que no son ESS. Por ejemplo pueden tratarse de mínimos (o no ser extremos) y no de máximos de la fitness con respecto a la segunda variable. Sin embargo las estrategias singulares son

relevantes desde el punto de vista adaptativo incluso cuando no son ESS. Por ejemplo, se introduce el concepto de estrategia atractora (*convergence stable strategy o CSS*). Una estrategia singular $\hat{x}$ se llama CSS si cualquier población residente con estrategia x en un entorno de $\hat{x}$ puede ser invadida por una población mutante con una estrategia y más próxima (en un sentido que debe precisarse en Ω_1) a $\hat{x}$ que x. Aquí *puede ser invadida* significa simplemente que $\tau(x,y) > 0$. Sin embargo, si $\hat{x}$ es también ESS y la dinámica de S^0 en el subespacio $X_{\{x,y\}}$ es sencilla, es decir, si los equilibrios (positivos) son sólo $V_x \times \delta_x$ y $V_y \times \delta_y$ y las soluciones tienden necesariamente a uno de los dos, como la condición $\tau(x,y) > 0$ implica que el primero es inestable entonces las soluciones tenderán al segundo (en particular deberá ser genéricamente $\tau(y,x) < 0$). La interpretación de este hecho es que una población residente con estrategia x es *sustituida* por una población invasora con estrategia y (que a su vez se convierte en residente). Si $\hat{x}$ es una estrategia evolutivamente estable y atractora, eventualmente se produce entonces una sucesión de sustitución de fenotipos: x (la correspondiente población de equilibrio $V_x \times \delta_x$) es sustituida por y (por $V_y \times \delta_y$), y por z, etc., cada uno de los fenotipos más cercanos a $\hat{x}$ que el anterior, y finalmente convergiendo a $\hat{x}$ (a la población de equilibrio $V_{\hat{x}} \times \delta_{\hat{x}}$). Cuando el espacio fenotípico Ω_1 es unidimensional, un análisis de la función $\tau(x,y)$ en un punto crítico $(\hat{x},\hat{x})$ (véase [44] Sect. 3.2) revela que las condiciones $\frac{\partial^2 \tau}{\partial y^2}(\hat{x},\hat{x}) < 0$ y $\frac{\partial^2 \tau}{\partial y^2}(\hat{x},\hat{x}) < \frac{\partial^2 \tau}{\partial x^2}(\hat{x},\hat{x})$ implican que $\hat{x}$ es una estrategia evolutivamente estable local y atractora o simplemente un *atractor evolutivo* (local). Si se cumple la primera de estas condiciones pero no la segunda, $\hat{x}$ es una ESS local (o sea, un "equilibrio" de la dinámica adaptativa) pero no es una estrategia atractora. Si, en cambio, se cumple sólo la segunda, $\hat{x}$ es llamado un punto de ramificación (*branching point*) porque la sucesión de sustituciones descrita anteriormente acaba produciendo una población dimórfica (véase [28]).

Con el resumen visto aquí, el lector debería llevarse acertadamente la impresión de que la dinámica adaptativa es una herramienta muy útil para la compresión del proceso de adaptación de los rasgos evolutivos cuantitativos en muchos contextos. Sin embargo pueden destacarse algunos inconvenientes de este modelo del cambio evolutivo. En primer lugar, y en el lenguaje de las secciones anteriores, mencionemos que la dinámica adaptativa no se basa en la construcción de un verdadero sistema dinámico en el espacio de estados $X = M(\Omega_0 \times \Omega_1)$ sino en la consideración de, en cada caso, una *sucesión* de dinámicas en los subespacios $X_{\{x,y\}}$ invariantes para el semigrupo S^0 (sin mutación) intercalada con la aparición de mutaciones (o invasiones) que, en cada paso, deben proporcionar el rasgo mutante y. Por lo tanto supone una separación completa de las escalas temporales: el *tiempo ecológico* durante el cual se produce una dinámica rápida en el espacio $X_{\{x,y\}}$ y el *tiempo evolutivo* en el que tiene lugar una lenta sucesión de mutaciones. Además la dinámica (rápida) debe ser sencilla de forma que el éxito inmediato de una población pequeña mutante de tipo y cuando la población residente es de tipo x ($\tau(x,y) > 0$) añadido al fracaso del tipo x tratando de invadir una población residente y ($\tau(y,x) < 0$) garantice que $V_y \times \delta_y$ es efectivamente el único atractor de S^0 restringido a $X_{\{x,y\}}$. En

fin, requiere que la tasa de mutación sea pequeña para que sea despreciable la posibilidad de aparición de nuevos fenotipos antes de que "termine" la dinámica en $X_{\{x,y\}}$.

Los requisitos anteriores no se cumplen en muchas situaciones de importancia biológica. Por ejemplo es obvia la coexistencia de muchísimos fenotipos cuando el carácter analizado es del tipo cuantitativo o continuo. Además, la tasa de mutación es significativamente elevada en muchos casos, especialmente tratándose de microorganismos, y es por ejemplo conocida la llamada a veces carrera armamentista (de mutación y selección) que se establece entre presas y depredadores (por ejemplo entre bacterias y virus bacteriófagos) o entre las células del sistema inmunitario y los agentes infecciosos. En estos procesos parece justificado el evitar la separación de escalas de tiempo y, aun a costa de arrastrar mayores dificultades técnicas, considerar, como haremos brevemente en la sección siguiente, la dinámica en el espacio completo X para $\varepsilon > 0$, es decir, con mutación.

5.3 Dinámica fisiológica-fenotípica con ε pequeño

En esta sección vamos a reseñar algunos de los resultados que se conocen sobre los procesos de cambios poblacionales en poblaciones con estructura fenotípica continua (Ω_1 un dominio de $\mathbb{R}^N$) y posiblemente con estructura fisiológica no trivial, en especial sobre los equilibrios de estos procesos. El caso de estructura fisiológica trivial (Ω_0 tiene un único elemento) y estructura fenotípica continua (Ω_1 el conjunto de los números reales) puede encontrarse ya en [36] donde se formula (cf. Sección 4) una ecuación para la densidad de probabilidad del efecto alélico promedio sobre el carácter fenotípico considerado. La ecuación tiene la forma

$$\frac{\partial p(x,t)}{\partial t} = (m(x) - \bar{m}(t))p(x,t) + \int_{\Omega_1} u(x,y)p(y,t)\mathrm{d}y - u_1(x)p(x,t), \qquad (10)$$

donde $m(x)$ es la función fitness Malthusiana (con un significado esencialmente equivalente al de la sección anterior), $\bar{m}(t) = \int_{\Omega_1} m(x)p(x,t)\mathrm{d}x$ denota la fitness promedio, $u(x,y)$ es la fracción de individuos de tipo x originados por mutación de individuos de tipo y por unidad de tiempo y, $u_1(x) = \int_{\Omega_1} u(y,x)\mathrm{d}y$, es la tasa relativa de mutación del tipo x. Aunque en el libro citado m y u tienen formas especiales, en [4] se prueba, bajo ciertas hipótesis que incluyen que el óptimo de m sea un punto "relativamente" regular, pero por lo demás, para m y u generales, que (10) tiene un (único) equilibrio (una densidad en L^1) globalmente atractor. En [5] (véase también [6]) se formula la ecuación (10) para medidas sobre Ω_1 y se prueba la existencia de soluciones estacionarias que no son absolutamente continuas cuando el punto de óptimo de la función m tiene una cúspide lo suficientemente pronunciada.

En los trabajos [7], [8], [10] se proponen modelos para densidades con respecto a variables evolutivas desde un punto de vista más ecológico. También la estructura fisiológica es trivial y la mutación se modeliza mediante un operador de difusión (hay que apuntar aquí que en [36] ya se aproxima la

ecuación (10) mediante una ecuación de difusión). En [9] se relacionan por primera vez los equilibrios de las ecuaciones de selección y mutación con tasa de mutación pequeña con las estrategias evolutivamente estables del sistema ecológico subyacente (en este caso un sistema presa-depredador). Más recientemente pueden citarse en este tema, aun con estructura fisiológica trivial [49], [41].

El estudio de las poblaciones de equilibrio del semigrupo S^ε se reduce, en muchos casos interesantes, a la determinación de los puntos fijos de una función definida en $\mathbb{R}^+ \times Z$ (y recuérdese que generalmente el espacio de los ambientes Z es de dimensión finita). En efecto en [23], véase también [21], se prueba que, si la esperanza de vida es acotada uniformemente con respecto a los estados individuales, una distribución de población V^ε en X es de equilibrio si y sólo si viene dada por

$$V^\varepsilon(\omega) = \int_0^\infty \int_{\Omega_b} u_{\rho(a)I}(x,\omega)\mathrm{d}m(x)\mathrm{d}a, \tag{11}$$

donde m es una medida en el subconjunto $\Omega_b \subset \Omega$ de los estados individuales en el nacimiento, interpretable como la distribución de la tasa de nacimientos, que es punto fijo del operador lineal

$$(T_I m)(\omega) = \int_{\Omega_b} \Lambda_I(x,\omega)\mathrm{d}m(x),$$

y donde I es un "input (constante, pues estamos interesados en equilibrios) definido en $[0,\infty)$, y punto fijo de la aplicación input-output.

$T_I m$ recibe el nombre de operador de la "generación siguiente" porque proporciona la distribución esperada de nacimientos en Ω_b de hijos de los individuos de una distribución dada m en Ω_b.

Para uso posterior, digamos que, dado I constante y bajo condiciones apropiadas, el radio espectral $R_0(I)$ del operador lineal T_I es menor que 1 (mayor que 1) si y sólo si la cota de crecimiento del semigrupo lineal $S_I(t)$ es menor que 0 (mayor que 0)([21]).

En el caso de poblaciones estructuradas fisiológica y fenotípicamente, el conjunto de los estados individuales en el nacimiento es a menudo de la forma $\Omega_b = \{x_b\} \times \Omega_1$. Este es el caso por ejemplo para estructura fisiológica con respecto a la edad (entonces $x_b = 0$), o, para poblaciones estructuradas por el tamaño, si se puede suponer que todos los individuos nacen con el mismo tamaño. Por otro lado, cualquier estado fenotípico debe ser posible en el nacimiento puesto que no cambia a lo largo de la vida. Así puede pensarse el operador $T_I^\varepsilon := T_I$ como un operador lineal en el espacio de las distribuciones de población sólo con respecto al fenotipo, $M(\Omega_1)$. Como Λ_I es de la forma siguiente (compárese con (9))

$$\Lambda_I((x_b,x_1),\omega_0 \times \omega_1) = r_I(x_1)\delta_{x_b}(\omega_0)B^\varepsilon_{x_1}(\omega_1),$$

donde $r_I(x_1)$ es el número esperado total de hijos de un individuo de fenotipo x_1, se tiene, para $m \in M(\Omega_1)$ y $\omega_1 \in \Omega_1$,

$$(T_I^\varepsilon m)(\omega_1) = \int_{\Omega_1} r_I(x_1) B_{x_1}^\varepsilon(\omega_1) \mathrm{d}m(x_1). \tag{12}$$

Para $\varepsilon > 0$, es decir, en presencia de mutación, generalmente se supone que la medida de probabilidad B_x^ε es tal que el operador T_I^ε resulta además de positivo, irreducible. La interpretación biológica de la última propiedad es que la descendencia (quizá no en primera generación) de cualquier distribución en Ω_1 va a contener a la larga individuos de todos los fenotipos. Esta hipótesis implica, por la teoria de matrices positivas si Ω_1 es finito, y más en general suponiendo además cierta compacidad del operador T_I^ε, por los teoremas tipo Perron-Frobenius en dimensión infinita ([50], [3]), que el radio espectral $R_0^\varepsilon(I)$ de T_I^ε es un valor propio dominante algebraicamente simple, con vector propio positivo (con norma 1) m_I^ε, y el único valor propio de T_I^ε que tiene un vector propio positivo. El valor propio $R_0^\varepsilon(I)$ recibe el nombre de ratio (o número) de reproducción básica y debe valer 1 en estado estacionario (obsérvese que mantenemos el subíndice 0 porque es la notación tradicional para este número, pero que no tiene ningún significado en el presente contexto). Así, un estado estacionario no trivial V^ε de S^ε queda caracterizado por una solución $(c, I) \in \mathbb{R}^+ \times Z$ del sistema

$$\begin{cases} R_0^\varepsilon(I) &= 1, \\ cLV_I^\varepsilon &= I, \end{cases} \tag{13}$$

donde V_I^ε está definido por la parte derecha de (11) con $m = m_I^\varepsilon$ y entonces $V^\varepsilon = cV_I^\varepsilon$.

Nótese que si el espacio de ambiente es unidimensional (por ejemplo si la única variable de interacción es la población total), entonces la primera ecuación de (13) determina I y como consecuencia, la segunda c. A menudo la ratio de reproducción básica R_0 es función monótona del ambiente (por ejemplo en la interacción únicamente por competencia, sería función decreciente de la población total), y entonces, el problema (13) tiene a lo sumo una solución.

En el caso de poblaciones estructuradas fisiológica y fenotípicamente, el "movimiento" en el espacio fisiológico Ω_0 es determinista, de forma que, si además el estado fisiológico en el momento del nacimiento es único (x_b), podemos llamar $x_{0,I}(a) \in \Omega_0$ al estado fisiológico de un individuo nacido en tiempo 0 al final del input $\rho(a)I$, es decir, cuando tiene edad a (cf. la sección 3). Por ejemplo, $x_{0,I}(a)$ sería el tamaño de los individuos de edad a sometidos a un input I si la estructura es con respecto al tamaño mientras que $x_{0,I}(a)$ es simplemente igual a a cuando la estructura es con respecto a la edad. Por otro lado, y como ya hemos mencionado, no hay "movimiento" en el espacio fenotípico a lo largo de la vida de los individuos. Por lo tanto, la probabilidad de que un individuo de fenotipo x_1 sobreviva hasta el final de un input I y tenga estado en ω, es decir, el núcleo $u_I((x_b, x_1), \omega)$, toma la forma de un producto de medidas de Dirac concentradas en $x_{0,I}(a)$ y en x_1, con un peso igual a la probabilidad de supervivencia, pongamos $\Pi_I(x_1, a)$. Además, siendo Z de

dimensión finita no es muy restrictivo suponer que el operador de "output" toma la forma siguiente:

$$LV = \int_{\Omega} \gamma(y_0, y_1) \mathrm{d}V(y_0, y_1), \tag{14}$$

donde γ es una función acotada definida en $\Omega = \Omega_0 \times \Omega_1$ y que toma valores en Z (véase [24]). Bajo estas condiciones, de (11) se obtiene, cambiando el orden de integración (y adviértase también la conveniencia de la notación $\mu(\mathrm{d}x)$ por $\mathrm{d}\mu(x)$ en la primera de las integrales), una forma más explícita para la segunda ecuación de (13):

$$\begin{aligned} LV_I^{\varepsilon} = \int_0^{\infty} \int_{\Omega_1} \int_{\Omega} \gamma(y_0, y_1) u_{\rho(a)I}((x_b, x_1), \mathrm{d}y_0 \mathrm{d}y_1) \mathrm{d}m_I^{\varepsilon}(x_1) \mathrm{d}a \quad = \\ \int_0^{\infty} \int_{\Omega_1} \int_{\Omega} \gamma(y_0, y_1) \Pi_I(x_1, a) \mathrm{d}\delta_{x_{0,I}(a)}(y_0) \mathrm{d}\delta_{x_1}(y_1) \mathrm{d}m_I^{\varepsilon}(x_1) \mathrm{d}a \quad = \\ \int_0^{\infty} \int_{\Omega_1} \gamma(x_{0,I}(a), x_1) \Pi_I(x_1, a) \mathrm{d}m_I^{\varepsilon}(x_1) \mathrm{d}a. \end{aligned}$$

Ahora consideramos la situación que se produce cuando la tasa de mutación o el tamaño medio de ésta tiende a 0. Más concretamente, vamos a suponer que la distribución B_x^{ε} de hijos de un individuo de fenotipo x tiende en algún sentido conveniente a la medida de Dirac concentrada en x cuando ε tiende a 0. Esto implicará, también en un sentido conveniente, que el operador T_I^{ε} (dado por (12)) tiende a un operador multiplicativo T_I^0 definido por

$$(T_I^0 m)(\omega_1) = \int_{\omega_1} r_I(x_1) \mathrm{d}m(x_1),$$

cuyo espectro es la clausura del conjunto imagen de la función positiva r_I y por lo tanto su cota espectral y su radio espectral, al que llamamos $R_0^0(I)$, son ambos iguales a $\sup_{\Omega_1} r_I$. Como consecuencia se tiene que el valor propio dominante $R_0^{\varepsilon}(I)$ tiende, para $\varepsilon \to 0$, a $\sup_{\Omega_1} r_I = R_0^0(I)$. Supongamos que este supremo se toma sólo en un punto x_I de Ω_1. Entonces el vector propio m_I^{ε} tiende (por lo menos en la topología débil estrella) a un vector propio de T_I^0 asociado a su radio espectral, por lo tanto, a una medida unitaria concentrada en x_I (véase [14] y [15] para este argumento en una situación muy similar). Como consecuencia, LV_I^{ε} tiende (formalmente) a

$$\int_0^{\infty} \int_{\Omega_1} \gamma(x_{0,I}(a), x_1) \Pi_I(x_1, a) \mathrm{d}\delta_{x_I}(x_1) \mathrm{d}a =$$

$$\int_0^{\infty} \gamma(x_{0,I}(a), x_I) \Pi_I(x_I, a) \mathrm{d}a = \int_0^{\infty} \int_{\Omega} \gamma(y_0, y_1) \Pi_I(x_I, a) \mathrm{d}\delta_{x_{0,I}(a)}(y_0) \mathrm{d}\delta_{x_I}(y_1) \mathrm{d}a.$$

Combinando las observaciones precedentes y usando (11) y (14), resulta que el sistema (13) tiende (formalmente) al sistema, también en $\mathrm{R}^+ \times Z$,

$$\begin{cases} (r_I(x_I) =) R_0^0(I) = 1 \\ cL(\int_0^\infty \Pi_I(x_I, a)\mathrm{d}\delta_{x_{0,I}(a)}\mathrm{d}\delta_{x_I}\mathrm{d}a) = I, \end{cases} \tag{15}$$

que es precisamente la condición de equilibrio del semigrupo S^0 restringido al subespacio invariante $X_{\{x_I\}}$ (véase la sección 5.1), es decir, un equilibrio del sistema puramente fisiológico para el valor x_I del fenotipo. Notemos ahora que si $(\hat{c}, \hat{I})$ es una solución de (15) entonces (y sólo entonces) $\hat{x} := x_{\hat{I}}$ es una ESS (véase la sección 5.2) puesto que la cota de crecimiento del semigrupo lineal $S^0_{I_{\hat{x}}}(t) = S^0_{\hat{I}}(t)$ restringido a $X_{\{y\}}$ es negativa siempre que y es distinto de $\hat{x}$ porque el radio espectral del operador de la generación siguiente (la ratio de reproducción básica) restringido a $X_{\{y\}}$ y con input $\hat{I}$, es decir, el número esperado de hijos $r_{\hat{I}}(y)$ de un individuo de fenotipo y sometido a un ambiente $\hat{I}$ es menor que 1 si $y \neq \hat{x} = x_{\hat{I}}$ (recuérdese que supusimos que x_I era el único punto de máximo de la función r_I y que por (15), $r_{\hat{I}}(x_{\hat{I}}) = 1$).

De lo dicho se deduce que los equilibrios del semigrupo S^ε tienden a concentrarse, para ε pequeño, alrededor de los valores ESS del fenotipo. Con más precisión, podemos afirmar que tienden a medidas en $X = M(\Omega_0 \times \Omega_1)$ de la forma $V^0 \times \delta_{\hat{x}}$, donde $V^0 \in M(\Omega_0)$ es un equilibrio de la dinámica fisiológica con parámetro $\hat{x} \in \Omega_1$ (por lo tanto, $V^0 \times \delta_{\hat{x}}$ es un equilibrio de S_0), y $\hat{x}$ es una estrategia evolutivamente estable.

Como se dijo al principio de la sección 5, en [14] y en [15] se considera un ejemplo de población estructurada fisiológicamente por la edad y fenotípicamente por la edad de maduración. Estos trabajos contienen resultados como los anteriormente expuestos aunque en la versión clásica de los modelos, es decir, mediante ecuaciones del tipo (7).

6 Conclusiones

La dinámica de poblaciones estructuradas es el marco adecuado para responder a preguntas sobre el crecimiento de las poblaciones en el caso muy frecuente en que el comportamiento de éstas en términos de tasas de fertilidad y mortalidad depende de estados individuales como la posición en el espacio, la edad, el tamaño, características fenotípicas, etc. El deseo de admitir como estados poblacionales distribuciones singulares como por ejemplo medidas concentradas que corresponden a poblaciones cuyos individuos tienen todos el mismo estado individual, a parte de dificultades técnicas ligadas a la construcción de sistemas dinámicos generados por las ecuaciones en derivadas parciales no locales que tradicionalmente modelizan la situación mencionada, ha llevado a varios autores ([21], [22], [23], [24], [26]) al desarrollo de un marco original para la dinámica de poblaciones estructuradas llamado formulación acumulativa en el que el modelo no se formula en términos de tasas instantáneas sino en términos de cantidades acumuladas en intervalos de tiempo finitos (y no infinitesimales) y referidas al comportamiento individual, como probabilidades de supervivencia y números esperados de descendientes directos. Así se construye el semigrupo

de la dinámica de poblaciones mediante el cómputo de cantidades de significado parecido pero extendidas a todas las generaciones.

La consideración de poblaciones estructuradas con respecto a características fenotípicas ofrece la posibilidad de explicar e incluso predecir los valores que la evolución biológica mediante mutación y selección natural determina para estas características. El punto de vista clásico en ese contexto procede de la teoría de juegos y el concepto principal es el de estrategia evolutivamente estable, que consiste en un fenotipo que, "adoptado" por una población *monomórfica*, la hace inmune a invasiones por poblaciones pequeñas con fenotipo distinto. Un punto de vista distinto, técnicamente más costoso pero también más satisfactorio en muchas situaciones de relevancia biológica consiste en la consideración directa de distribuciones de individuos con respecto al fenotipo, modelizando simultáneamente selección y mutación. En la parte final del trabajo se han mostrado las relaciones que pueden establecerse entre los equilibrios de esta dinámica de selección y mutación y las estrategias evolutivamente estables cuando la tasa de mutación o el tamaño medio de ésta es pequeño.

Ciertamente, queda mucho camino por recorrer en estos temas, empezando por la explotación del conocimiento de los sistemas puramente fisiológicos para probar existencia y unicidad de equilibrios de la dinámica de selección y mutación para mutación pequeña, y la comprensión de las relaciones entre la estabilidad asintótica de estos equilibrios (véase [17]) y las propiedades definidas por la dinámica adaptativa de los sistemas puramente fisiológicos subyacentes.

También conviene intentar la extensión del marco de la formulación acumulativa a sistemas con ambiente en espacios de dimensión infinita. Sobre la importancia biológica de la consideración de variables de interacción dependientes del estado individual (y por lo tanto tomando valores en espacios infinito dimensionales si el espacio de estados individuales no es finito) véase [38] y [39] en el caso de competencia por la luz en formaciones forestales y [29] y [30] en el caso de la modelización del canibalismo.

Otra extensión interesante pero que seguramente presenta grandes dificultades técnicas es la consideración de fenotipos que toman valores en espacios de dimensión infinita ("function valued phenotypical trait", en inglés). Sobre este tema en el contexto de estructura únicamente fisiológica y dinámica adaptativa, pueden verse los trabajos [12], [37], [47] y [48].

Agradecimientos. Este trabajo ha sido parcialmente financiado por el proyecto BFM 2002-04613.

Referencias

[1] A.S. Ackleh, D.F. Marshall, H.E. Heatherly, B.G. Fitzpatrick. Survival of the fittest in a generalized logistic model. *Math. Models Methods Appl. Sci*, 9, no. 9: 1379-1391, 1999.

[2] A.S. Ackleh, B.G. Fitzpatrick, H.R. Thieme. Rate distributions and survival of the fittest: a formulation on the space of measures. *Discrete Contin. Dyn. Syst. Ser. B*, 5, no. 4: 917-928, 2005.

[3] W. Arendt, A. Grabosch, G. Greiner, U. Groh, H.P. Lotz, U. Moustakas, R. Nagel, F. Neubrander, U. Schlotterbeck. *One-parameter semigroups of positive operators*. Lecture Notes in Mathematics, 1184. Springer-Verlag, Berlin, 1986.

[4] R. Bürger. Perturbations of positive semigroups and applications to population genetics. *Math. Z*, 197, no. 2: 259-272, 1988.

[5] R. Bürger, I.M. Bomze. Stationary distributions under mutation-selection balance: structure and properties. *Adv. in Appl. Probab*, 28, no. 1: 227-251, 1996.

[6] R. Bürger. *The mathematical theory of selection, recombination, and mutation.* Wiley Series in Mathematical and Computational Biology. John Wiley and Sons, Ltd., Chichester, 2000.

[7] À. Calsina, C. Perelló. Modelos matemáticos de la evolución darwiniana. *Actas de la Reunión Matemática en honor de A. Dou*, Universidad Complutense de Madrid 63-75, 1989.

[8] À. Calsina, C. Perelló. La matemática de la evolución biológica *Proceedings del XI Congreso de Ecuaciones diferenciales y aplicaciones / Primer Congreso de Matemática Aplicada. (Málaga, 1989),* 73-82, Univ. Málaga, Málaga, 1990.

[9] À. Calsina, C. Perelló, J. Saldaña. Non-local reaction diffusion equations modelling predator-prey coevolution. *Publ. Mat*, 38 no. 2: 315-325, 1994.

[10] À. Calsina, C. Perelló. Equations for biological evolution. *Proc. Roy. Soc. Edinburgh Sect. A*, 125, no. 5: 939-958, 1995.

[11] À. Calsina, J. Saldaña. A model of physiologically structured population dynamics with a nonlinear individual growth rate. *J. Math. Biol*, 33, no. 4: 335-364, 1995.

[12] À. Calsina, J. Saldaña. Global dynamics and evolutionarily stable life history of a size structured population. *SIAM Journal of Applied Mathematics*, Vol 59, No. 5: 1667- 1685, 1999.

[13] À. Calsina, S. Cuadrado. A model for the adaptive dynamics of the maturation age. *Ecological Modelling*, 133: 33-43, 2000.

[14] À. Calsina, S. Cuadrado. Small mutation rate and evolutionarily stable strategies in infinite dimensional adaptive dynamics. *J. Math. Biol*, 48: 135-159, 2004.

[15] À. Calsina, S. Cuadrado. Stationary solutions of a selection mutation model: the pure mutation case. *Math. Models Meth. Appl. Sci*, 15, no 7: 1091-1117, 2005.

[16] À. Calsina, C. Perelló, M. Sanchón, Modelling some aspects of the darwinian theory: internal competition. Prepublicacions Departament de Matemàtiques UAB, núm 31, 2005.

[17] À. Calsina, S. Cuadrado. Asymptotic stability of equilibria of selection-mutation equations. Prepublicacions Departament de Matemàtiques UAB, núm 02, 2006.

[18] J.A. Carrillo, S. Cuadrado, B. Perthame. Adaptive dynamics via Hamilton-Jacobi approach and entropy methods for a juvenile-adult model. Prepublicacions Departament de Matemàtiques UAB, núm 11, 2006.

[19] J.F. Crow, M. Kimura. The theory of genetic loads. *Proc XIth Int. Congr. Genetics* 495-505, 1964.

[20] J. M. Cushing. *An introduction to structured population dynamics*. CBMS-NSF Regional Conference Series in Applied Mathematics, 71. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[21] O. Diekmann, M. Gyllenberg, J.A.J. Metz, H.R. Thieme. On the formulation and analysis of general deterministic structured population models. I. Linear theory. *J. Math. Biol*, 36, no. 4: 349-388, 1998.

[22] O. Diekmann, M. Gyllenberg, H. Huang, M. Kirkilionis, J.A.J. Metz, H.R. Thieme. On the formulation and analysis of general deterministic structured population models. II. Nonlinear theory. *J. Math. Biol*, 43, no. 2: 157-189, 2001.

[23] O. Diekmann, M. Gyllenberg, and J.A.J. Metz. Steady state analysis of structured population models. *Theor. Pop. Biol.*, 63: 309-338, 2003.

[24] O. Diekmann, Ph. Getto. Boundedness, global existence and continuous dependence for nonlinear dynamical systems describing physiologically structured populations. *J. Differential Equations*, 215, no. 2: 268-319, 2005.

[25] O. Diekmann, P.-E. Jabin, S. Mischler, B. Perthame. The dynamics of adaptation: an illuminating example and a Hamilton Jacobi approach. *Theor. Pop. Biol.*, 67(4): 257-271, 2005.

[26] O. Diekmann, M. Gyllenberg, J.A.J. Metz. Physiologically structured population models: Towards a general mathematical theory. In Takeuchi, Sato, and Iwasa eds. *Mathematics for Ecology and Environmental Sciences*. Springer Verlag, en prensa.

[27] R.A. Fisher. *The genetical theory of natural selection*. Clarendon Press, 1930.

[28] S.A.H. Geritz, É. Kisdi, G. Meszéna, J.A.J. Metz. Evolutionary singular strategies and the adaptive growth and branching of the evolutionary tree, *Evol. Ecol.* **12** : 35-57, 1998.

[29] Ph. Getto, O. Diekmann, A.M. de Roos. On the (dis) advantages of cannibalism, *J. Math. Biol.*, 51, no. 6: 695-712, 2005.

[30] Getto, Ph. Steady state analysis for a size structured cannibalism model with two dynamic resources. In: *On some quasilinear structured population models*, Ph. D. thesis. Universidad de Utrecht, 2005.

[31] I. Gudelj, C.D. Coman, R.E. Beardmore. Classifying the role of trade-offs in the evolutionary diversity of pathogens. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci*, 462, no. 2065: 97-116, 2006.

[32] M.E. Gurtin, R.C. MacCamy. Non-linear age-dependent population dynamics. *Arch. Rational Mech. Anal.*, 54: 281-300, 1974.

[33] J. Hofbauer, K. Sigmund. *The theory of evolution and dynamical systems: mathematical aspects of selection*. Cambridge University Press, 1988.

[34] M. Iannelli. *Mathematical theory of age-structured population dynamics*, Applied Mathematics Monographs C.N.R. vol. 7, Giardini, Pisa, 1995.

[35] N. Kato. A general model of size-dependent population dynamics with nonlinear growth rate. *J. Math. Anal. Appl.*, 297 no. 1: 234-256, 2004.

[36] M. Kimura. A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proc. Natl. Acad. Sci. U.S.A* , 54: 731-736, 1965.

[37] J. G. Kingsolver, R. Gomulkiewicz, P. A. Carter. Variation, selection and evolution of function-valued traits. *Genetica*, 112-113, no. 1: 87-104, 2001.

[38] M. Kirkilionis, J. Saldaña. A height-structured forest model *Preprint 2001-03 (SFB 359), IWR, University of Heidelberg*, 1-36, 2001.

[39] E.A. Kraev. Existence and uniqueness for height structured hierarchical population models. *Natural resource modeling*, 14, no. 1: 45-70, 2001.

[40] L.R. Lawlor, J. Maynard-Smith, J. The coevolution and stability of competing species. *Amer. Natur*, 110: 79-99, 1976.

[41] P. Magal, G.F. Webb. Mutation, selection and recombination in a model of phenotype evolution. *Discrete Contin. Dynam. Systems*, 6, no. 1, 221-236, 2000.

[42] J. Maynard Smith, G.R. Price. The logic of animal conflict. *Nature*, 246: 15-18, 1973.

[43] J.A.J. Metz, O. Diekmann, O. *The dynamics of physiologically structured populations*. Lecture Notes in Biomath., 68, Springer, Berlin, 1986.

[44] J.A.J. Metz, S.A.H. Geritz, G. Mészena, F.J.A. Jacobs, J.S. van Heerwaarden. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. *Stochastic and spatial structures of dynamical systems (Amsterdam, 1995),* 183–231, Konink. Nederl. Akad. Wetensch. Verh. Afd. Natuurk. Eerste Reeks, 45, *North-Holland, Amsterdam,* 1996.

[45] M. A. Nowak, D. C. Krakauer, A. Klug, R. M. May. Prion Infection Dynamics. *Integrative Biology: Issues, News and Reviews*, 1 (1): 3-15, 1998.

[46] J.M. Palmada. Equacions de selecció i mutació en una població estructurada per l'edat. Trabajo de investigación. UAB, 2005.

[47] K. Parvinen, U. Dieckmann, M. Heino. Function-valued adaptive dynamics and the calculus of variations. J. Math. Biol. 52: 1-26, 2006.

[48] J. Ripoll. *Evolution of sex ratio in structured population dynamics*. Tesis doctoral. Universidad de Barcelona, 2005.

[49] J. Saldaña, S.F. Elena, R.V. Solé. Coinfection and superinfection in RNA virus populations: a selection-mutation model. *Math. Biosci*, 183, no. 2, 135-160, 2003.

[50] H.H. Schaefer. *Banach lattices and positive operators*. Die Grundlehren der mathematischen Wissenschaften, Band 215. Springer-Verlag, New York-Heidelberg, 1974.

[51] F.R. Sharpe, A.J. Lotka. A problem in age distribution, *Phil. Mag*, 21: 435-438, 1911.

[52] A. Turing. The chemical basis of morphogenesis. *Phil. Trans. Roy. Soc.* B 237: 37-72, 1952.

[53] G. Webb. *Theory of nonlinear age-dependent population dynamics*. Monographs and Textbooks in Pure and Applied Mathematics, 89. Marcel Dekker, Inc., New York, 1985.

Bol. Soc. Esp. Mat. Apl. n°36(2006), 125–149

Fourier y sus coeficientes

A. Cañada

Departamento de Análisis Matemático, Universidad de Granada

acanada@ugr.es

1 Introducción

Cuando se hace alguna consulta histórica sobre los llamados métodos de Fourier y su influencia en la historia de la matemática, un aspecto común suele ser el comentario que se refiere al procedimiento usado por Fourier en el cálculo de los coeficientes del desarrollo considerado. Es más o menos, así:

Para calcular los coeficientes, Fourier usó el desarrollo en serie de potencias de la función dada y de las funciones trigonométricas consideradas. Reordenó estos desarrollos con objeto de igualar los términos que multiplican a las respectivas potencias y llegó a un sistema lineal de infinitas ecuaciones con infinitas incógnitas. Entonces consideró un sistema lineal finito con m ecuaciones y m incógnitas (sistema formado con las primeras m filas y las primeras m columnas del sistema infinito original). Resolvió este sistema finito e hizo tender m a ∞. ***Después de un análisis largo y complicado****, alcanzó su célebre fórmula para los coeficientes.*

Jean Baptiste J.Fourier (1768-1830)

Cuando se me propuso realizar esta colaboración para el boletín de SëMA, pensé que podría tener interés transcribir a nuestros días algunos de los razonamientos originales de Fourier. En nuestro Departamento tenemos una copia de una edición que Dover llevó a cabo en 1955 ([16]) sobre la versión inglesa de 1878 del libro original de Fourier (publicado en francés en 1822). Me puse manos a la obra (y he de confesar que ya lo había intentado varias veces, pero no había perseverado lo suficiente). Lo pasé mal (porque muchos de los razonamientos que hace Fourier son realmente difíciles de entender para mí y porque la notación que usa es muy complicada para nosotros) y bien (cuando después de algunas horas de trabajo, conseguí entenderlos). El resultado final es muy positivo. Por una parte tienes ocasión de comparar lo que entendemos hoy

en día por rigor matemático con el rigor de la época de Fourier. Por otra, te das cuenta de que leer escritos originales de grandes matemáticos es muy formativo y al mismo tiempo placentero. No se daban por vencidos, desarrollaban unas tremendas dosis de ingenio para conseguir su objetivo y cuando, al final, haces balance de lo que has aprendido intentando entender lo que allí hay escrito, te das cuenta de que ésa es una parte fundamental de la auténtica matemática, una parte que aúna conceptos, resultados profundos, etc. con sus orígenes históricos. En las notas que aparecen a continuación, en la sección tercera intento trasladar a nuestros días algunas de las ideas de Fourier. En la mayoría de las ocasiones hay que prescindir del rigor en los razonamientos, tal y como lo entendemos hoy en día. Es también inevitable, aunque sea muy someramente, escribir algo sobre el origen de los métodos de Fourier (sección segunda) y la influencia que las ideas de Fourier han tenido en la historia de la matemática (sección cuarta).

2 El origen de las series de Fourier: las ecuaciones de ondas y del calor

Uno de los problemas más interesantes del que se ocuparon los científicos del siglo XVIII (y que posteriormente motivó el estudio de muchos otros similares) fue el problema de la cuerda vibrante. Si tomamos como referencia el estupendo texto de M. Kline ([25]), el primer matemático que elaboró un modelo apropiado para estudiar este problema fue Jean Le Rond d'Alembert en 1747 (para esta breve sección puede consultarse el texto citado para documentarse de manera muy precisa sobre fechas, revista científica donde se realizaron las publicaciones, volumen, páginas, etc. También son útiles [9] y [19]).
En su versión más sencilla, D'Alembert demostró que si la función $u(x,t)$ representa el desplazamiento vertical de la cuerda, en la coordenada x (suponemos $0 \leq x \leq \pi$ por simplicidad) y el tiempo t, entonces, si la posición inicial de la cuerda viene dada por una función f y la velocidad inicial de la misma es cero, la función u satisface un problema de tipo mixto de la forma

$$\begin{gathered} \frac{\partial^2 u(x,t)}{\partial t^2} = \frac{\partial^2 u(x,t)}{\partial x^2},\ 0 < x < \pi, t > 0 \\ u(x,0) = f(x),\ u_t(x,0) = 0,\ 0 \leq x \leq \pi \\ u(0,t) = u(\pi,t) = 0, t \geq 0 \end{gathered} \tag{1}$$

D'Alembert demostró además que la solución de (1) viene dada por

$$u(x,t) = \frac{1}{2}[\tilde{f}(x+t) + \tilde{f}(x-t)] \tag{2}$$

donde $\tilde{f}$ es la extensión a $\mathbb{R}$, impar y $2\pi-$ periódica de la función f.

La fórmula (2) fue también demostrada por Euler en 1749. Euler difería de D'Alembert en el tipo de funciones iniciales f que podían tenerse en cuenta. De hecho, estas diferencias pueden considerarse como una de las primeras manifestaciones escritas sobre los problemas que ha llevado consigo la definición

de la noción de "función", un concepto que hoy en día presumimos de tener muy claro. Mientras que para D'Alembert, f debería tener una fórmula concreta (una única expresión analítica), Euler defendía que no había ninguna razón física para no admitir como posiciones iniciales f a aquellas que, en diferentes partes de $[0, \pi]$, tuviesen expresiones distintas, siempre que al considerarlas unidas la posición inicial resultante tuviese una apropiada regularidad. Parece ser que tal discusión entre D'Alembert y Euler provenía del hecho de que en su tiempo se admitía que cada función daba lugar a una gráfica, pero no recíprocamente (cuando la gráfica considerada tenía diferentes expresiones en distintos intervalos). En resumen, Euler defendía que cualquier gráfica podía considerarse como curva inicial, tesis que no era compartida por D'Alembert. A este respecto puede consultarse la versión castellana de un interesante artículo de Luzin sobre el concepto de función ([28]).
Otra manera de obtener la solución del problema (1), completamente distinta (al menos a primera vista), fue propuesta por Daniel Bernoulli en 1753. La idea clave es obtener la solución de (1) como superposición de ondas más sencillas, concretamente aquellas que son de la forma

$$u_n(x,t) = sen(nx)\cos(nt),\ \forall\ n \in \mathbb{N}, \tag{3}$$

donde $\mathbb{N}$ es el conjunto de los números naturales. Para cada tiempo t fijo, la anterior función es un múltiplo de la función $sen(nx)$, que se anula exactamente en $n-1$ puntos del intervalo $(0, \pi)$. Así, si pudiésemos observar la vibración de la cuerda correspondiente a las ondas u_n, tendríamos $n-1$ puntos, llamados nodos, en los que la cuerda se mantendría constantemente fija en el eje de abscisas (como en los extremos del intervalo $[0, \pi]$). Entre dichos nodos, la cuerda oscilaría de acuerdo con (3).

¿Cómo concibió Bernoulli esta idea? Parece ser que una posibilidad es que usase sus conocimientos musicales. Para ello se basó en que el sonido que emite una cuerda vibrante es, en general, superposición de armónicos, es decir, superposición de funciones de la forma $u_n(x,t)$. Tales funciones representan, para $n = 1$ el tono fundamental y para $n > 1$ sus armónicos, y desde el punto de vista musical se corresponden con los tonos puros. Así, Bernoulli afirmó que cualquier sonido que produjese la vibración de la cuerda debe ser superposición de tonos puros. Desde el punto de vista matemático, ello significa que la solución de (1) puede representarse de la forma:

$$u(x,t) = \sum_{n=1}^{\infty} f_n sen(nx)\cos(nt), \tag{4}$$

donde los coeficientes f_n han de elegirse adecuadamente para que se satisfagan todas las relaciones de (1). Si la solución propuesta por *Bernoulli* fuese correcta, ello implicaría que

$$f(x) = \sum_{n=1}^{\infty} f_n sen(nx),\ \forall\ x \in [0, \pi], \tag{5}$$

para una adecuada elección de los coeficientes f_n. Este punto de vista expuesto por Bernoulli no tuvo aceptación en su tiempo. En particular, recibió duras contestaciones por parte de D'Alembert y Euler quienes no admitían que una función inicial f, más o menos arbitraria, pudiera representarse en la forma (5). Representativo de esto que decimos puede ser el artículo de D'Alembert titulado "*Fondamental*" contenido en el volumen séptimo de la famosa "*Encyclopédie*". La controversia se prolongó durante años.

Parece ser que las ideas de Bernoulli fueron fuente de inspiración para Jean Baptiste-Joseph Fourier, matemático y físico francés y profesor de análisis de la Escuela Politécnica. Fourier se interesó por la teoría de la conducción del calor en los cuerpos sólidos. En 1807 envió un artículo a la Academia de Ciencias de París (Mémoire sur la propagation de la chaleur), que trataba sobre dicho tema. En su versión más elemental (véase de nuevo [25]), Fourier se interesó por un problema de tipo mixto para la ecuación del calor de la forma

$$\begin{gathered}\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{\partial u(x,t)}{\partial t},\ 0 < x < \pi,\ 0 < t < T,\\ u(0,t) = u(\pi,t) = 0,\ 0 \le t \le T,\ u(x,0) = f(x),\ 0 \le x \le \pi.\end{gathered} \tag{6}$$

Como Bernoulli, Fourier buscó las soluciones más sencillas que puede presentar este problema usando el método de separación de variables y afirmó que la solución de (6) viene dada como superposición de ellas. Más precisamente, Fourier propuso como solución de (6) a la función u dada por la serie

$$u(x,t) = \sum_{n=1}^{\infty} f_n \exp(-n^2 t) sen(nx), \tag{7}$$

donde

$$f_n = \frac{2}{\pi}\int_0^{\pi} f(x) sen(nx)\ dx,\ \forall\ n \in \mathbb{N}. \tag{8}$$

Sin duda, el hecho de haber alcanzado la fórmula anterior para los coeficientes f_n es una de las contribuciones fundamentales de Fourier, y marca una diferencia significativa respecto del trabajo previo de Bernoulli sobre este tema.

El artículo de Fourier fue estudiado por los miembros de la Academia Francesa y, en términos generales, recibió serias críticas de los mismos, siendo su principal objeción la falta de rigor. No obstante, los científicos de tan prestigiosa institución estaban convencidos de la importancia que tenían los problemas relacionados con la propagación del calor y, los resultados teóricos presentados por Fourier tenían una gran concordancia con diversos experimentos llevados a cabo previamente. Por este motivo, convocaron un premio sobre el tema. Dicho premio fue otorgado a Fourier en 1812, pero a pesar de esto se continuó criticando su falta de rigor, de tal manera que aunque obtuvo el citado premio, Fourier no consiguió el propósito de publicar su trabajo en

la célebre serie "Mémoires"de la Academia Francesa. Fourier publicó por su cuenta su famoso libro *Théorie Analytique de la Chaleur*, en 1822 en París, donde incorporó parte de su artículo de 1812 prácticamente sin cambio. Este libro es actualmente una de las obras clásicas en matemáticas. Dos años más tarde consiguió el cargo de Secretario de la Academia Francesa y al fin pudo publicar el mencionado artículo en la serie "Mémoires".

Leyendo el libro original de Fourier, no es de extrañar la reacción de los miembros de la Academia Francesa. Me gustaría que el lector pensase sobre ello después de leer con detalle la siguiente sección, donde se presentan algunos de los razonamientos originales de Fourier.

CHAPITRE III. 211

en une suite infinie de sinus ou de cosinus d'arcs multiples. Cette question est liée à la théorie des équations aux différences partielles et a été agitée dès l'origine de cette analyse. Il était nécessaire de la résoudre pour intégrer convenablement les équations de la propagation de la chaleur; nous allons en exposer la solution.

On examinera, en premier lieu, le cas où il s'agit de réduire en une série de sinus d'arcs multiples, une fonction dont le développement ne contient que des puissances impaires de la variable. Désignant une telle fonction par φx, on posera l'équation

$$\varphi x = a \sin. x + b \sin. 2x + c \sin. 3x + d \sin. 4x + \ldots, \text{ etc.}$$

et il s'agit de déterminer la valeur des coëfficients a, b, c, d, etc. On écrira d'abord l'équation

$$\varphi x = x\varphi' 0 + \frac{x^2}{2}\varphi'' 0 + \frac{x^3}{2.3}\varphi''' 0 + \frac{x^4}{2.3.4}\varphi'' 0 + \frac{x^5}{2.3.4.5}\varphi' 0 + \ldots \text{ etc.}$$

dans laquelle $\varphi' 0, \varphi'' 0, \varphi''' 0, \varphi'' 0$, etc. désignent les valeurs que prennent les coëfficients

$$\frac{d.\varphi x}{dx}, \frac{d^2.\varphi x}{dx^2}, \frac{d^3.\varphi x}{dx^3}, \frac{d^4.\varphi x}{dx^4}, \text{ etc.}$$

lorsqu'on y suppose $x = 0$. Ainsi en représentant le développement selon les puissances de x par l'équation

$$\varphi x = Ax - B\frac{x^3}{2.3} + C\frac{x^5}{2.3.4.5} - D\frac{x^7}{2.3.4.5.6.7} + E\frac{x^9}{2.3.4.5.6.7.8.9} - \text{etc.}$$

3 Razonamientos e ideas de Fourier en el cálculo de los coeficientes

En esta sección intento transcribir algunas de las ideas originales de Fourier en la deducción de la fórmula (8). Más concretamente me refiero a las contenidas en los párrafos 207 a 223, de la sección VI de [16] (por favor olvídense del rigor tal y como lo entendemos hoy en día, pero disfruten con el ingenio y atrevimiento de Fourier).

3.1 Funciones desarrollables en series de potencias

Como hemos comentado en la sección anterior, Fourier se planteó, entre otros, desarrollos del tipo

$$f(x) = \sum_{n=1}^{\infty} f_n \operatorname{sen}(nx) \tag{9}$$

para funciones f que en principio supuso que eran impares y desarrollables en series de potencias y para valores de la variable x comprendidos entre 0 y π. Su objetivo era lograr una fórmula para el cálculo de los coeficientes $f_n,\ n \in \mathbb{N}$ que permitiese afirmar que (9) es verdad. Como f es impar, las derivadas de orden par de f en el origen son cero, es decir $f^{2k)}(0) = 0,\ \forall\ k \in \mathbb{N} \cup \{0\}$. Por tanto

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{2n+1)}(0)}{(2n+1)!}\ x^{2n+1} \tag{10}$$

En este punto hemos de decir que el desarrollo de Taylor de una función era conocido al menos desde 1715, cuando Taylor publicó un trabajo titulado "Methodus incrementorum directa et inversa", donde aparecía la conocida hoy en día como fórmula de Taylor.

Regresando a nuestro tema, teniendo en cuenta el desarrollo en serie de potencias de la función $\operatorname{sen} x$, tenemos

$$\operatorname{sen}(nx) = \sum_{k=0}^{\infty} (-1)^k \frac{(nx)^{2k+1}}{(2k+1)!}$$

Por tanto

$$\sum_{n=1}^{\infty} f_n \operatorname{sen}(nx) = \sum_{n=1}^{\infty} f_n \left(\sum_{k=0}^{\infty} (-1)^k \frac{(nx)^{2k+1}}{(2k+1)!} \right) =$$

$$\sum_{k=0}^{\infty} \left(\frac{(-1)^k}{(2k+1)!} x^{2k+1} (\sum_{n=1}^{\infty} f_n n^{2k+1}) \right)$$

Volviendo a las relaciones (9) y (10) e igualando los coeficientes de las respectivas potencias x^{2k+1} obtenemos

$$(-1)^k f^{2k+1)}(0) = \sum_{n=1}^{\infty} f_n\ n^{2k+1},\ \forall\ k \in \mathbb{N} \cup \{0\} \tag{11}$$

Lo anterior constituye un sistema de infinitas ecuaciones con infinitas incógnitas (los coeficientes f_n). Aquí Fourier "corta por lo sano" (perdón por la expresión, pero no se me ocurre otra mejor), considerando, para cada $m \in \mathbb{N}$ el siguiente sistema de m ecuaciones con m incógnitas (los elementos $g_n^m,\ 1 \le n \le m$)

$$(-1)^k f^{2k+1)}(0) = \sum_{n=1}^{m} g_n^m\ n^{2k+1},\ 0 \le k \le m-1 \tag{12}$$

Obsérvese que el anterior sistema y las incógnitas $g_n^m,\ 1 \leq n \leq m$ cambian con m (observación realizada por Fourier). Como primera afirmación conflictiva, Fourier dice que los coeficientes f_n que está buscando se obtienen como límite de los anteriores, es decir

$$f_n = \lim_{m \to +\infty} g_n^m,\ \forall\ n \in \mathbb{N} \tag{13}$$

(Parece ser que, entonces, el problema no ofrece ninguna dificultad. Sigan leyendo por favor)
Haciendo una pequeña pausa en nuestro objetivo, diré que existe algo de similitud entre la afirmación de Fourier (fórmula (13) y la idea usada por Taylor en la obtención de su famosa fórmula. En efecto, Taylor obtuvo su fórmula como "un caso límite" de la fórmula de interpolación de Newton ([10],[32]).

En adelante y para evitar complicaciones innecesarias nos concentraremos en el caso $n = 1$ que nos conducirá al primer coeficiente de Fourier f_1 (los razonamientos son muy similares para n general).
Para resolver el sistema lineal finito (12), Fourier usó un método elemental de eliminación de incógnitas: multiplicó la primera ecuación por un número conveniente y le restó la segunda con objeto de eliminar la última incógnita g_m^m. Hizo lo propio con la segunda ecuación y le restó la tercera, etc. Así obtuvo un sistema con $m-1$ ecuaciones y $m-1$ incógnitas. Iterando el procedimiento $m-1$ veces obtuvo g_1^m. Esta es la parte de "obrero" y les aseguro que no tiene nada de interés. Simplemente hay que tener un poco de tiempo y paciencia para hacer los cálculos. Se obtiene así

$$g_1^m = \frac{m^2(m-1)^2 \ldots 2^2}{(m^2-1)((m-1)^2-1))...(2^2-1)} H(m) \tag{14}$$

donde

$$H(m) = \sum_{k=0}^{m-1} f^{2k+1)}(0) \left[\sum_{2 \leq n_1 < n_2 < \ldots < n_k} \frac{1}{n_1^2 n_2^2 ... n_k^2} \right] \tag{15}$$

Para $k = 0$, el corchete anterior se entiende que es uno. Como

$$\lim_{m \to +\infty} \frac{m^2(m-1)^2 \ldots 2^2}{(m^2-1)((m-1)^2-1))...(2^2-1)} = \lim_{m \to +\infty} \frac{2 \cdot 2 \cdot 3 \cdot 3 \cdot 4 \cdot 4 \cdot 5 \cdot 5 \ldots}{1 \cdot 3 \cdot 2 \cdot 4 \cdot 3 \cdot 5 \cdot 4 \cdot 6 \ldots} = 2, \tag{16}$$

tendríamos una primera fórmula para f_1 que merece la pena ser destacada. En ella se obtiene el coeficiente f_1 en función de las derivadas (de orden impar) de la función f en el origen.

Fórmula número uno para el primer coeficiente

$$\frac{f_1}{2} = \sum_{k=0}^{\infty} f^{2k+1)}(0) \left[\sum_{2 \leq n_1 < n_2 < \ldots < n_k} \frac{1}{n_1^2 n_2^2 \ldots n_k^2} \right] \tag{17}$$

(Los que conozcan algo de series de Fourier, deben estar preguntándose qué tiene que ver esto con la expresión usual de f_1 dada por (8). Un poco de paciencia, por favor).

Nota 1 *Observemos que en la expresión anterior, los coeficientes de* $f^{2k+1)}(0)$ *se forman con elementos del conjunto* $\{\frac{1}{(m+1)^2},\ m \in \mathbb{N}\}$ *siguiendo una ley muy clara. En el cálculo del segundo coeficiente* f_2 *la ley es la misma usando el conjunto* $\{\frac{1}{(m+1)^2},\ m \in (\mathbb{N} \cup \{0\}) \setminus \{1\}\}$. *En general, el* $n-$*ésimo coeficiente de Fourier* f_n *se obtendría de manera análoga usando el conjunto* $\{\frac{1}{(m+1)^2},\ m \in (\mathbb{N} \cup \{0\}) \setminus \{n-1\}\}$.

A continuación Fourier calculó de manera explícita las sumas de las series que aparecen como coeficientes de $f^{2k+1)}(0)$ en la expresión (17).

Esto está íntimamente conectado con el llamado problema de Basilea, sobre la suma

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

resuelto brillantemente por Euler y que con anterioridad había sido objetivo (sin éxito) de otros grandes matemáticos (véase [14] para los detalles).

Las ideas principales se describen a continuación.
Como hemos mencionado con anterioridad, era conocido que

Leonhard Euler(1707-1783)

$$\operatorname{sen} x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!},\ \forall\ x \in \mathbb{R}$$

de donde se obtiene, para cualquier x que no sea cero,

$$\frac{\operatorname{sen} x}{x} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k+1)!}$$

La parte derecha de la expresión anterior era para Euler un "polinomio de grado infinito", cuyos ceros son los mismos que los de la función $\operatorname{sen} x$, salvo $x = 0$, es

decir, el conjunto de números reales $\{k\pi,\ k \in \mathbb{Z}\backslash\{0\}\}$. Por tanto, este polinomio infinito se puede factorizar de la forma siguiente

$$\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k}}{(2k+1)!} = \prod_{k=1}^{+\infty}\left(1-\frac{x}{k\pi}\right)\left(1-\frac{x}{-k\pi}\right) = \prod_{k=1}^{\infty}\left(1-\frac{x^2}{k^2\pi^2}\right) \tag{18}$$

Hay veces en las que merece la pena poner las fórmulas de manera desarrollada para apreciar su belleza, evitando sumas y productos infinitos. Por ejemplo, la fórmula anterior queda de la forma siguiente:

$$1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots = \left(1-\frac{x^2}{1^2\pi^2}\right)\left(1-\frac{x^2}{2^2\pi^2}\right)\left(1-\frac{x^2}{3^2\pi^2}\right)\ldots$$

Igualando el coeficiente de la potencia x^2, Euler obtuvo en 1735

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$$

(Tengo que reconocer que la primera vez que vi esta demostración en [14] me quedé maravillado de su sencillez y belleza).
Euler estudió además el valor de la serie

$$\sum_{n=1}^{\infty}\frac{1}{n^{2k}} \tag{19}$$

para diferentes valores de k. Por ejemplo, probó que

$$\sum_{n=1}^{\infty}\frac{1}{n^4} = \frac{\pi^4}{90},\ \sum_{n=1}^{\infty}\frac{1}{n^6} = \frac{\pi^6}{945},\ldots,\ \sum_{n=1}^{\infty}\frac{1}{n^{26}} = \frac{2^{24}76977927\pi^{26}}{27!}$$

Además encontró una relación clara (para $k \in \mathbb{N}$ arbitrario) entre la suma anterior y los llamados números de Bernoulli, que no son sino los coeficientes del desarrollo

$$\frac{x}{\exp x - 1} = \sum_{n=0}^{\infty}\frac{B_n x^n}{n!}$$

De hecho, para cualquier natural k se tiene

$$\sum_{n=1}^{\infty}\frac{1}{n^{2k}} = (-1)^{k-1}((2\pi)^{2k}/2(2k)!)B_{2k}$$

donde B_{2k} denota el número de Bernoulli de orden $2k$. Como estos números son racionales, se deduce inmediatamente que la suma (19) es irracional para cualquier valor de $k \in \mathbb{N}$.
El caso de la suma de las series del tipo

$$\sum_{n=1}^{\infty}\frac{1}{n^{2k+1}} \tag{20}$$

con k un número natural sigue siendo un misterio (uno más, de los muchos que rodean a la función zeta de Riemann). En 1978, Roger Apéry ([5]; véanse también [12] y [30]) demostró que cuando $k = 1$, la suma es un número irracional, pero aún no se tiene un resultado similar para valores mayores de k, aunque también se sabe que el conjunto de los valores de k para los que (20) es irracional, es un conjunto infinito ([15]).

Volvamos a nuestro tema. Fourier igualó los coeficientes de las respectivas potencias de x en la expresión (18) obteniendo

$$\sum_{1\leq n_1<n_2<\ldots<n_k} \frac{1}{n_1^2 n_2^2 \ldots n_k^2} = \frac{\pi^{2k}}{(2k+1)!},\ \forall\ k \in \mathbb{N} \tag{21}$$

Si llamamos R_k al coeficiente de $f^{2k+1)}(0)$ en (17) y S_k a la suma anterior, entonces se tiene

$$R_k = S_k - R_{k-1},\ \forall\ k \in \mathbb{N}.$$

Como, por definición $R_0 = 1$, se obtiene fácilmente que

$$R_k = (-1)^k \sum_{n=0}^{k} (-1)^n \frac{\pi^{2n}}{(2n+1)!},\ \forall\ k \in \mathbb{N}.$$

Combinando esto con (17) obtenemos un segundo resultado sobre el coeficiente f_1, que merece la pena destacarse (puesto que ya aparece el número π).

Fórmula número dos para el primer coeficiente

$$\frac{f_1}{2} = \sum_{k=0}^{\infty} (-1)^k f^{2k+1)}(0) \left[\sum_{n=0}^{k} (-1)^n \frac{\pi^{2n}}{(2n+1)!} \right] \tag{22}$$

Nota 2 *Fourier no usaba las notaciones anteriores, con sumas infinitas, productos infinitos, etc. al menos en la parte que yo he estudiado. Obtenía varios casos particulares y después escribía algo parecido a esto: "es fácil darse cuenta de la ley general que siguen estas relaciones". Podemos hacernos una buena idea de lo que quiero decir, si leemos con detenimiento el razonamiento que expongo a continuación de la nota 3.*

Nota 3 *Una vez obtenida por Fourier una fórmula similar a (22) para coeficientes arbitrarios f_n, Fourier incluyó en su libro numerosos casos concretos. Así concluyó, por ejemplo, que en el intervalo $(-\pi, \pi)$ se tiene*

$$\frac{x}{2} = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\operatorname{sen}(nx)}{n}$$

y comentó que con anterioridad este resultado había sido obtenido por Euler. También consideró Fourier el caso más general dado por $f(x) = x^{2p+1}$, para algunos valores de p.

En un golpe más de audacia (mis conocimientos no me permiten calificar esto de otra forma), Fourier dio otra expresión para f_1 que fue clave para la deducción de la fórmula definitiva. Es como sigue.
Partiendo de (22) se tiene

$$\frac{f_1}{2} = \sum_{k=0}^{\infty}(-1)^k f^{2k+1)}(0)\left[\sum_{n=0}^{k}(-1)^n\frac{\pi^{2n}}{(2n+1)!}\right] =$$

$$f'(0) - f'''(0)\left[1-\frac{\pi^2}{3!}\right] + f^{v)}(0)\left[1-\frac{\pi^2}{3!}+\frac{\pi^4}{5!}\right]$$

$$-f^{vii)}(0)\left[1-\frac{\pi^2}{3!}+\frac{\pi^4}{5!}-\frac{\pi^6}{7!}\right]+... =$$

$$\left[f'(0)+f'''(0)\frac{\pi^2}{3!}+f^{v)}(0)\frac{\pi^4}{5!}+...\right]-$$

$$\left[f'''(0)+f^{v)}(0)\frac{\pi^2}{3!}+f^{vii)}(0)\frac{\pi^4}{5!}+...\right]...$$

Ahora dice Fourier que usando nuevamente el desarrollo de Taylor, el primer corchete de la expresión anterior es claramente (¿?) el desarrollo en serie de potencias en torno al origen de $\dfrac{f(\pi)}{\pi}$, el segundo corchete es el desarrollo en serie de potencias en torno al origen de $\dfrac{f''(\pi)}{\pi}$ y así sucesivamente con lo que obtiene

$$\frac{f_1}{2} = \frac{1}{\pi}\sum_{n=0}^{\infty}(-1)^n f^{2n)}(\pi) \tag{23}$$

o bien, para los que están un poco impacientes,

Fórmula número tres para el primer coeficiente

$$f_1 = \frac{2}{\pi}\sum_{n=0}^{\infty}(-1)^n f^{2n)}(\pi) \tag{24}$$

(al menos ya aparece en la fórmula de f_1 el número $\frac{2}{\pi}$).

3.2 ¿Funciones arbitrarias?

Como hemos tenido oportunidad de apreciar, en los razonamientos de la sección anterior hay de todo: una parte que puede calificarse de obrera, en el sentido de que es cálculo y más cálculo sin ideas significativas (la resolución del sistema lineal finito-dimensional). Otras deducciones son ingeniosas (como el paso de la fórmula número uno a la fórmula número dos en la obtención del primer

coeficiente) y otras son simplemente audaces (por ejemplo, el paso de la fórmula número dos a la fórmula número tres). No obstante, aquellos que tienen alguna familiaridad con series de Fourier estarán todavía preguntándose qué tienen que ver las expresiones (17), (22) y (24) con la fórmula bien conocida para f_1. Tengan un poco de paciencia porque ahora viene lo mejor.

Alcanzada la fórmula (24) Fourier hace una (¿otra?) afirmación sorprendente. Dice más o menos lo siguiente: Hasta ahora hemos supuesto que la función f puede desarrollarse en serie de potencias de la variable x. No obstante, podemos hacer que el resultado previo sea válido para funciones cualesquiera enteramente arbitrarias, incluso discontinuas. Para establecer la veracidad de esta afirmación es preciso que examinemos con detalle la naturaleza de los coeficientes de $\operatorname{sen} x$, $\operatorname{sen}(2x)$, ... *en el desarrollo (9).*

232 THÉORIE DE LA CHALEUR.

qui seraient discontinues et entièrement arbitraires. Pour établir clairement la vérité de cette proposition, il est nécessaire de poursuivre l'analyse qui fournit l'équation précédente (B) et d'examiner quelle est la nature des coëfficients qui multiplient sin. x, sin. $2x$, sin. $3x$, sin. $4x$. En désignant par $\frac{s}{n}$ la quantité qui multiplie dans cette équation $\frac{1}{n}$ sin. nx, si n est impair, et $-\frac{1}{n}$ sin. nx, si n est pair; on aura

$$s = \varphi\pi - \frac{1}{n^2}\varphi''\pi + \frac{1}{n^4}\varphi^{IV}\pi - \frac{1}{n^6}\varphi^{VI}\pi + \text{etc.}$$

Considérant s comme une fonction de π, différentiant deux fois, et comparant les résultats, on trouve $s + \frac{1}{n^2}\frac{d^2 s}{d\pi^2} = \varphi\pi$; équation à laquelle la valeur précédente de s doit satisfaire. Or, l'équation $s + \frac{1}{n^2}\frac{d^2 s}{dx^2} = \varphi x$, dans laquelle s est considérée comme une fonction de x, a pour intégrale

$$s = a\cos. nx + b\sin. nx + n\sin. nx \int \cos. nx.\varphi x.dx - n\cos. nx \int \sin. nx.\varphi x\, dx.$$

n étant un nombre entier, et la valeur de x étant égale à π, on a $s = \pm n\int \varphi x.\sin. nx\, dx$. Le signe + doit être choisi lorsque n est impair, et le signe — lorsque ce nombre est pair. On doit supposer x égal à la demi-circonférence π, après l'intégration indiquée; ce résultat se vérifie, lorqu'on développe au moyen de l'intégration par parties, le terme

$$\int \varphi x \sin. nx.dx$$

Cuidado con la afirmación anterior. Parece claro que una "función cualesquiera enteramente arbitraria" no era para Fourier lo que entendemos hoy en día por ello. Por ejemplo, Fourier consideraba a una función del tipo

$$f(x) = \begin{cases} \exp(-x), & x < 0, \\ \exp(x), & x \geq 0 \end{cases}$$

como discontinua. En este sentido conviene tener en cuenta que fue Cauchy quien definió rigurosamente en su Cours d'analyse (1823,1829) los conceptos de función, límite, continuidad, derivada e integral tal y como los entendemos hoy en día, mientras que la primera versión del tratado de Fourier se presentó a la Academia Francesa en 1807, aunque la publicación de su famoso libro se retrasó (por causas bien conocidas para los aficionados a la historia) hasta 1822.

Para tratar de entender el razonamiento de Fourier nos concentramos de nuevo en la expresión (24) que define el coeficiente f_1. Fourier consideró que f_1 era una función de π y si denotamos por $s(\pi)$ a la función $s(\pi) = \sum_{n=0}^{\infty}(-1)^n f^{2n)}(\pi)$ entonces, derivando dos veces respecto de π se obtiene

$$s''(\pi) = \sum_{n=0}^{\infty}(-1)^n f^{2n+2)}(\pi) \tag{25}$$

con lo que

$$s''(\pi) + s(\pi) = f(\pi) \tag{26}$$

Fourier razona a continuación sobre la ecuación diferencial

$$s''(x) + s(x) = f(x) \tag{27}$$

en la que s se considera una función de x. De hecho, $s(x) = \sum_{n=0}^{\infty}(-1)^n f^{2n)}(x)$. Era conocido (el razonamiento de Fourier es un poco más complicado que el que expongo aquí) que $s(x)$ debe ser de la forma

$$s(x) = a\cos x + b\,\mathrm{sen}\, x + \mathrm{sen}\, x \int_0^x f(s)\cos s\ ds - \cos x \int_0^x f(s)\,\mathrm{sen}\, s\ ds \tag{28}$$

para ciertas constantes a y b. Por tanto

$$s(\pi) = -a + \int_0^{\pi} f(x)\,\mathrm{sen}\, x\ \ dx$$

Como además $s(0) = 0$, la constante a debe ser cero, alcanzándose la fórmula

$$s(\pi) = \int_0^{\pi} f(x)\,\mathrm{sen}\, x\ dx$$

con lo que

$$f_1 = \frac{2}{\pi}\int_0^{\pi} f(x)\,\mathrm{sen}\, x\ dx \tag{29}$$

Como ya hemos comentado, los razonamientos de Fourier son similares para obtener el coeficientes f_n, obteniendo

$$f_n = \frac{2}{\pi}\int_0^{\pi} f(x)\,\mathrm{sen}(nx)\ dx,\ \forall\ n \in \mathbb{N}. \tag{30}$$

Una vez que Fourier acabó los razonamientos que le permitieron alcanzar la fórmula (30), hace también una afirmación muy curiosa. Dice que el resultado obtenido se puede comprobar (¿?) de manera muy simple: basta multiplicar la expresión (9) por la función $\mathrm{sen}(nx)$ e integrar de 0 a π.
El mismo Fourier dice en su libro que la fórmula (30) es un resultado destacable, puesto que la función considerada puede ser enteramente arbitraria, siempre que (30) se pueda calcular. Precisamente el intentar dar sentido a los llamados

coeficientes de Fourier ha motivado de manera significativa los diferentes conceptos de integral (véase [25] para fechas históricas concretas).

En efecto, para el caso en que f es una función continua, Cauchy introdujo lo que hoy en día se conoce con el nombre de sumas de Riemann, es decir sumas de la forma

$$\sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) \tag{31}$$

donde $x_0 = 0 < x_1 < ... < x_n = \pi$ es cualquier partición del intervalo $[0, \pi]$ y $x_{i-1} \leq \xi_i \leq x_i$, $1 \leq i \leq n$. Aunque de manera no totalmente rigurosa (pues no expuso explícitamente el concepto de continuidad uniforme), Cauchy demostró que si f es continua en $[0, \pi]$ y las longitudes de todos los subintervalos de la partición considerada tienden a cero, entonces las anteriores sumas convergen a un límite llamado la integral de la función.


Henri Léon Lebesgue (1875-1941)

Riemann también se interesó por el tema afirmando que era importante, al menos para los matemáticos aunque no necesariamente para las aplicaciones físicas, establecer las condiciones más amplias posibles bajos las cuales tienen sentido las fórmulas de los coeficientes de Fourier. Introdujo así lo que llamamos hoy en día integral de Riemann, cuya idea básica es por una parte no asumir necesariamente que f es continua, y por otra establecer condiciones lo más generales posibles para que las sumas (31) tengan un único límite cuando las longitudes de todos los subintervalos de la partición considerada tienden a cero. Esto le permitió integrar funciones con un número infinito de discontinuidades. No obstante, hubo que esperar a los trabajos de Lebesgue sobre la medida de un conjunto, para tener una caracterización precisa de las funciones que pueden integrarse según Riemann. De hecho, la que se considera actualmente como integral definitiva en muchos aspectos, es la introducida por Lebesgue en 1902 en su tesis doctoral: "Intégrale, longueur, aire". El punto de partida, respecto de la noción de integral de Cauchy o de Riemann es completamente diferente, pues lo que se intentaba era medir, de alguna forma, el conjunto de puntos de discontinuidad de una función dada (véase [4]). La noción de integral de Lebesgue permitió probar con gran generalidad muchas conclusiones sobre series de Fourier que, con anterioridad a Lebesgue, eran conocidas para tipos particulares de funciones (lema de Riemann-Lebesgue, igualdad de Parseval, criterios de convergencia puntual, etc.). Además, muchos resultados de la teoría de integración de Lebesgue se expresan con una gran simplicidad y claridad respecto de las teorías de integración anteriores (teoremas de convergencia, teorema

de Fubini, etc.), de tal forma que el conocimiento de la teoría de la integral de Lebesgue es, hoy en día, imprescindible, para poder entender y presentar adecuadamente la teoría de series de Fourier.

4 Algunos temas relacionados con series de Fourier

Las ideas expuestas por Fourier en su libro plantearon de manera inmediata innumerables interrogantes y han originado, a lo largo de dos siglos, gran cantidad de investigación. Han sido muchas las partes de la matemática que se han desarrollado a partir de ellas. Comentamos algunas a continuación.

4.1 Funciones continuas no derivables

Las nociones de continuidad y diferenciabilidad de una función real de variable real están hoy en día perfectamente establecidas. Sin embargo, históricamente no ha ocurrido así. De hecho, el primitivo concepto de derivada debido a Newton y Leibnitz era bastante más complicado de expresar del que conocemos en la actualidad. Fue Cauchy ([25]) quien, unificando las notaciones de Newton y Leibnitz, y basado en una definición anterior de Bolzano de 1817, introdujo en 1823 la definición que hoy en día se da en todos los libros de texto

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{32}$$

Durante bastante tiempo se estuvo convencido de que cualquier función continua debía ser derivable, excepto posiblemente en conjuntos "aislados"de puntos. Pero, insistamos, ¿en cuántos puntos puede una función continua no ser derivable? La respuesta a esta pregunta estuvo relacionada desde el principio con la siguiente cuestión sobre series de Fourier: ¿en cuántos puntos puede no converger la serie de Fourier de una función continua dada? De hecho, después de la publicación, en 1822, del libro de Fourier, Dirichlet se ocupó durante varios años del problema de la convergencia de las series de Fourier, dando por primera vez de forma rigurosa un conjunto de hipótesis para garantizar la convergencia de las mismas. Este conjunto de hipótesis incluía la continuidad. Durante aproximadamente los cincuenta años siguientes, se pensó que la continuidad de la función debería ser suficiente para la convergencia de su serie de Fourier. Sin embargo, algunos matemáticos sospechaban que ello no debía ser así y todo esto motivó el estudio de funciones "raras" en el sentido de que tales funciones fuesen continuas, pero no derivables "en el máximo número de puntos posibles". Riemann definió en 1868 una función f, integrable en cualquier intervalo real finito, pero que tiene un conjunto infinito de discontinuidades en cualquier intervalo real no trivial. Además, para esta función f definida por Riemann, una integral indefinida cualquiera es continua en cualquier punto de $\mathbb{R}$, y sin embargo no es derivable en ningún punto de discontinuidad de f.
Posteriormente, Weierstrass, estudiando el tipo de funciones que podían representarse o desarrollarse en serie de Fourier, presentó en 1872 un ejemplo sorprendente a la Real Academia de Ciencias de Berlín: una función real,

de variable real, continua en cualquier punto y no derivable en ninguno. Concretamente, el ejemplo de Weierstrass está dado por la función

$$f(x) = \sum_{n=0}^{\infty} b^n \cos(a^n \pi x) \tag{33}$$

donde $0 < b < 1$ y a es cualquier entero impar tal que $ab > 1 + (3\pi/2)$. El resultado de Weierstrass fue generalizado por diferentes matemáticos, destacando el resultado de Hardy ([20]) que demostró que se tiene la misma conclusión suponiendo hipótesis más generales: $0 < b < 1$ y $ab \geq 1$ (véase también [6]). Posteriormente se han dado numerosos ejemplos de funciones continuas no derivables. Algunas de las más sencillas pueden verse en [3], [27] y [29].

Puede pensarse que el tipo de funciones anteriores es excepcional. Nada más lejos de la realidad. El análisis funcional, la disciplina matemática por excelencia del siglo XX, permite probar que las anteriores situaciones son las que "usualmente cabe esperar". ¿Cómo es esto? La herramienta clave para entenderlo es lo que se conoce con el nombre de Teorema de la Categoría de Baire (Baire, 1899) que comentamos a continuación. Sea X un espacio de Banach real cualquiera. Si $M \subset X$, diremos que M es de primera categoría en X, si M es alguna unión numerable de subconjuntos M_n de X tales que cada M_n verifica la propiedad $int\ \overline{M_n} = \emptyset$, donde $int\ \overline{M_n}$ denota el interior de la clausura de M_n y $\emptyset$ indica el conjunto vacío. Un subconjunto M de X se dice de segunda categoría en X, si M no es de primera categoría en X. El teorema de la categoría de Baire afirma que X es de segunda categoría en sí mismo.
Consideremos ahora $X = C([a,b], \mathbb{R})$, es decir, el espacio de las funciones reales y continuas, definidas en un intervalo dado $[a,b]$ de $\mathbb{R}$, con la norma uniforme. Sea

$$M = \{f \in X : \exists\ x \in [a,b) :\ \text{existe}\ f'(x+)\}$$

Pues bien, Banach y Mazurkiewicz probaron en 1931 que el conjunto M es de primera categoría en X y por tanto $X \setminus M$ es de segunda categoría en X. Este resultado es de gran belleza. No obstante hemos de ser precavidos si pensamos que puede haber alguna relación entre la noción de categoría y la noción intuitiva de tamaño o medida de un conjunto. De hecho, usando los conjuntos ternarios de Cantor ([4]) no es difícil dar ejemplos de subconjuntos de $\mathbb{R}$ que son de primera categoría en $\mathbb{R}$ y con medida (de Lebesgue) positiva. Asimismo existen subconjuntos de $\mathbb{R}$ de segunda categoría en $\mathbb{R}$ y con medida cero.

En lo que respecta a las series de Fourier de funciones continuas, Du Bois-Reymond dió en 1873 un ejemplo de una función continua cuya serie de Fourier no convergía en un conjunto denso de puntos.
Llegados aquí, la pregunta puede ser: ¿en cuántos puntos puede no converger la serie de Fourier de una función continua? Hubo que esperar hasta 1966, año en que Carleson demostró que se da la convergencia salvo posiblemente en un

conjunto de medida cero ([11]). Este resultado puede considerarse como uno de los más destacados de la matemática del siglo XX. La demostración de Carleson es realmente complicada y la referencia [1] puede ser de gran ayuda para aquellos que tengan interés en entenderla. Por cierto que el Premio Abel 2006 ha sido concedido a Carleson, entre otras cosas por sus importantes contribuciones al análisis armónico.
En 1966 también, Kahane y Katznelson probaron que dado cualquier conjunto A de medida nula existe una función continua cuya serie de Fourier diverge en cada punto de A ([23]). Estas son las cosas bonitas de la matemática.

4.2 Unicidad de la representación de una función en serie trigonométrica

Pasando a otro tema, la teoría de conjuntos de Cantor, base y fundamento de lo que se conoce con el nombre de matemática moderna, estuvo en buena parte motivada por el estudio de los puntos de convergencia o divergencia de las series trigonométricas. Fue este problema lo que llevó a Cantor a definir algunas de las primeras nociones de topología conjuntista, como las de conjunto cerrado y punto de acumulación. En efecto, cuando Cantor comenzó a trabajar en la Universidad de Halle, Heine estaba interesado en aquella época en la cuestión de la unicidad de la representación de una función dada en serie trigonométrica. Una serie trigonométrica es una serie de funciones de la forma

$$a_0/2 + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \operatorname{sen}(nx)) \tag{34}$$

donde $a_n, b_n \in \mathbb{R}$. Una función $f : \mathbb{R} \to \mathbb{R}$ se dice que admite un desarrollo en serie trigonométrica si existe alguna serie trigonométrica como (34) tal que

$$f(x) = a_0/2 + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \operatorname{sen}(nx)), \ \forall\ x \in \mathbb{R}. \tag{35}$$

Por ejemplo, sabemos que esto es así si f es $2\pi-$ periódica y de clase C^1. En este caso, los coeficientes a_n y b_n son los coeficientes de Fourier definidos como

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(nt)\ dt,\ b_n = \frac{1}{\pi} \int_0^{2\pi} f(t) \operatorname{sen}(nt)\ dt$$

El problema que Heine planteó en 1869 a Cantor (con 24 años de edad) fue: ¿es el desarrollo en serie trigonométrica único? Es decir, si

$$f(x) = a_0/2 + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \operatorname{sen}(nx)) =$$
$$= a_0'/2 + \sum_{n=1}^{\infty} (a_n' \cos(nx) + b_n' \operatorname{sen}(nx)),\ \forall\ x \in \mathbb{R},$$

¿es verdad que $a_0 = a_0'$, $a_n = a_n'$, $b_n = b_n'$, $\forall\ n \in \mathbb{N}$? Este problema no era fácil y antes habían intentado resolverlo, sin éxito, el mismo Heine, Dirichlet, Lipschitz y Riemann, entre otros. Es claro que el problema es equivalente al siguiente: si

$$0 = a_0/2 + \sum_{n=1}^{\infty}(a_n \cos(nx) + b_n \operatorname{sen}(nx)),\ \forall\ x \in \mathbb{R}, \tag{36}$$

¿es verdad que $a_0 = 0$, $a_n = 0$, $b_n = 0$, $\forall\ n \in \mathbb{N}$?

Cantor probó en 1870 que ello era así y que incluso, se puede renunciar a la convergencia de la serie (36) en un conjunto finito de puntos. La pregunta que Cantor se hizo a continuación era obvia: ¿en cuántos puntos podemos renunciar a la convergencia de la serie (36) y sin embargo seguir teniendo el mismo resultado de unicidad? Según mis conocimientos, este problema sigue sin resolverse hoy en día en toda su generalidad, a pesar de que se han realizado numerosos estudios sobre ello, comenzando por varios de Cantor, el primero fechado en 1871.

Georg Cantor(1845-1918)

En este trabajo Cantor demostró que el conjunto de puntos excepcionales, es decir, aquellos donde no se tiene necesariamente convergencia de la serie trigonométrica (36), puede estar formado por infinitos elementos, siempre que tal conjunto sea "de orden finito". ¿Cómo definió Cantor el orden de un conjunto? De la siguiente manera: dado cualquier subconjunto E de números reales, Cantor introdujo el concepto de punto de acumulación de E, tal y como se entiende hoy en día. Al conjunto de todos los puntos de acumulación, conjunto derivado de E, lo notó por E'. Análogamente puede definirse el segundo conjunto derivado de E, E'', como el conjunto derivado de E', y así sucesivamente. Claramente se tienen las inclusiones $...E''' \subset E'' \subset E'$. Entonces, un conjunto es de orden finito si algún derivado suyo es finito. También Cantor definió los conjuntos cerrados como aquellos que contienen a su derivado. Demostró además que un conjunto de orden finito es o finito o puede ponerse en correspondencia biyectiva con el conjunto $\mathbb{N}$ de los números naturales. A estos últimos conjuntos les dió el nombre de infinitos numerables, interesándose a continuación por la existencia de subconjuntos de números reales infinitos no numerables, para seguir con el estudio de subconjuntos del espacio $\mathbb{R}^n$. Por cierto, que posteriormente fue demostrado que cualquier conjunto numerable es válido también como conjunto de puntos excepcionales donde puede fallar la convergencia de la serie trigonométrica (36) y seguir teniendo la representación única (Bernstein en 1908 y Young en 1909). También se han

dado ejemplos de conjuntos excepcionales no numerables. Realmente, este es uno de los problemas abiertos más interesantes y difíciles en la actualidad ([2]) y está relacionado con muchas otras áreas del análisis clásico, teoría de la medida, análisis funcional, teoría de números, teoría de conjuntos, etc. Un estupendo y completo trabajo sobre el tema es [24].

4.3 Valores propios, funciones propias y coeficientes de Fourier

Comentemos por último algunos aspectos relacionados con valores propios, dominios isoespectrales, etc. Como hemos mencionado con anterioridad, el interés de Fourier por desarrollos de la forma (9) estuvo motivado por la aplicación del método de separación de variables al problema (6). Más precisamente, si se buscan soluciones elementales de (6) de la forma $u(x,t) = X(x)P(t)$, ello origina el problema de valores propios

$$X''(x) + \lambda X(x) = 0,\ x \in (0,\pi),\ X(0) = X(\pi) = 0 \tag{37}$$

Es conocido que (37) tiene solución no trivial si y solamente si $\lambda \in \{n^2,\ n \in \mathbb{N}\}$. Además, si $\lambda = n^2$, para algún n natural, el conjunto de soluciones de (37) es un espacio vectorial real de dimensión uno engendrado por la función $\text{sen}(nx)$.

Para el problema de la propagación del calor, las condiciones de contorno que se consideran pueden ser mucho más generales que las establecidas en (6). De hecho, desde el punto de vista de las aplicaciones, tienen gran interés condiciones de contorno tales como

$$\begin{aligned} &\alpha_1 u(0,t) + \alpha_2 u_x(0,t) = 0,\ t \geq 0, \\ &\beta_1 u(\pi,t) + \beta_2 u_x(\pi,t) = 0,\ t \geq 0, \end{aligned}$$

donde u_x indica la derivada parcial respecto de la variable x y $\alpha_1, \alpha_2, \beta_1, \beta_2$ son números reales dados. Esto conduce a la posibilidad de desarrollos en serie que usen las funciones propias de problemas de contorno muy generales. A este respecto, la teoría de problemas de contorno del tipo Sturm-Liouville proporciona, de manera bastante general, bases del espacio $L^2(a,b)$ (el espacio de funciones de cuadrado integrable en el sentido de Lebesgue) que pueden usarse en los problemas a estudiar. Estas ideas fueron desarrolladas, en el siglo XIX (concretamente entre 1829 y 1837) por Sturm, profesor de Mecánica en la Sorbona y por Liouville, profesor de matemáticas en el College de Francia. Con la ayuda del lenguaje de hoy en día, sus resultados pueden resumirse de la forma siguiente: consideremos un problema de contorno de la forma (λ es un parámetro real):

$$\begin{aligned} &\frac{d}{dt}\left[\ p(t)\frac{dx(t)}{dt}\ \right] + (\lambda - q(t))x(t) = 0,\ t \in [a,b] \\ &\alpha_1 x(a) + \alpha_2 x'(a) = 0 \\ &\beta_1 x(b) + \beta_2 x'(b) = 0, \end{aligned} \tag{38}$$

donde suponemos las siguientes hipótesis:
1) $p \in C^1([a,b],\mathbb{R})$; además $p(t) > 0,\ \forall\ t \in [a,b]$; $\ q \in C([a,b],\mathbb{R})$.

3) $\alpha_1, \alpha_2, \beta_1$ y β_2 son números reales dados tales que $|\alpha_1| + |\alpha_2| > 0$ y $|\beta_1| + |\beta_2| > 0$.

Sturm y Liouville demostraron:
a) Cualquier valor propio de (38) es de multiplicidad 1.
b) Cualquier par de funciones propias x e y, asociadas respectivamente a valores propios distintos λ y μ, son ortogonales, es decir,

$$\int_a^b x(t)y(t)\ dt = 0$$

c) El conjunto de valores propios de (38) es infinito numerable. El sistema ortonormal de funciones propias asociado $\{\phi_n,\ n \in \mathbb{N}\}$, es una base de $L^2(a,b)$.
d) Sea $g \in C^2[a,b]$ cualquier función satisfaciendo las condiciones de contorno dadas en (38). Entonces

$$g(t) = \sum_{n=1}^{\infty} \langle g, \phi_n \rangle \phi_n(t),\ \forall\ t \in [a,b]$$

donde la serie converge de manera absoluta y uniforme en $[a,b]$ ($\langle , \rangle$ indica el producto escalar usual de funciones).

Una de las maneras más bonitas y sencillas de probar los resultados de Sturm y Liouville es usando el concepto de función de Green. Ello permite transformar (38) en una ecuación integral equivalente y trabajar, a partir de ahí, con operadores integrales. De esta forma van surgiendo de manera natural una serie de propiedades que, puestas de manera abstracta, dan lugar a la teoría de operadores compactos y autoadjuntos. Esta teoría, debida en gran parte a Fredholm y Hilbert, tuvo su origen a finales del siglo XIX ([25]) y principios del XX y proporcionó muchas ideas claves para el nacimiento del análisis funcional. Permite generalizar de manera destacada la teoría de los desarrollos de Fourier, y legitima el uso de métodos análogos en problemas aparentemente muy diferentes de los aquí considerados. Por ejemplo, si estamos tratando el problema de la conducción del calor en un dominio (conexo, abierto y acotado) Ω de $\mathbb{R}^3$ en lugar de en una varilla unidimensional como en (6), tendríamos el problema

$$\begin{aligned} \Delta_x u(x,t) &= \frac{\partial u(x,t)}{\partial t},\ (x,t) \in \Omega \times (0,T), \\ u(x,t) &= 0,\ \forall\ (x,t) \in \partial\Omega \times [0,T], \\ u(x,0) &= f(x),\ \forall\ x \in \overline{\Omega}, \end{aligned} \tag{39}$$

siendo Δ_x el operador laplaciano con respecto a $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Por su parte, $\partial\Omega$ indica la frontera del conjunto Ω.

La aplicación del método de separación de variables al problema anterior, origina, en lugar de (37), que es un problema de ecuaciones diferenciales ordinarias, el problema

$$\Delta X(x) + \lambda X(x) = 0,\ x \in \Omega;\ X(x) = 0,\ x \in \partial\Omega. \tag{40}$$

Ahora puede aplicarse la teoría espectral de operadores compactos y autoadjuntos ([7]) para demostrar que el conjunto de valores propios de (40) es infinito numerable y que el conjunto de funciones propias asociadas, convenientemente ortonormalizadas, $\{X_n(x),\ n \in \mathbb{N}\}$, forma una base del espacio $L^2(\Omega)$. Esto justifica el hecho de que la condición inicial f se exprese como

$$f(x) = \sum_{n=1}^{\infty} a_n X_n(x), \tag{41}$$

para coeficientes convenientes a_n. Así, la solución de (39) es de la forma

$$u(x,t) = \sum_{n=1}^{\infty} a_n P_n(t) X_n(x), \tag{42}$$

para funciones P_n convenientes. Ideas parecidas pueden aplicarse al estudio del problema de la cuerda vibrante (1) en dimensiones superiores, así como a otros problemas de naturaleza diferente.

Existen en la actualidad muchas cuestiones de interés en torno al problema de valores propios (40), que lo consideraremos en adelante para un dominio (abierto y acotado) de $\mathbb{R}^n$. Una de ellas se relaciona con el conjunto de valores propios $\{\lambda_n(\Omega),\ n \in \mathbb{N}\}$ (al que denotaremos en adelante por espectro de Ω) y fue planteada por M. Kac en 1966 ([21]) en un famoso artículo titulado: can one hear the shape of a drum? Dos dominios Ω_1 y Ω_2 se dicen isoespectrales si tienen el mismo espectro. Dos dominios se dicen isométricos, si son congruentes en el sentido de la geometría euclídea. De las propiedades del operador laplaciano se deduce trivialmente que dos dominios isométricos son isoespectrales. La conjetura planteada por Kac fue: dos dominios isoespectrales, ¿son necesariamente isométricos? Esta misma cuestión puede plantearse para condiciones de contorno más generales y para otros tipos de operadores diferentes del laplaciano ([31]).
En 1982, Urakawa ([33]) mostró un ejemplo de dominios isoespectrales en $\mathbb{R}^n,\ n \geq 4$, que no son isométricos. En 1992, Gordon, Webb y Wolpert ([18]) expusieron un contraejemplo en $\mathbb{R}^2$. Como se puede comprender, hay muchos problemas abiertos aún.

Un último aspecto, relacionado con los coeficientes de Fourier, que vamos a comentar implica dos conceptos aparentemente alejados: el grado topológico de Brouwer de una aplicación continua y los coeficientes de Fourier de dicha aplicación (si viviésemos lo suficiente, al final tendríamos oportunidad de comprobar que todo en matemáticas está relacionado). Si B_1 es la bola cerrada

unidad de $\mathbb{R}^2$ y $g : B_1 \to \mathbb{R}^2$ es una aplicación continua que no se anula en la frontera de B_1, sabemos que su grado topológico está bien definido (véase, por ejemplo [13]). Denotémoslo por $deg(g)$. Si definimos la función $2\pi-$ periódica $h(x) = g(\exp(ix))$, puede demostrarse que, si g es una función de clase C^1, entonces

$$deg(g) = \sum_{n=-\infty}^{+\infty} n|h_n|^2, \tag{43}$$

donde h_n son los coeficientes de Fourier de la función h respecto de la base $\{\exp(inx),\ n \in \mathbb{N}\}$. De hecho, (43) fue probado por Brezis y Nirenberg bajo condiciones más generales ([8]). También conjeturaron que (43) debería ser verdad para funciones continuas g. La respuesta negativa a esta conjetura ha sido dada por Korevaar en 1999 ([26]). No obstante, esto ha planteado nuevos interrogantes como el propuesto por Brezis con el título siguiente: can one hear the degree of continuous maps? De manera más precisa: si h y k son dos aplicaciones continuas de la frontera de B_1 en sí misma, con coeficientes de Fourier respectivos h_n y k_n verificando $|h_n| = |k_n|,\ \forall n \in \mathbb{N}$, ¿es verdad que $deg(h) = deg(k)$?. Puede consultarse el reciente trabajo de Brezis ([8]) sobre este tema. Es claro que cuestiones relacionadas con los coeficientes de Fourier continúan desempeñando un papel importante en la historia de la matemática.

No cabe duda de que la teoría de series de Fourier es una de las creaciones más grandes de la Historia de la Ciencia. Ha tenido, además, una gran influencia en el nacimiento y desarrollo de numerosas técnicas y conceptos matemáticos. En la actualidad, la teoría de series de Fourier sigue teniendo una gran importancia y su conocimiento es de gran utilidad en disciplinas muy diversas como matemáticas, física, biología, ingeniería, economía, etc. Tales series están siempre presentes en todos aquellos procesos naturales de tipo oscilatorio, de difusión o de naturaleza periódica. Por mencionar algunos, los métodos de Fourier se emplean en problemas tan diversos como los relacionados con: el ciclo de las manchas solares, predicción de mareas, mejora de la calidad de las imágenes de los objetos celestes tomadas desde el espacio, física de plasmas, física de semiconductores, acústica, sismografía, oceanografía, confección de imágenes en medicina (escáner TAC), estudio del ritmo cardíaco, análisis químicos, estudios de rayos X (usando el análisis de Fourier, los astrónomos pueden estudiar las variaciones en intensidad de las señales de rayos X de un objeto celeste), etc.

Me gustaría acabar con las palabras de Lord Kelvin, que siguen teniendo plena actualidad: *Los métodos de Fourier no son solamente uno de los resultados más hermosos del análisis moderno, sino que puede decirse además que proporcionan un instrumento indispensable en el tratamiento de casi todas las cuestiones de la física actual, por recónditas que sean.*

Agradecimientos. Deseo dar las gracias a J. M. Vegas, por haberme ofrecido la posibilidad de realizar esta colaboración de carácter histórico

sobre Fourier. Asimismo, deseo agradecer a J. Alaminos su información sobre la página web donde se puede encontrar el texto original del libro de Fourier y su imprescindible ayuda en la inclusión de figuras en este trabajo. Las fotografías que aparecen en el texto han sido obtenidas de la dirección `http://www-groups.dcs.st-and.ac.uk/~history/index.html`. Las reproducciones de las páginas del libro original de Fourier han sido obtenidas de la dirección `http://gallica.bnf.fr/ark:/12148/bpt6k29061r`

Referencias

[1] J. Arias de Reyna. *Pointwise convergence of Fourier series.* Lecture Notes in Mathematics, nº 1785, Springer-Verlag, Berlin, 2002.

[2] J. M. Ash y S.T. Tetunashvili. *New uniqueness theorems for trigonometric series.* Proc. Amer. Math. Soc., 128, 2000, 2627-2636.

[3] R. Beals. *Analysis. An Introduction.* Cambridge University Press, Cambridge, 2004.

[4] J. Alaminos, C. Aparicio, P. Muñoz y A.R. Villena. *Un recorrido histórico del teorema fundamental del cálculo.* Sometido a publicación.

[5] R. Apery. *Irrationalité de* $\zeta(2)$ *et* $\zeta(3)$. Astérisque 61, 1979, 11-13.

[6] A. Baouche y S. Dubuc. *La non-dérivabilité de la fonction de Weierstrass.* Enseign. Math., 38, 1992, 89-94.

[7] H. Brezis. *Análisis Funcional.* Madrid, Alianza Universidad Textos, 1984.

[8] H. Brezis. *New questions related to the topological degree.* The unity of mathematics, Progr. math. nº 244, Birkhäuser Boston, Boston, 2006, 137-154.

[9] A. Cañada. *Una perspectiva histórica de las series de Fourier: de las ecuaciones de ondas y del calor a los operadores compactos y autoadjuntos.* Relime, Revista Latinoamericana de investigación en Matemática educativa, 3, 2000, 293-320.

[10] A. Cañada. *Brook Taylor, tercer centenerario de su nacimiento.* Epsilon, 4, 1985, 104-111.

[11] L. Carleson. *On convergence and growth of partial sums of Fourier series.* Acta Math., 116, 1966, 135-157.

[12] A. Córdoba. *Disquisitio numerorum.* Gac. R. Soc. Mat. Esp. 4, 2001, 249-260.

[13] K. Deimling *Nonlinear Functional Analysis.* Springer-Verlag, Berlin, 1985.

[14] W. Dunham. *Euler. The master of us all.* The Mat. Ass. Am. Dolciani Mat. Expositions, 22, 1999. Traducido al español por Jesús Fernández, con comentarios de Antonio Pérez Sanz. (Euler. El maestro de todos los matemáticos). (Spanish) La Matemática en sus Personajes. 6. Madrid: Nivola Libros Ediciones, 2000.

[15] S. Fischler. *Irrationalité de valeurs de zeta (d'après Apéry, Rivoal,...)* Astérisque, 294, 2004, 27-62.

[16] J.B. Fourier. *The analytical theory of heat.* Dover Publ., New York, 1955.

[17] E.A. González-Velasco. *Connections in mathematical analysis: the case of Fourier series.* Am. Math. Mon. 99, 1992, 427-441.

[18] C. Gordon, D. Web y S. Wolpert. *Isospectral plane domains and surfaces via Riemannian orbifolds.* Invent. Math., 110, 1992, 1-22.

[19] M. de Guzmán. *Impactos del análisis armónico.* Discurso de ingreso en la Real Academica de Ciencias Exactas, Físicas y Naturales de Madrid. http://www.mat.ucm.es/deptos/am/guzman/impactoanalisisarmonico.htm, 1983.

[20] G. H. Hardy. *Weierstrass's Non-Differentiable Function.* Trans. Amer. Math. Soc. 17, 1916, 301-325.

[21] M. Kac. *Can one hear the shape of a drum?* Amer. Math. Monthly, 73, 1966, 1-23.

[22] J.P. Kahane *A century of interplay between Taylor series, Fourier series and Brownian motion.* Bull. Lond. Math. Soc., 29, 1997, 257-279.

[23] J.P. Kahane y Y. Katznelson. *Sur les ensembles de divergence des séries trigonométriques.* Studia Math. 26, 1966, 305-306.

[24] A. S. Kechris *Set theory and uniqueness for trigonometric series.* Preprint, 1997. http://www.math.caltech.edu/people/kechris.html

[25] M. Kline. *Mathematical thought from ancient to modern times.* Oxford University Press, New York, 1972. Versión española en Alianza Editorial, Madrid, 1992.

[26] J. Korevaar. *On a question of Brezis and Nirenberg concerning the degree of circle maps.* Sel. Math., New Ser. 5, 1999, 107-122.

[27] T. W. Körner. *Fourier analysis.* Cambridge University Press, Cambridge, 1988.

[28] N.N. Luzin. *Función.* Gac. R. Soc. Mat. Esp., 6, 2003, 201-225.

[29] H. Okamoto. *A remark on continuous, nowhere differentiable functions.* Proc. Japan Acad. 81, Ser. A, 2005, 47-50.

[30] A. van der Poorten *A proof that Euler missed...Apéry's proof of the irrationality of* $\zeta(3)$. Math. Intelligencer, 1, 1978/79, 195-203.

[31] M.H. Protter. *Can one hear the shape of a drum? Revisited.* SIAM Rev. 29, 1987, 185-197.

[32] D.J. Struik (editor). *A sourcebook in Mathematics, 1200-1800.* Hardvard University Press, XIV, Cambridge, Massachusetts, 1969.

[33] H. Urakawa. *Bounded domains which are isospectral but not congruent.* Ann. Scient. Ecole Norm. Sup. 15, 1982, 441-456.

Estimados socios:

El **premio SEMA al joven investigador** de este año ha recaído en Jorge Cortés de la Universidad de California Santa Cruz. El premio de "Divulgación de la Matemática Aplicada" ha quedado desierto en esta edición.

Recordad que podéis consultar toda la información de la Sociedad desde la *renovada* página web de SEMA (`www.sema.org.es`).

Un cordial saludo,

Carlos Castro.

Tipo de evento:	Congreso
Nombre:	Second Chilean Workshop on Numerical Analysis of Partial Differential Equations (WONAPDE 2007)
Lugar:	Concepción (Chile)
Fecha:	del 15 al 19 de enero de 2007
E-mail:	wonapde_2007@ing-mat.udec.cl
WWW:	www.ing-mat.udec.cl/wonapde2007

Tipo de evento:	Encuentro
Nombre:	The Future of Computational Acoustics
Lugar:	Londres
Fecha:	22 y 23 de febrero de 2007
Organiza:	Professor Keith Attenborough (Hull) and Professor Simon Chandler-Wilde (Reading)
E-mail:	K.Attenborough@hull.ac.uk; s.n.chandler-wilde@reading.ac.uk
WWW:	www.newton.cam.ac.uk/programmes/HOP/hop_london.html

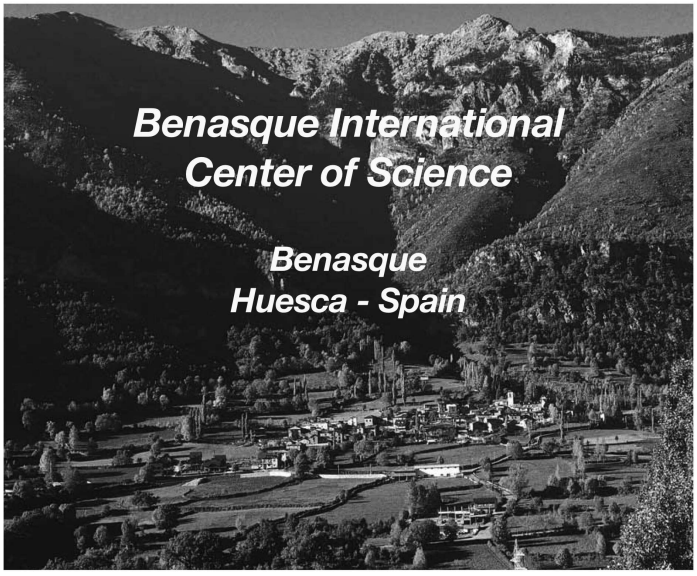
Tipo de evento:	Congreso
Nombre:	Stochastic Dynamical Systems and Control
Lugar:	Berkeley, CA
Fecha:	del 19 al 23 de marzo de 2007
Organiza:	Jonathan Mattingly (Duke), Igor Mezic (UCSB-Chair), Andrew Stuart (Warwick)
E-mail:	msri-workshops@msri.org
WWW:	www.msri.org/calendar/workshops/WorkshopInfo/387/show_workshop

Tipo de evento:	Jornada
Nombre:	World Congress on Computational Finance. The First Decade
Lugar:	Londres
Fecha:	26 de marzo de 2007
Organiza:	Mathematical Sciences Research Institute (MSRI), Berkeley, CA
Información:	Domingo Tavella en tavella@octanti.com o Jesper Andreasen en jesper.andreasen@bankofamerica.com
E-mail:	msri-workshops@msri.org
WWW:	www.msri.org/specials/compfinance

Tipo de evento:	Congreso
Nombre:	Mathematical Issues in Stochastic Approaches for Multiscale Modeling
Lugar:	Berkeley, CA
Fecha:	del 21 al 25 de mayo de 2007
Organiza:	Roberto Camassa (UNC - Chapel Hill), Jinqiao Duan (Illinois Institute of Technology - Chicago), Peter E. Kloeden (U of Frankfurt, Germany), Jonathan Mattingly (Duke U), Richard McLaughlin (UNC - Chapel Hill)
Información:	Domingo Tavella en tavella@octanti.com o Jesper Andreasen en jesper.andreasen@bankofamerica.com
E-mail:	msri-workshops@msri.org
WWW:	www.msri.org/calendar/workshops/WorkshopInfo/398/show_workshop

Tipo de evento:	Congreso
Nombre:	6th International Congress on Industrial and Applied Mathematics (ICIAM 07)
Lugar:	Zurich (Suiza)
Fecha:	del 16 al 20 de julio de 2007
Organiza:	Sociedad Matemática Suiza (SMG)
Información:	Domingo Tavella en tavella@octanti.com o Jesper Andreasen en jesper.andreasen@bankofamerica.com
E-mail:	msri-workshops@msri.org
WWW:	www.iciam07.ch

Partial Differential Equations, Optimal Design and Numerics



Summer School & Workshop

2007, August 27 - September 7

Scientific Organizers:

G. BUTTAZZO
Università di Pisa

E. ZUAZUA
Universidad Autónoma de Madrid

Confirmed Scientists:

L. AMBROSIO (Pisa, Italy),
J.A. CARRILLO (Barcelona, Spain)
V. CASELLES (Barcelona, Spain),
F. COLOMBINI (Pisa, Italy),
G. DAL MASO (Trieste, Italy),
R. DONAT (Valencia, Spain),
H. KOCH (Bonn, Germany)
G. LEUGERING (Erlangen, Germany),
J. P. PUEL (Versailles, France),
G. TOSCANI (Pavia, Italy)
M. VANNINATHAN (Bangalore, India)

The workshop is intended to provide a fruitful atmosphere for discussions and joint research work on themes involving partial differential equations and their applications to shape optimization, optimal control problems, singularities in fracture mechanics and fluid dynamics, and numerical analysis.

The work will be focused on new research trends in the fields above, in order to stimulate collaborations mong participants by means of various activities on the basis of a daily programme (talks, seminars, minicourses, discussions,...)

The activity is mainly intended for young scientists as PhD students and post-docs, and will be held thanks to the participation of a number of world leader mathematicians.

For further information, registration, accomodation, etc see http://benasque.ecm.ub.es

The deadline for registration is March 31 2007 and for housing reservations May 31st. Partial financial support will be available for young researchers.

Those willing to participate may send an email to the scientific organizers
buttazzo@dm.unipi.it, enrique.zuazua@uam.es

with a short CV, a letter of presentation describing the interest in the activity and a letter of recommendation.
For further local information about Benasque and the region: http://www.benasque.com

CONSOLIDER INGENIO 2010

CONSOLIDER-MATHEMATICA

CONSOLIDER MATHEMATICA is a Spanish project proposing an ambitious research program for the period 2006-2011, with the main objective of promoting strategic action to increase both qualitative and quantitatively the presence of Mathematics in science, technology and innovation.

CONSOLIDER MATHEMATICA is an initiative promoted and supported by the spanish Ministry of Education and Research with 7.500.000€.

CONSOLIDER MATHEMATICA is a network structured around a Research Coordinator, a Management Center, a Board of Directors, five nodes (CESGA, CIEM, CRM, ICM and IMUB) and more than 340 research groups.

The MATHEMATICA Research Program

From basic research to applications
- Enigmas to decipher: Fundamental Mathematics and Cryptography.
- From Classical Analysis to the digital era.
- New frontiers in Algebra and Geometry.

How to understand the physical world
- From Geometry and Topology to Physics and Cosmological models.
- Mathematics, Dynamics and Complexity.
- Partial Differential Equations as a tool for modeling.

The essential computational support
- Intelligent structures and materials: design and control.
- Math software: Mathematics as the basic foundation for Scientific and High Performance Computing.
- New techniques and horizons in Computational Mathematics.

Direct applications to society
- Mathematics and the information society and communications.
- Statistical models and their applications.
- Stochastic modeling of complex evolving phenomena and configurations in random media.
- Optimization and decision-making support techniques.
- New horizons in Mathematics education and training.
- Bringing Mathematics closer to society.

The Objectives of MATHEMATICA

To improve the role of mathematical research in the Spanish system of science, technology and innovation.

To increase and promote the activities of transference of knowledge and technology of the Spanish mathematicians.

To promote the use of computational methods both inside and outside

MINISTERIO DE EDUCACIÓN Y CIENCIA

CONSOLIDER INGENIO 2010

mathematical research.

To achieve greater recognition for Spanish research groups at an international level and to increase the presence of Spanish mathematicians in strategic areas.

To create a Doctorate School of international status.

To use research and innovation to improve education and mathematical training at all levels.

To make the results of mathematical research more accessible both from within and from outside Mathematics.

The Tools of MATHEMATICA

The Platforms of MATHEMATICA
- The MATHEMATICA FUTURE platform
- The MATHEMATICA CONSULTING platform
- The MATHEMATICA COMPUTING platform
- The MATHEMATICA EDU platform
- The MATHEMATICA WEB platform

The Thematic Projects of MATHEMATICA
- The MATHEMATICA International Graduate School
- MATHEMATICA Programs of Intensive Research

The MATHEMATICA Cross-sectional Support Services
- Meetings and Workshops
- Research Institutes and Thematic Networks
- The MATHEMATICA Virtual House

The MATHEMATICA Board of Directors

Alfredo Bermudez de Castro, Joaquim Bruna, Eduardo Casas, Antonio Duran, Laureano Gonzalez-Vega, Manuel de León,Marco Antonio Lopez-Cerda, Ignacio Luengo, Consuelo Martinez, Marta Sanz-Sole, Oriol Serra, Carles Simo, Luis Vega, Enrique Zuazua

Research Coordinator
Enrique Zuazua
Management Center
Universidad de Cantabria

More Information:
mathematica@unican.es
http://www.mathematica.unican.es

With the support of:

UC UNIVERSIDAD DE CANTABRIA

IDICAN Sociedad Regional Cantabria de I+D+i

Belmonte Beima, Juan
Estudiante (Becario). *Líneas de investigación:* Matemática Aplicada: Dinámica no lineal – Univ. de Castilla - La Mancha – E.T.S.I. Industriales – Dpto. de Matemáticas – Avda. Camilo José Cela, s/n; 13071 Ciudad Real.
Tlf.: . *Fax:* .
e-mail: Juan.belmonte@uclm.es

Climent Coloma, Joan Josep
Catedrático de Universidad. *Líneas de investigación:* Teoría de códigos, criptografía, seguridad informática – Univ. de Alicante – Escuela Politécnica Superior – Dpto. de Ciencia de la Computación e Inteligencia Artificial – Campus de Sant Vicent del Raspeig, Ap. Correos 99; 03080 Alicante.
Tlf.: 965903655. *Fax:* 965903902.
e-mail: jcliment@dccia.ua.es
http://www.dccia.ua.es/∼jcliment

Direcciones útiles

Consejo Ejecutivo de SëMA

Presidente:

Juan Ignacio Montijano. (monti@unizar.es).
Dpto. de Matemática Aplicada.Facultad de Ciencias. Edificio de Matemáticas. Ciudad Universitaria s/n. 50016 Zaragoza. *Tel:* 976 761 120.

Secretario:

Carlos Castro Barbero. (ccastro@caminos.upm.es).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

Tesorera:

María Pilar Laburta Santamaría. (laburta@unizar.es).
Dpto. de Matemática Aplicada. Centro Pol. Superior. Univ. de Zaragoza. Edificio Torres Quevedo. C/ María de Luna 3. 50018 Zaragoza. *Tel:* 976 762 006.

Vocales:

Rafael Bru García. (rbru@mat.upv.es)
Dpto. de Matemática Aplicada. E.T.S.I. Agrónomos. Univ. Politécnica de Valencia. Camí de Vera, s/n. 46022 Valencia. *Tel:* 963 879 669.

José Antonio Carrillo de la Plata. (carrillo@mat.uab.es)
Dpto. de Matemáticas. Univ. Autónoma de Barcelona. Edifici C. 08193 Bellaterra (Barcelona). *Tel:* 93 581 2413.

Javier Chavarriga Soriano. (chava@eup.udl.es).
Dpto. de Matemática. E.U. Politécnica. Univ. de Lleida. Avda. Jaume II, 69. 25001 Lleida. *Tel:* 973 702 777.

Inmaculada Higueras Sanz. (higueras@unavarra.es).
Dpto de Matemática e Informática Univ. Pública de Navarra. Campus de Arrosadía, s/n. *Tel:* 948 169 526. 31006 Pamplona.

Pablo Pedregal Tercero. (Pablo.Pedregal@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real.

Ireneo Peral Alonso. (ireneo.peral@uam.es).
Dpto. de Matemáticas, C-XV. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Ctra. de Colmenar, km. 14. 28049 Madrid. *Tel:* 913 974 204.

José Javier Valdés García. (valdes@orion.ciencias.uniovi.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. de Oviedo. Avda. de Calvo Sotelo, s/n. 33007 Oviedo. *Tel:* 985 103 340.

Enrique Zuazua Iriondo. (enrique.zuazua@uam.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Ctra. de Colmenar, km. 14. 28049 Madrid. *Tel:* 914 974 368.

Comité Científico del Boletín de SëMA

Grupo Editor del Boletín de SëMA

Luis Ferragut Canals. (ferragut@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400, ext. 1522.

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Francisco Andrés Pérez. (franc@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400 ext. 1537.

M. Isabel Asensio Sevilla. (mas@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias Químicas. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400 ext. 1537.

M. Teresa de Bustos Muñoz. (tbustos@usal.es).
Dpto. de Matemática Aplicada. E.T.S. Ing. Ind. de Béjar. Univ. de Salamanca. Avda. Fernando Ballesteros, 2. 37700 Béjar, Salamanca. *Tel:* 923 408 080 ext. 2263.

Antonio Fernández Martínez. (anton@usal.es).
Dpto. de Matemática Aplicada. E. Politécnica Superior Zamora. Univ. de Salamanca. Avda. Requejo, 33. Campus Viriato. 49022 Zamora. *Tel:* 980 545 000 ext. 4459.

Responsables de secciones del Boletín de SëMA

Artículos:

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Matemáticas e Industria:

Mikel Lezaun Iturralde. (mepleitm@lg.ehu.es).
Dpto. de Matemática Aplicada, Estadística e I. O. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Educación Matemática:

Roberto Rodríguez del Río. (rr_delrio@mat.ucm.es).
Dpto. de Matemática Aplicada. Fac. de Químicas. Univ. Compl. de Madrid. Ciudad Universitaria. 28040 Madrid. *Tel:* 913 944 102.

Resúmenes de libros:

Fco. Javier Sayas González. (jsayas@posta.unizar.es).
Dpto. de Matemática Aplicada. Centro Politécnico Superior . Universidad de Zaragoza. C/María de Luna, 3. 50015 Zaragoza. *Tel:* 976 762 148.

Noticias de SëMA:

Carlos Castro Barbero. (ccastro@caminos.upm.es).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

Anuncios:

Óscar López Pouso. (oscarlp@usc.es).
Dpto. de Matemática Aplicada. Fac. de Matemáticas. Univ. de Santiago de Compostela. Campus sur, s/n. 15782 Santiago de Compostela *Tel:* 981 563 100, ext. 13228.

Responsables de otras secciones de SëMA

Gestión de Socios:

María Pilar Laburta Santamaría. (laburta@unizar.es).
Dpto. de Matemática Aplicada. Centro Politécnico Superior. Univ. de Zaragoza. Edificio Torres Quevedo. C/ María de Luna 3. 50018 Zaragoza. *Tel:* 976 762 006.

Página web: www.sema.org.es/**:**

J. Rafael Rodríguez Galván. (rafael.rodriguez@uca.es).
Dpto. de Matemáticas. Fac. de CC. EE. y Empresariales. Univ. de Cádiz. C/ Duque de Nájera, 6. 11002 Cádiz. *Tel:* 956 015 478.

1. Los artículos publicados en este Boletín podrán ser escritos en español o inglés y deberán ser enviados por correo certificado a

Prof. E. FERNÁNDEZ CARA

Presidente del Comité Científico, Boletín SëMA

Dpto. E.D.A.N., Facultad de Matemáticas

Aptdo. 1160, 41080 SEVILLA

También podrán ser enviados por correo electrónico a la dirección

`boletin_sema@usal.es`

En ambos casos, el/los autor/es deberán enviar por correo certificado una carta a la dirección precedente mencionando explícitamente que el artículo es sometido a publicación e indicando el nombre y dirección del autor corresponsal. En esta carta, podrán sugerirse nombres de miembros del Comité Científico que, a juicio de los autores, sean especialmente adecuados para juzgar el trabajo.

La decisión final sobre aceptación del trabajo será precedida de un procedimiento de revisión anónima.

2. Las contribuciones serán preferiblemente de una longitud inferior a 24 páginas y se deberán ajustar al formato indicado en los ficheros a tal efecto disponibles en la página web de la Sociedad (`http://www.sema.org.es/`).

3. El contenido de los artículos publicados corresponderá a un área de trabajo preferiblemente conectada a los objetivos propios de la Matemática Aplicada. En los trabajos podrá incluirse información sobre resultados conocidos y/o previamente publicados. Se anima especialmente a los autores a presentar sus propios resultados (y en su caso los de otros investigadores) con estilo y objetivos divulgativos.

Ficha de Inscripción Individual

Sociedad Española de Matemática Aplicada SēMA

Remitir a: SEMA, Despacho 520, Facultad de Matemáticas,
Universidad Complutense. 28040 Madrid.
Fax: 913 944 607. CIF: G-80581911

Datos Personales

- Apellidos: .....
- Nombre: .....
- Domicilio: .....
- C.P.: ..... Población: .....
- Teléfono: ..... DNI/CIF: .....
- Fecha de inscripción: .....

Datos Profesionales

- Departamento: .....
- Facultad o Escuela: .....
- Universidad o Institución: .....
- Domicilio: .....
- C.P.: ..... Población: .....
- Teléfono: ..... Fax: .....
- Correo electrónico: .....
- Página web: `http://` .....
- Categoría Profesional: .....
- Líneas de Investigación: .....

.....

Dirección para la correspondencia: ☐ Profesional ☐ Personal

Cuota anual para el año 2005

☐ Socio ordinario: 30 EUR. ☐ Socio de reciprocidad con la RSME: 12 EUR.
☐ Socio estudiante: 15 EUR. ☐ Socio extranjero: 25 EUR.

Datos bancarios

. . . de . . . . . de 200. .

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de S$\vec{e}$MA (Sociedad Espa nola de Matemática Aplicada) sean pasados al cobro en la cuenta cuyos datos figuran a continuación

Entidad (4 dígitos)	Oficina (4 dígitos)	D.C. (2 dígitos)	Número de cuenta (10 dígitos)

- Entidad bancaria: . . . . . .
- Domicilio: . . . . . .
- C.P.: . . . . . Población: . . . . . .

Con esta fecha, doy instrucciones a dicha entidad bancaria para que obren en consecuencia.

Atentamente,

Fdo. . . . . .

Para remitir a la entidad bancaria

. . . de . . . . . de 200. .

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de S$\vec{e}$MA (Sociedad Espa nola de Matemática Aplicada) sean cargados a mi cuenta corriente/libreta . . . . . en esa Agencia Urbana y transferidas a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander

Atentamente,

Fdo. . . . . .

Ficha de Inscripción Institucional

Sociedad Española de Matemática Aplicada SëMA

Remitir a: SEMA, Despacho 520, Facultad de Matemáticas,
Universidad Complutense. 28040 Madrid.
Fax: 913 944 607. CIF: G-80581911

Datos de la Institución

- Departamento: .....
- Facultad o Escuela: .....
- Universidad o Institución: .....
- Domicilio: .....
- C.P.: ..... Población: .....
- Teléfono: ..... DNI/CIF: .....
- Correo electrónico: .....
- Página web: `http://` .....
- Fecha de inscripción: .....

Forma de pago

La cuota anual para el año 2005 como Socio Institucional es de 150 EUR.
El pago se realiza mediante transferencia bancaria a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander